MYKOLO ROMERIO UNIVERSITETAS

Tatjana BILEVIČIENĖ
Steponas JONUŠAUSKAS

STATISTINIŲ METODŲ TAIKYMAS RINKOS TYRIMUOSE

MYKOLO ROMERIO UNIVERSITETAS

Tatjana Bilevičienė, Steponas Jonušauskas

STATISTINIŲ METODŲ TAIKYMAS RINKOS TYRIMUOSE

Vadovėlis

Vilnius 2011

UDK 311:339.13(075.8)
Bi238

2010 m. gruodžio 3 d. Nr. 10-16
Aukštųjų mokyklų vadovėlių kokybės vertinimo komisijos rekomenduota

R e c e n z a v o:
Vilniaus universiteto Kauno humanitarinio fakulteto Informatikos katedros doc. dr. **Eugenijus Telešius**;
Mykolo Romerio universiteto Ekonomikos ir finansų valdymo fakulteto Verslo ekonomikos katedros katedros prof. dr. **Vitalija Rudzkienė**

Vadovėlis svarstytas Mykolo Romerio universiteto Ekonomikos ir finansų valdymo fakulteto tarybos 2010 m. rugpjūčio 30 d. posėdyje (protokolas Nr. 1EFV-11) ir rekomenduotas spausdinti

Vadovėlis svarstytas Mykolo Romerio universiteto Socialinės informatikos fakulteto tarybos 2010 m. birželio 22 d. posėdyje (protokolas Nr. 1SI-26) ir rekomenduotas spausdinti

Vadovėlis svarstytas Mykolo Romerio universiteto Socialinės informatikos fakulteto Informatikos ir programų sistemų katedros 2010 m. birželio 10 d. posėdyje (protokolas Nr. 1INFPSK-7) ir rekomenduotas spausdinti

Mykolo Romerio universiteto mokslinių-mokomųjų leidinių aprobavimo spaudai komisija 2010 m. rugsėjo 14 d. posėdyje (protokolas Nr. 2L-1) vadovėlį patvirtino spausdinti

Visos leidinio leidybos teisės saugomos. Šis leidinys arba kuri nors jo dalis negali būti dauginami, taisomi arba kitu būdu platinami leidėjui nesutikus

ISBN 978-9955-19-242-8 © Mykolo Romerio universitetas, 2011

TURINYS

PRATARMĖ..... 6

1. STATISTINIO TYRIMO PROJEKTAS..... 8
1.1. Tyrimo tikslas ir tyrimo modelis..... 8
1.2. Kokybiniai ir kiekybiniai tyrimai. Tyrimo tipo pasirinkimas..... 13
1.3. Anketinės apklausos metodas..... 17
1.3.1. Kintamųjų tipai ir matavimo skalės pasirinkimas..... 19
1.3.2. Imties sudarymo būdas ir dydis..... 26
1.3.3. Anketos kodavimo schemos sudarymas. Duomenų kodavimas..... 33
2. DUOMENŲ ĮVEDIMAS Į KOMPIUTERĮ..... 37
2.1. SPSS duomenų bazės pildymas..... 37
2.2. Duomenų įkėlimas iš kitų šaltinių..... 56

3. PIRMINĖ STATISTINĖ DUOMENŲ ANALIZĖ..... 65
3.1. Duomenų keitimas ir atranka..... 66
3.2. Duomenų filtravimas..... 76
3.3. SPSS paketo sintaksės redaktorius..... 86
3.4. Aprašomoji statistika..... 90
3.4.1. Vienmatė dažnių analizė..... 95
3.4.2. Porinės dažnių lentelės..... 97
3.4.3. Pagrindiniai grafikai: histogramos, stulpelinės ir skritulinės diagramos..... 99
3.4.4. Sudėtingesnių grafikų braižymas. Trimatės (pirminės) ir sumuojančios diagramos..... 107
3.4.5. Padėties ir sklaidos charakteristikos..... 133

4. STATISTINĖS HIPOTEZĖS..... 145
4.1. Imties skirstiniai. Įverčiai..... 145
4.2. Statistinių hipotezių tikrinimas..... 146
4.3. Kintamojo skirstinio normalumo tikrinimas SPSS paketu..... 147

5. RYŠIŲ ANALIZĖ..... 154
5.1. Kintamųjų, išmatuotų nominaliąja arba rangų skale su nedideliu kategorijų skaičiumi, statistiniai ryšiai..... 154
5.1.1. Pirsono $\chi 2$ suderinamumo kriterijus..... 154
5.1.2. Pirsono $\chi 2$ suderinamumo kriterijaus skaičiavimas SPSS terpėje..... 158
5.2. Koreliacinė analizė..... 165
5.2.1. Spirmeno (*Spearman*) koreliacijos koeficientas..... 166
5.2.2. Spirmeno koreliacijos koeficiento skaičiavimas SPSS terpėje..... 167
5.2.3. Kiti kokybinių kintamųjų ryšio koeficientai..... 172
5.2.4. Pirsono koreliacijos koeficientas..... 176
5.2.5. Pirsono koreliacijos koeficiento skaičiavimas SPSS terpėje..... 180

6. REGRESINĖ ANALIZĖ..... 189
6.1. Tiesinis (paprastasis) regresijos modelis..... 189
6.2. Netiesinė (kreivinė) regresija..... 192
6.2.1. Regresinė analizė SPSS terpėje..... 192
6.3. Daugialypės tiesinės regresijos modelis..... 199
6.3.1. Daugialypė regresija SPSS terpėje..... 199
6.4. Logistinė regresija..... 205
6.4.1. Logistinė regresija SPSS terpėje..... 206

7. SKIRTUMŲ ANALIZĖ..... 212
7.1. Hipotezės apie vidurkių lygybę..... 212
7.2. *T* kriterijus vienai imčiai..... 213
7.2.1. *T* kriterijaus vienai imčiai skaičiavimas SPSS terpėje..... 214
7.3. *T* kriterijus dviem nepriklausomoms imtims..... 216
7.3.1. *T* kriterijaus dviem nepriklausomoms imtims skaičiavimas SPSS terpėje..... 220
7.4. *T* kriterijaus dviem priklausomoms (porinėms) imtims skaičiavimas SPSS terpėje..... 225
7.5. Dispersinė analizė..... 227
7.5.1. Vienfaktorinė dispersinė analizė SPSS terpėje..... 232
7.5.2. Dviejų veiksnių dispersinė analizė SPSS terpėje..... 243

8. FAKTORINĖ ANALIZĖ..... 255
8.1. Faktorinė analizė SPSS terpėje..... 268

9. KLASTERINĖ ANALIZĖ..... 281
9.1. Klasterinė analizė SPSS terpėje..... 290

LITERATŪRA..... 298

PRIEDAI..... 300
1 lentelė. χ^2 skirstinio α lygmens kritinės reikšmės..... 300
2 lentelė. Stjudento skirstinio α lygmens kritinės reikšmės $t_\alpha(n)$..... 301

ATSAKYMAI..... 302

PRATARMĖ

Šiuolaikiniame globaliame pasaulyje sparčiai didėja konkurencija. Norėdami suvokti įmanomas rizikas ir galimybes, verslo atstovai siekia kuo tikslesnės informacijos apie jų srities rinką, galimus vartotojus ir konkurentus.

Naujoji rinkodara – tai strategijų ir metodų rinkinys, galintis paskatinti potencialius klientus patikėti paslaugomis ir prekėmis.

Rinkos tyrimų rezultatai yra pagrindas priimti įmonės steigimo sprendimą, įsisavinti naują rinkos segmentą, praplėsti prekybinius plotus, įsteigti filialą arba atstovybę kitame regione, plėtoti prekių liniją ir t. t. Visi šie sprendimai siejami su pinigų investavimu, o tai reiškia ir su jų praradimo rizika. Kad būtų galima sumažinti investuojamų lėšų praradimo tikimybę, pagrįsti investicijas ir pateikti labiau apibrėžtus jų grąžinimo laikotarpius yra vykdomi rinkos tyrimai. Tyrimų rezultatai pateikiami kartu su plėtojamo projekto verslo planu.

Pastaraisiais dešimtmečiais net ir mažos įmonės atlieka rinkos tyrimus, o gautą informaciją taiko rinkos dydžiui, konkurencinei aplinkai įvertinti, taip pat naujos rinkos strategijai (produkto, kainos, paskirstymo, pardavimo ir kt.) diegti. Klausti ir pasitikėti vartotojais būtina – gamintojai gauna vertingą informaciją, t. y. atgalinį ryšį iš savo vartotojų: gamintojas turi domėtis, ar jo sukurtas produktas tenkina jo vartotojus, ar produktas tenkina juos labiau negu konkurentų gaminiai.

Rinkos tyrimų pramonė nustato aukštas profesines ir etikos normas. Sėkmingas rinkos tyrimas priklauso nuo visuomenės pasitikėjimo, jog tyrimas atliekamas sąžiningai, objektyviai, kad jis paremtas pasirengimu bendradarbiauti. Profesiniai praktikos kodeksai, kurie numato rinkos tyrimo projektų atlikimo būdus, skatina šį pasitikėjimą. Pirmąjį tokį kodeksą publikavo Europos viešosios nuomonės ir rinkos tyrimų asociacija ESOMAR (angl. *European Society for Opinion and Marketing*) 1948 metais. Vėliau parengta daug kitų kodeksų, kuriuos paruošė nacionalinės rinkodaros organizacijos, tokios kaip Tarptautiniai prekybos rūmai (ICC – *International Chamber of Commerce*), atstovaujantys tarptautinei rinkodaros bendruomenei. Šiuo metu daugiau kaip 5 000

ICC/ESOMAR narių visame pasaulyje laikosi tarptautinio rinkos ir socialinių tyrimų praktikos kodekso reikalavimų.

Rinkos tyrimų procesas apima duomenų rinkimą, analizę ir interpretavimą. Rinkos tyrimų esmę sudaro specialiosios metodikos, pradedant paprasčiausiais ir baigiant naujausiais sudėtingais tyrimo metodais, paremtais sudėtingais matematiniais skaičiavimais, kurių beveik neįmanoma atlikti neturint šiuolaikinės kompiuterinės ir programinės įrangos. Susiduriama ir su statistikos programinių paketų pasirinkimo problema: jų yra daugybė – nuo paprastų iki profesionalių, nuo pigių iki kainuojančių dešimtis tūkstančių eurų, nuo įgyvendintų atviru kodu iki įgyvendintų vartotojui nesuprantama programavimo kalba.

Autoriai pateikia skaitytojams (nebūtinai tik studentams) vadovėlį, kuriame glaustai, be sudėtingos matematikos mėginama aprašyti, kaip atlikti rinkos tyrimus, kokią statistiką taikyti pateikiant ir interpretuojant gautus rezultatus.

Vadovėlyje pateikiami studentų tyrimo darbų pavyzdžiai, kuriems spręsti pasirinktas SPSS (angl. *Statistical Package for the Social Science*) programų paketas. Šio paketo licenciją turi nemažai universitetų, jį naudoja ir verslo įmonės. SPSS programinis paketas taikomas sociologijos, psichologijos, biologijos, medicinos, rinkodaros ir kokybės valdymo procesuose. Knygoje pats paketas aprašomas tiek, kiek reikia skaitytojams praktikams, t. y. vadovėlis nepretenduoja į išsamią SPSS programinio paketo apžvalgą, juo labiau į išsamų jo mokymą.

Tikimės, kad šis vadovėlis padės skaitytojams sukurti tyrimų projektą, surinkti duomenis, juos visapusiškai išanalizuoti ir parengti vertingą ataskaitą.

1. STATISTINIO TYRIMO PROJEKTAS

Atliekant statistinį tyrimą, visais atvejais būtinas tyrimo projektas. Projekte pateikiama tyrimo struktūra, kuri susieja visus tyrimo elementus. Tyrimo projektą sudaro šie pagrindiniai elementai:

- tikslas ir uždaviniai;
- metodika;
- tyrimo procedūros parinkimas;
- matavimo skalės parinkimas;
- imties sudarymo būdas ir dydis;
- biudžetas ir laiko diagrama;
- duomenų kodavimas ir analizė;
- ataskaitos parengimas.

Atsižvelgiant į tyrimo apimtis ir kitus veiksnius, tyrimo projekte turi atsispindėti trys pagrindinės statistinio tyrimo fazės: duomenų rinkimas, duomenų įvedimas ir jų analizė bei ataskaitos parengimas.

1.1. Tyrimo tikslas ir tyrimo modelis

Rinkos tyrimo planavimą reikia pradėti nuo nagrinėjamos problemos apibrėžimo ir tyrimo tikslų formulavimo. Tiksliai apibrėžti problemą daugeliu atvejų apsunkina dalykinėms situacijoms būdingas neapibrėžtumas ir problemų neaiškumas. Be to, dažnai problemų ir jų sprendimų būdų būna keletas, todėl kurį būdą tyrėjui pasirinkti yra nelengva užduotis. Taigi, prieš pradedant tyrimą, svarbiausias uždavinys – tiksliai ir aiškiai suformuluoti problemą ir nustatyti, kokia informacija būtina jai spręsti. Paprastai surinkti ir išanalizuoti duomenis dažniausiai patikima statistikams, tačiau suformuluoti problemas gali tik rinkos procesus gerai išmanantis vadybininkas.

Tyrimo tikslas – tai rezultatas, kurį norima gauti, taikant tam tikrus tyrimo metodus. Pavyzdžiui, „Nustatyti produkto, kurį pasiūlysime artimiausiu metu, paklausias ir nepaklausias savybes“. Tyrimo uždaviniuose apibrėžiama, kokius konkrečius žingsnius reikia atlikti norint pasiekti užsibrėžtą tikslą.

Teisingai suformuluotas tikslas visada prasideda veiksmažodžiu, reiškiančiu veiksmą. Apibrėžiant tyrimo tikslą, vartotini tipiniai veiksmažodžiai. Pavyzdžiui, problemą arba galimybes galima *nustatyti*, reklamos priemones arba koncepcijas galima *parinkti* ir panašiai. Tyrimo tikslas turi būti suformuluotas trumpai, aiškiai ir konkrečiai. Taip suformuluotas tikslas turi šiuos pranašumus: aiškiai matoma, ką norima pasiekti (taupomas laikas, pinigai ir kiti ištekliai); dažnai paaiškėja, kad suformuluotas tikslas atspindi tik dalį to, ką buvo norima pasiekti.

Kiekvienai tyrimo stadijai būdingi saviti tikslai. Iš 1.1 lentelės matyti, kokie yra tinkamiausi tyrimo tikslą apibūdinantys veiksmažodžiai.

1.1 lentelė. **Sprendimų priėmimų stadijos, tyrimo tikslai ir metodai**

Stadija	*Tikslas*	*Pagrindiniai tyrimo metodai*	*Pagalbiniai metodai*
Situacijos įvertinimas. Kas vyksta?	Nustatyti, išsiaiškinti, atsekti	Antriniai tyrimai, vizitai pas vartotojus	Fokus grupės, apklausos
Alternatyvų suformulavimas. Kokios galimybės?	Sukurti, ištirti, aprašyti	Fokus grupės, vizitai pas vartotojus	Antriniai tyrimai
Alternatyvos parinkimas. Kiek mes gausime? Kas geriau?	Įvertinti, testuoti, išrinkti, nustatyti prioritetus	Eksperimentai, apklausos, parinkimo modeliai, naudingumo testai	Antriniai tyrimai
Sėkmės įvertinimas. Kaip gerai mes tai atlikome?	Išmatuoti, pažingsniui praeiti	Apklausos, antriniai tyrimai	Vizitai pas vartotojus

Tyrimo tikslai gali būti:

- *žvalgybiniai*, t. y. aiškinantys problemą, o galbūt ir padedantys suformuluoti pradines hipotezes;
- *aprašomieji*, t. y. padedantys aprašyti ir išaiškinti konkrečius reiškinius, pavyzdžiui, kiek žmonių naudojasi tam tikra paslaugos rūšimi ir kiek jų yra girdėję apie užsakovo firmą;

- *eksperimentiniai*, t. y. tikrinantys hipotezes apie priežasčių ir pasekmių ryšius, pavyzdžiui, kad paslaugų perkėlimas į internetą padidintų norinčių ją gauti skaičių 10 procentų.

Kaip matyti iš 1.1 lentelės, tyrimo tikslams pasiekti reikia taikyti tam tikrus tyrimo metodus. Kiekvienas tyrimo metodas gali būti pagrindinis esant vienoms stadijoms, siekiant tam tikro tikslo, bet gali būti ir pagalbinis kitais atvejais. Remiantis šia lentele, galima greitai nustatyti, kada verta parinkti vieną ar kitą metodą. Lentelę galima panaudoti planuojant tyrimus viso sprendimo priėmimo proceso metu.

Nustačius tyrimo tikslus, svarbu apibrėžti *informacijos šaltinius*. Būtina išsiaiškinti, kokia reikalinga informacija ir kaip efektyviai rinkti duomenis.

Tyrėjas gali naudotis pirminiais arba antriniais duomenimis (žr. 1.2 lentelę) arba ir vienais, ir kitais vienu metu.

Pirminiai duomenys. Kai nėra tinkamų antrinių duomenų arba jie neišsamūs, atliekami tyrimai ir renkami pirminiai duomenys (lot. *ad hoc* – specialus, šiuo tikslu). Tai duomenys, kurie renkami konkrečiai problemai spręsti. Daugeliu rinkos tyrimų atvejų renkami išoriniai pirminiai duomenys. Deja, dažnai renkant šiuos duomenis, sugalvojama keletas klausimų ir suieškomi asmenys, kuriuos patogiausia apklausti. Taip surinkti duomenys pasirodo esantys labai mažai naudingi, o kartais net ir klaidingi. Todėl pirminiams duomenims rinkti reiktų sudaryti atskirą planą, kuriame būtų numatytas imties dydis, apklausos metodas ir planuojami statistinio tyrimo metodai.

Antriniai duomenys. Tai duomenys, kurie buvo renkami tam tikram tikslui, nesusijusiam su dabartine problema, t. y. duomenys, kurie jau buvo surinkti anksčiau ir kuriuos galima iš dalies panaudoti sprendžiant dabartinę problemą. Kadangi antriniai tyrimai remiasi jau turima informacija, jie dar vadinami *kabinetiniais tyrimais* (angl. *desk research*). Antriniai duomenys gali būti *vidiniai ir išoriniai*. Plačiausiai naudojami *išoriniai antriniai duomenys* – gyventojų surašymai ir valstybės institucijų registruojama statistika, taip pat visiems suprantama informacija, pateikiama duomenų bazėse, laikraščiuose ir žurnaluose. *Vidinius antrinius duomenis* galima rasti savo įmonės buhalterijoje, vidinio įmonės audito ataskaitose, vartotojų pateikiamuose pasiūlymuose ir skunduose.

Beveik visada antrinius duomenis naudoti yra pigiau ir rezultatas gaunamas greičiau, nei atliekant savo tyrimus. Praktiškai kiekviename

projekte pirmiausia norima sužinoti, kokie antriniai tyrimai yra prieinami ir ką jais remiantis galima išsiaiškinti.

1.2 lentelė. **Pirminių ir antrinių duomenų palyginimas**

	Pirminiai duomenys	*Antriniai duomenys*
Rinkimo tikslas	Tyrimo problemai spręsti	Kitiems tikslams
Rinkimo procesas	Reikalauja nemažai pastangų	Greitas, lengvas
Rinkimo kaina	Didelė	Nedidelė
Rinkimo laikas	Ilgas	Trumpas

Labai svarbų vaidmenį nustatant tyrimo problemą ir tikslus gali atlikti tinkamo tiriamo nagrinėjamo reiškinio, proceso modelio tipo pasirinkimas.

Modeliai iš esmės yra abstrakčios konstrukcijos, kuriomis mėginama pavaizduoti tik kai kurias realaus pasaulio savybes. Modelio sąvokos apibrėžimai pateikti ne tik moksliniuose, bet ir populiariuose šaltiniuose. Pavyzdžiui, Tarptautinių žodžių žodyne teigiama, kad „modelis – originalo atvaizdas, tapatus pasirinktu struktūros lygmeniu arba pasirinktomis funkcijomis“. Klausimas, ar modeliai yra empiriškai teisingi, neturi prasmės, nes to nė negali būti. Reikėtų kalbėti ne apie modelio empirinį teisingumą, bet apie nagrinėjamo modelio tinkamumą konkrečiu empiriniu atveju.

Dažniausiai skiriami penki modelių tipai:

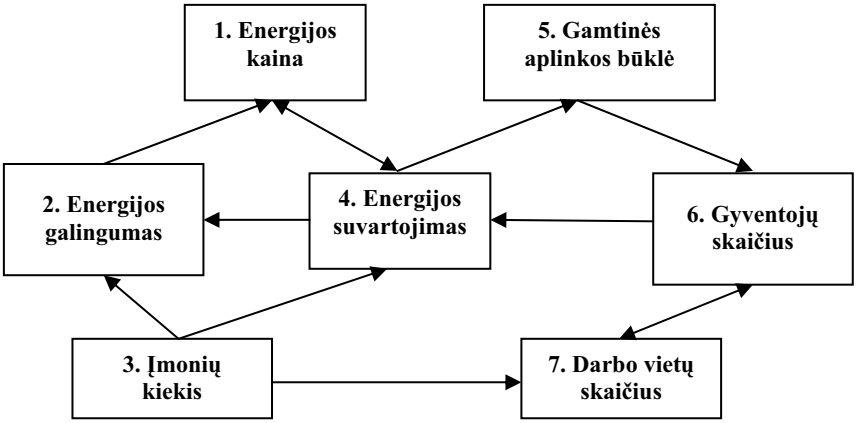
- verbalinis (žodinis);
- grafinis;
- statistinis;
- matematinis;
- konceptualus (sąvokinis).

Panagrinėkime konceptualų (sąvokinį) modelį. Konceptualus modelis – vaizdinis metodas (diagrama), vaizduojantis priežastinius ryšius tarp veiksnių (faktorių), kurie, tyrėjų manymu, yra reikšmingi nagrinėjamai problemai ar objektui. Geras modelis turėtų aiškiai ir pagrįstai susieti objektą su jį tiesiogiai veikiančiais faktoriais (tiesioginėmis grėsmėmis ar galimybėmis), netiesioginiais faktoriais ir strateginėmis veiklomis, kurios turi įtakos šiems faktoriams. Konceptualus modelis turi

būti lydimas tekstinio aprašymo, kuris sudarytą diagramą paaiškina žodžiais.

Konceptualaus modelio tikslas – atskleisti, kaip įvairios veiklos susijusios viena su kita, kaip jos logiškai yra išdėstytos ir sutvarkytos. Konceptualiuose modeliuose konceptams (abstrakčioms idėjoms ar bendriems principams, išreikštiems kaip trumpos frazės ar tekstas) jungti tarpusavyje taikomas kognityvinių žemėlapių (angl. *cognitive map* – trimatis „smegenų erdvės“ žemėlapis) metodas.

Kognityvinis žemėlapis kuriamas specialistų, gerai suprantančių tiriamą problemą. Pirmajame etape jie suformuluoja konceptų aibę su frazėmis, iš kurių reikia išrinkti svarbiausius veiksnius. Antrajame tyrimo etape nustatomi šių kintamųjų ryšiai. Trečiajame etape aptariama šių ryšių prigimtis ir detalės.



1.1 pav. **Elektros energijos suvartojimo analizės kognityvinis modelis**

Statistinio tyrimo procesas tampa suprantamesnis ir lengvesnis, kai pradedama nuo konceptualaus modelio, kuris vartoja žmonėms lengvai suvokiamas sąvokas (koncepcijas) ir kalbą.

1.2. Kokybiniai ir kiekybiniai tyrimai. Tyrimo tipo pasirinkimas

Visi metodai, taikomi tyrimo procesui, skirstomi į dvi dideles klases: *kiekybinius* ir *kokybinius*. *Kiekybiniai* metodai taip vadinami todėl, kad jų tikslas – gauti kiekybinę informaciją apie didelį tyrimo objektų (vartotojų, įmonių ar tiekėjų) skaičių. Tokie tyrimai atliekami, kai reikia gauti tikslius, statistiškai patikimus skaitinius duomenis. Kiekybiniai tyrimai susiduria su išmatuojamais (objektyviais) dalykais, skaičiais ir logika. Remiantis moksliniais metodais sudaroma imtis, žmonės atsako į klausimus ir nustatomas jų atsakymų dažnis bei kitos statistinės atsakymų charakteristikos. Jeigu imtis statistiškai patikima, iš jos gauti rezultatai taikomi visai populiacijai. Tradicinis empirinis kiekybinis tyrimų metodas yra *apklausa*. Jos tikslas – iš nedidelės populiacijos dalies (imties) gauti išvadas apie visos populiacijos elgesį. Visa populiacija dažniausiai tiriama atliekant gyventojų surašymą.

Ilgą laiką tyrėjai praktikavosi tik analizuojant kiekybinius duomenis. Taip atsitiko todėl, kad socialiniai mokslai (psichologija, ekonomika, sociologija) savo tyrimo metodus perėmė iš fizinių mokslų, skirtų fizinių objektų analizei ir nepritaikytų sudėtingam žmonių elgesiui vertinti.

Tačiau kiekybiniai tyrimai nepritaikyti paaiškinti žmonių elgesio, jų skonio, suvokimo ir produkto vertinimo. Tiems trūkumams kompensuoti buvo išplėtoti *kokybiniai* tyrimo metodai. Kokybiniai metodai nagrinėja žodžius, paveikslus ir įsivaizduojamus (subjektyvius) dalykus.

Dažniausiai kokybiniai tyrimai yra kiekybinių tyrimų pradžia. Jie taikomi problemai apibrėžti, pradinėms hipotezėms formuluoti, lemiamiems faktoriams išskirti ir kiekybinio tyrimo planui sudaryti. Kokybiniai tyrimai yra greiti ir nebrangūs. Atliekant šiuos tyrimus apsiribojama mažu respondentų skaičiumi, todėl tyrimo metodų rezultatai negali būti apibendrinti visai populiacijai. Kokybiniai tyrimai suteikia daug vertingos informacijos, padedančios išaiškinti ir tiksliai apibrėžti problemas, todėl juos taiko daugelis tyrėjų.

Kokybinius tyrimo metodus rekomenduojama taikyti tada, kai būtina:

- sudaryti giluminį vartojimo modelį, suprasti vartotojo elgseną ir veiksnius, nulemiančius vartotojo pasirinkimą;
- suvokti vartotojo sprendimo pirkti priėmimo procesą;
- aprašyti vartotojo požiūrį į produktą, kompaniją ir prekės ženklą;
- įvertinti pasitenkinimo esamais produktais lygį.

Labai svarbus yra kokybinių tyrimų vaidmuo kuriant naujus produktus. Šiuo atveju kokybiniai tyrimai leidžia:

- suprasti, ar rinkoje yra niša naujam produktui;
- įvertinti naujo produkto prekinį ženklą (brendą): pavadinimą, pakuotę, simbolį ir kita;
- suprasti požiūrį į naują produktą;
- nustatyti ir patikslinti pozicionavimo strategiją.

Kokybiniai tyrimai apima labai skirtingas tyrimo metodikas, kurios skirstomos į keletą didelių grupių:

- **Grupinės diskusijos (fokus grupės)**. Fokus grupių metodas naudingas tyrimo pradžioje, kai tiriama situacija ir siūlomos idėjos. Fokus grupes dažniausiai sudaro 6–12 asmenų grupės. Pašnekesys trunka 1,5–2 valandas. Jis dažnai vyksta specialioje patalpoje, kuri įrengta taip, kad ją galima stebėti iš vienpusio veidrodžio ir naudoti garso bei vaizdo užrašus. Grupinei diskusijai pagal sudarytą planą vadovauja patyręs vadovas. Dalyviai vienas su kitu susiduria tiesiogiai. Grupinės diskusijos metu tarp nepažįstamų dalyvių vyksta du pagrindiniai procesai: grupės jungimasis ir skirstymasis. Pradžioje respondentai stengiasi sukurti vieną grupę, bet pamažu jį suskyla į atskiras grupeles. Šios dvi priešingos tendencijos rinkodaros tyrimuose duoda labai efektyvius rezultatus. Jungimosi etape išryškėja bendros produkto savybės. Skilimo etape atsiskleidžia produkto trūkumai ir skirtumai nuo kitų produktų.
- **Giluminiai (asmeniniai) interviu**. Tai ilgai trunkantis interviu, kai respondentui pateikiami atvirojo tipo klausimai. Dažniausiai pokalbis vyksta „vienas su vienu“, tačiau gali vykti ir triadomis (vienas klausėjas – du respondentai). Galima sakyti, kad giluminio interviu metodas yra priešingas anketavimui: respondentui nepateikiami klausimai, o leidžiama laisvai pasisakyti, netaupant laiko. Tam reikalingas specialiai apmokytas apklausos tyrėjas, žinantis, ką jis gali daryti ir ko jokiu būdu daryti negalima. Idealiu atveju respondentas iš viso neturi žinoti, ką būtent norima išsiaiškinti, apie kokią prekę (prekės ženklą ar firmą) kalbama.

- **Pusiau struktūrizuoti (formalizuoti) interviu**. Kitaip nei giluminis interviu, šiam metodui būdinga vidinė struktūra, tačiau respondentui leidžiama netrukdomai reikšti savo mintis. Pusiau struktūrizuotas interviu taikomas apklausiant specialistus, nes jie nelinkę kalbėti nežinoma tema. Kaip ir giluminiame interviu, apklausos tyrėjas privalo laikytis taip, kad neturėtų jokios įtakos respondentui. Interviu metu naudojamos tarsi „špargalkos“, kur surašomos pagrindinės temos, etapai, momentai, kad nebūtų praleista kokia nors svarbi problema. Pusiau struktūrizuotas interviu dažniausiai trunka trumpiau nei giluminis interviu, nes pokalbis vyksta laikantis iš anksto sudaryto plano. Čia ypač svarbu stengtis, kad pagalbinės apklausos tyrėjo frazės nebūtų panašios į atsakymus.
- **Stebėjimai**. Stebėjimas – tai toks pradinių duomenų rinkimo būdas, kai tyrėjas tiesiogiai stebi žmones ir aplinką. Stebėjimai vykdomi įvairiais tikslais. Jie gali tarnauti kaip informacijos šaltinis formuluojant pradines hipotezes, tikrinant kitais metodais tiriamą objektą. Tyrėjas gali apsistoti vietose, susijusiose su nagrinėjama problema. Pavyzdžiui, stebėti gyventojų elgesį vykdomų akcijų metu. Stebėjimas – daug laiko reikalaujantis metodas. Stebėjimo rezultatų aprašo įforminimas kartais užima dukart daugiau laiko, nei pats stebėjimas. Pagal supančią aplinką skiriami lauko sąlygomis vykdomi stebėjimai, kai procesai stebimi natūralioje aplinkoje (parduotuvėje, prie vitrinų ar lauko skydų), ir laboratoriniai, kai vykdomi dirbtinai sukurtoje aplinkoje. Stebėjimų rezultatai fiksuojami garso ar vaizdo technikos priemonėmis.
- **Projekciniai metodai**. Projekciniai metodai perimti iš klinikinės psichologijos arsenalo. Šie metodai paremti projekcijos idėja, t. y. priskyrimu kitiems idėjų, minčių, nuomonių, norų, motyvų ar ribojančių veiksnių, kurių žmogus nemano esant savais ir nuo kurių nesąmoningai atsiriboja. Apklausos metu prašoma paaiškinti ar išsakyti savo nuomonę apie nagrinėjamą situaciją. Projekciniai metodai gali būti kelių pavidalų:

Tematinis apercepcinis testas. Vietoj paprasto klausimo respondentui pateikiama viena ar kelios situacijos ir prašoma jas pakomentuoti. Pavyzdžiui, žinoma, kad pirkėjas, vertindamas produkto kokybę, dažnai turi omenyje jo kainą, nors žodžiais tvirtina priešingai.

Sakinio užbaigimas. Taikant šį metodą respondentui pateikiama nežinomos frazės, turinčios tiesioginį ryšį su nagrinėjama tema, pradžia ir prašoma pabaigti sakinį. Šis metodas dažnai taikomas tiriant prekės išvaizdą.

Piešinio užbaigimas. Respondentui duodamas piešinėlis, panašus į komiksą, kuriame vaizduojamas personažų monologas arba dialogas nagrinėjama tema. Respondentas turi užpildyti tuščias vietas, paliktas veikėjų pasisakymams.

- **Asociatyviniai metodai**. Bendras šių metodų tikslas – nustatyti, kokias asociacijas sukelia prekė, prekinis ženklas, kompanija ir t. t. Iš asociatyvinių metodų dažniausiai taikomi šie:

Laisvoji asociacija. Respondentui siūloma laisvai ir spontaniškai išreikšti žodžiais, kokias idėjas sukelia jam vienas arba kitas daiktas ar problema.

Žodinė asociacija. Respondentui pateikiama daug požymių (pavyzdžiui, atsparus, greitai dūžtantis, minkštas, greitas, lėtas) ir siūloma susieti kiekvieną iš jų su tiriama preke, prekiniu ženklu ir panašiai.

Vaizdinė asociacija. Šis metodas gali būti taikomas visoms respondentų grupėms, neišskiriant vaikų ir užsieniečių. Tiriant šiuo metodu, nekalbant rodomas tiriamas objektas (prekė, modelis ar prekinis ženklas), o paskui lengvai atpažįstami stereotipiniai pirkėjų vaizdai (darbininkas, ūkininkas, sekretorė, sportininkas, tarnautojas, verslininkas ar namų šeimininkė). Paskutinis etapas – sujungti šiuos vaizdus su tiriamu daiktu.

Kiniško portreto tipo asociacija. Šis metodas taikomas kūrybinėms grupėms, bet tinka ir rinkos tyrimams tais atvejais, kai tiriamas prekės, prekinio ženklo ar kompanijos įspūdis. Respondento prašoma nustatyti objektą su gyvūnu, mitologiniu personažu, gėle ir t. t. Tuo tikslu dažnai vartojami šio tipo sakiniai: „Jeigu (prekės, modelio ar prekinio ženklo pavadinimas) būtų gyvūnas, tai šis gyvūnas būtų…“, „Jeigu (prekės, modelio ar prekinio ženklo pavadinimas) būtų mitologinis personažas, tai šis mitologinis personažas būtų…“

- **Eksperimentai**. Eksperimento tikslas – patikrinti kokiomis sąlygomis dominantis stimulas sukelia pageidautiną reakciją. Eksperimentas retai kada taikomas tyrimo pradžioje. Dažniausiai juo tikslinamas galimybių pasirinkimas. Pavyzdžiui, tipiniai eksperimento metodikos taikymo atvejai yra tinkamos reklamos varianto ar produkto dizaino parinkimas. Tuo tikslu testuojamos atskiros gyventojų grupės.

Laboratoriniai eksperimentai vykdomi nenatūraliomis sąlygomis, kurios padeda išvengti šalutinių veiksnių įtakos. Pavyzdžiui, vertinant pirkėjų reakciją į įvairius reklamos tipus galima pakviesti tam tikros lyties, amžiaus ar socialinės padėties pirkėjus. Pastaraisiais metais labora-

toriniams eksperimentams atlikti vis plačiau naudojami kompiuteriai. Sukurtos kompiuterių programos, kurios leidžia vartotojui „vaikščioti“ po parduotuvę ir rinktis prekes. Kompiuteryje sumontuota technika ne tik registruoja pirkinius, bet ir matuoja žmonių reakciją į atskirus elementus: kainą, dizaino spalvą ir formą, prekių išdėstymą.

Lauko eksperimentai atliekami natūraliomis sąlygomis: parduotuvėse, vartotojų namuose. Nors šių eksperimentų rezultatai turėtų būti patikimesni, nei laboratorinių, jiems didelę įtaką turi šalutiniai faktoriai, kurie ilgiau užtrunka ir brangiau kainuoja. Lauko sąlygomis atliekami įvairių tipų eksperimentai vadinami „bandomąja rinkodara“ ar rinkos testavimu. Firmos lauko eksperimentus gali atlikti iškart keliose parduotuvėse ar net skirtinguose miestuose.

- **Ekspertinis vertinimas**. Taikant ekspertinio vertinimo metodą atliekami šie žingsniai:

1. sudaromas ekspertinio tyrimo planas;
2. parenkamas ekspertų apklausos metodas ir parengiami būtini dokumentai;
3. atrenkami ekspertai;
4. atliekama ekspertinio vertinimo procedūra;
5. analizuojami ekspertinio vertinimo duomenys, pašalinamos klaidos ir prieštaravimai;
6. interpretuojami rezultatai;
7. pateikiami oficialūs ekspertinio vertinimo rezultatai ir išvados.

Rinkos tyrimuose taikomi ir kiekybiniai, ir kokybiniai tyrimo metodai, taip pat jų deriniai. Šie tyrimo metodai dažnai papildo vienas kitą.

1.3. Anketinės apklausos metodas

Apklausą atliekantys tyrėjai skirstomi į dvi grupes: atsakingieji tyrėjai ir tyrimo vykdytojai arba apklausėjai.

Apklausai atlikti taikomos trys pagrindinės formos:

- pašnekesys;
- interviu;
- anketavimas.

Dažniausiai taikomos kelios anketinės apklausos rūšys:

- anketinė apklausa el. paštu arba paštu;
- apklausa telefonu;
- apklausa, naudojant masines informacijos priemones;
- ekspertų apklausa, kai atrenkama žmonių grupė (ekspertai), turinti kurios nors srities žinių. Ekspertai analizuoja problemą ir kiekybiškai arba kokybiškai ją įvertina.

Dažniausiai (bet ne visada!) anketose būna šios dalys: įvadinė, padėties, pagrindinė ir baigiamoji.

Įvadinėje anketos dalyje formuluojamas kreipimasis į respondentą, pateikiamos apklausos sąlygos ir atsakymų teikimo taisyklės.

Padėties dalį sudaro klausimai apie socialinę demografinę respondento padėtį. Šie klausimai skirti analizės rezultatų interpretavimui palengvinti.

Pagrindinę dalį sudaro klausimai apie nagrinėjamą temą.

Baigiamojoje dalyje respondentui gali būti suteikta galimybė laisvai pasisakyti apie tiriamą problemą.

Padėties ir baigiamoji dalys nėra privalomos ir jos gali būti praleistos. Įvadinė dalis gali būti pakeista žodiniu paaiškinimu.

Pagrindinės anketos dalies klausimai skirstomi į tris grupes:

- dichotominiai (klausimai, į kuriuos galima atsakyti tik taip arba ne. Kartais jie patikslinami papildomu klausimu: „jei taip, tai kodėl?“);
- atvirieji (numato laisvos formos atsakymą. Pateikiant šio tipo klausimus galima gauti visą informaciją, tačiau atsakymai sunkiai analizuojami ir pateikiami);
- uždarieji (numato daug parengtų atsakymų. Išrenkamas tas, kuris atitinka respondento nuomonę. Rekomenduotini ne mažiau kaip keturi atsakymų variantai).

Jeigu viešosios nuomonės tyrimai yra skelbiami viešai, tai prie jų turi būti pridėta oficiali tyrimo suvestinė, o joje nurodyta:

- tyrimą atlikusios organizacijos pavadinimas;
- medžiagos rinkimo laikas;
- apibrėžta tiriamoji populiacija;
- taikytas atrankos metodas, imties dydis ir apklausos vieta. Taip pat turi būti nurodyta, kuris regionas buvo tiriamas, kaip jis buvo suskirstytas smulkiau ir kaip imtis išdėstyta geografiškai;

- apklausos būdas ir sutikusiųjų dalyvauti tyrime respondentų procentas. Jei šis rodiklis yra nedidelis, reiktų išsiaiškinti nenoro dalyvauti tyrime priežastis;
- norint išvengti galimų nesusipratimų, turėtų būti paskelbtas teiktų klausimų sąrašas;
- jeigu pateikiami rezultatai labai skiriasi nuo pradinių surinktų duomenų, turi būti nurodyta, ar buvo taikyti statistiniai svorių metodai bei kitos statistinės procedūros.

Be abejo, iškyla apklausų patikimumo problema. Yra trys svarbiausi veiksniai, kurie gali sukelti sistemines paklaidas:

- praktikoje nepatikrintos apklausos procedūros;
- nepakankamos pastangos įtikinti respondentus dalyvauti apklausoje ir todėl didelis atsisakiusiųjų dalyvauti apklausose procentas;
- nerūpestingai atlikta apklausa ir nepakankama kokybės kontrolė.

1.3.1. Kintamųjų tipai ir matavimo skalės pasirinkimas

Požymiai – grupavimo pagrindas duomenų visumai suskirstyti į pogrupius. Jų parinkimas priklauso nuo grupavimo tikslo ir preliminarinės analizės. Išskiriami kokybiniai (atributiniai) ir kiekybiniai požymiai.

Parinkus duomenų visumos nariams tam tikrus grupavimo požymius, juos tenka matuoti. Kiekvieno požymio matuojamos reikšmės sudaro duomenų aibę. Tuos požymius vadinsime *kintamaisiais*. Duomenų visumos ar dalies aibė – tai atitinkamų kintamųjų reikšmių aibė. Kintamieji kaip ir požymiai skirstomi į *kokybinius* ir *kiekybinius*.

Su kintamųjų tipu siejasi matavimo skalės. Kokybinių kintamųjų reikšmės gali būti matuojamos *nominaliąja* (*pavadinimų*) arba *rangų* skale, kiekybinių – *metrine* (intervalų ar santykių) skale. Matavimų skalė lemia duomenų analizės metodo parinkimą.

Pavadinimų (*nominalioji*) *skalė* (angl. *nominal measurement scale*). Ši skalė taikoma tada, kai imties objektus galima tik klasifikuoti, t. y. priskirti vienai ar kitai grupei. Imties nariams arba jų grupėms priskiriami kodai arba kiti simboliai. Grupės numeruojamos laisvai, o jų numeriai turi ne skaičių, bet tam tikrų grupių (vardų) prasmę.

Kintamieji, kurie matuojami pavadinimų skale, vadinami *nominaliaisiais kintamaisiais*. Pavyzdžiui, lytis, socialinis statusas, tautybė,

kraujo grupė, studentų numeris mokomųjų dalykų žurnale, komandos žaidėjo numeris, telefono numeris, akademinės grupės numeris ir t. t.

Nominalieji kintamieji tarpusavyje nepalyginami, o jų reikšmių aritmetinis vidurkis neturi prasmės.

Dichotominė matavimų skalė (angl. *dichotomous measurement scale*). Taikant šią skalę, požymių reikšmės yra priskiriamos prie vienos iš dviejų galimų kategorijų. Pavyzdžiui, gyvas arba miręs, rūko – nerūko ir kita.

Rangų (*tvarkos*) *skalė* (angl. *ordinal measurement scale*). Ši skalė taikoma tada, kai galima nustatyti objektų tiriamojo požymio skirtumus ir pagal tai išdėstyti objektus į eilę. Kintamieji, matuojami rangų skale, vadinami *ranginiais kintamaisiais*. Pavyzdžiui, konkurso dalyviams pagal pasiektus rezultatus skiriamos vietos, pedagoginiai ir moksliniai vardai (dėstytojas, docentas, daktaras, habilituotas daktaras ir t. t.). Prie klasikinių rangų skalės kintamųjų priskiriami kintamieji, gauti siejant duomenis į klases, pavyzdžiui, pajamas skirstant intervalais: iki 500 Lt; nuo 501 iki 1 000 Lt; nuo 1 001 iki 2 000 Lt; daugiau kaip 2 000 Lt.

Rangų skale matuojami dydžiai gali būti lyginami tik eiliškumui nustatyti. Tai nėra kiekybinės charakteristikos: pirmą vietą krose užėmęs sportininkas nebėga tris kartus greičiau už trečioje vietoje likusį. Taikant šią skalę, galima teigti, kad vieno objekto požymio reikšmė didesnė, lygi arba mažesnė už kito objekto, nors neįmanoma nustatyti, kiek ji didesnė ar mažesnė. Idealioje rangų skalėje sutampančių verčių neturėtų būti. Praktikoje sutampančių rangų būna. Tokiais atvejais, taikant ranginius metodus, neretai tenka naudoti specialias pataisas. Gerokai apytikslesnis šios skalės variantas – balų matavimo skalė. Pavyzdžiui, žinių įvertinimas balais, ligos stadija, dangos korozinis pažeistumas ir panašiai. Joje sutampančių verčių būna daug.

Jei rangų skalę sudaro 15 ir daugiau pozicijų, kintamąjį galima nagrinėti kaip kiekybinį.

Ranginiai ir nominalieji kintamieji dar vadinami *kategoriniais*.

Intervalų skalė (angl. *interval measurement scale*). Ši skalė taikoma tada, kai imties objektai yra skaitiniai ir juos galima įvertinti kiekybiškai.

Nulinis taškas ir intervalo ilgis yra laisvai parenkami. Dviejų šios skalės intervalų santykis nepriklauso nei nuo numatomų vienetų, nei nuo nulinio taško.

Intervalinių kintamųjų pavyzdžiai: temperatūra, kalendorinis laikas (nulinis taškas – Kristaus gimimas, Romos įkūrimas, laiko intervalai pagal Saulę, Mėnulį ir kita), gabumų, intelekto testų rezultatai.

Intervalų skalėje nurodomi kiekybiniai kintamųjų reikšmių skirtumai, išreikšti matavimo vienetais (metrais, sekundėmis, laipsniais, taškais ir panašiai). Šie skirtumai gali būti tarp atskirų intervalų arba nuo kurio nors pasirinkto atskaitos taško, t. y. nulinė reikšmė dar nereiškia, kad tiriamasis požymis visai nepasireiškia, o tiktai, kad jis nesiskiria nuo sąlyginio atskaitos nulio. Duomenis intervalų skalėje (intervalinius kintamuosius) galima apdoroti visais be apribojimų statistikos metodais.

Santykių skalė (angl. *ratio measurement scale*). Ši skalė skiriasi nuo intervalų skalės tik tuo, kad joje yra apibrėžtas nulinis taškas (ataskaitos pradžia). Santykių skalės kintamųjų pavyzdžiai: kaina, laikas, svoris, amžius ir kita. Skaičių, gautų matuojant požymius, santykis parodo kiekybinį matuojamojo požymio santykį, kuris nepriklauso nuo matavimo vienetų.

Atliekant statistinę analizę, skirtumai tarp intervalų ir santykių skalės kintamųjų paprastai laikomi nereikšmingi ir šių kintamųjų analizei taikomi tie patys metodai. SPSS programinio paketo taikymo praktikoje skirtumas tarp kintamųjų, matuojamų pagal intervalų arba santykių skalę, yra neesminis. Todėl toliau nagrinėsime matuojamus pagal intervalų skalę (intervalinius) kintamuosius.

Teorijoje pateikiami keturi pagrindiniai matavimo skalių tipai:

- lygių intervalų arba Terstouno (*Thurstone*) skalė;
- suminė arba Likerto skalė;
- kaupiamoji arba Gutmano skalė;
- semantinė diferencialinė skalė.

Šias matavimo skales panagrinėsime išsamiau.

Lygių intervalų arba Terstouno skalė. 1929 m. L. L. Terstounas (*Thurstone*) pasiūlė rangų skalės intervalus, padalytus lygiais intervalais, vertinti kaip iš tiesų lygius intervalus ir priskirti juos ne kokybinei rangų skalei, o kiekybinei intervalų skalei.

Ši skalė išliko kaip metodologinis pavyzdys ir turi teorinę vertę, bet tyrimuose taikoma retai.

Sudarant lygių intervalų skalę, kaip ir daugeliu kitų atvejų, pradedama nuo didelės aibės formuluočių (teiginių) apie tiriamą požymį. Šios formuluotės turi būti panašios gramatinės struktūros. Kiekviena formuluotė gali būti teiginys, su kuriuo galima sutikti arba nesutikti. Pavyz-

džiui, nagrinėjant, kokios savybės būtinos, norint tapti lyderiu, galima formuluoti šias savybes (teiginius):

- bendras kultūros lygis;
- sugebėjimas mokytis;
- logika;
- savikritika;
- atsakomybė;
- energija ir aktyvumas;
- tikslo siekimas ir kryptingumas;
- savitvarda;
- ištvermingumas;
- asmenybės branda;
- padorumas;
- sugebėjimas bendrauti;
- pakantumas kitų nuomonei;
- lankstus elgesys;
- sugebėjimas sukurti palankų įvaizdį.

Svarbiausia Terstouno technikos procedūra yra teiginių ekspertinis išdėstymas pagal tai, ar teiginys tiriamo požymio atžvilgiu yra palankus ar nepalankus. Pavyzdžiui, teiginys „Aš nenoriu, kad gėjai gyventų mano kaimynystėje“ kintamojo „tolerancija“ atžvilgiu būtų žemai.

Išdėstymui pateikiamos kelių ekspertų nuomonės. Kiekvienas ekspertas teiginius išdėsto pagal eilės numerius. Formuojant skalę, pirmenybė teikiama tiems teiginiams, dėl kurių ekspertų nuomonės sutapo.

Suminė arba Likerto skalė. Šią skalę sukūrė R. Likertas. Tęsdamas L. L. Terstouno darbus, 1932 m. jis pasiūlė metodą kaip išreikšti skalę taip, kad vieno klausimo taškus būtų galima sudėti su kitų klausimų taškais ir gauti bendrus (suminius) taškus.

Likertas sudarė standartinę 5 taškų skalę pavartodamas būdvardžius (vadinamuosius inkarus – angl. *anchors*), įgyjančius reikšmes nuo „visai sutinku“ iki „visai nesutinku“ su atitinkamomis skaitinėmis reikšmėmis nuo 5 iki 1.

Likerto skalė – viena iš dažniausiai naudojamų priverstinio pasirinkimo (angl. *forced-choice*) skalių. Ji yra plačiai taikoma apklausose, ypač matuojant nuostatas, įsitikinimus ir nuomones.

Pagrindinės skalės kūrimo idėjos yra šios:

a) nagrinėjamas klausimas pateikiamas kaip konstatuojamasis sakinys;
b) tada pateikiama keletas atsakymo variantų arba pasirinkimų, kurie galėtų parodyti sutikimo ar pritarimo šiam sakiniui mastą.

Pagal šią skalę respondentai išreiškia savo požiūrį į teiginius pasirinkdami vieną iš jiems pateiktų (iš anksto parengtų) skalės reikšmių.

Likerto pasiūlyta skalė tapo dažniausiai naudojama (ir plūstama) matavimo skale.

Pagal Likerto reikalavimus sudaryta 5 taškų matavimo skalė privalo turėti du teigiamus taškus, neutralųjį tašką ir du neigiamus taškus.

Standartinė 5 taškų Likerto skalė reiškia:

5 = griežtai teigiamas požiūris;
4 = šiek tiek teigiamas požiūris;
3 = neapsisprendęs (neutralus);
2 = šiek tiek neigiamas požiūris;
1 = griežtai neigiamas požiūris.

Likerto skalės „klausimai“ yra pateikiami tvirtinimo forma, su kuriais respondentas gali sutikti arba nesutikti.

Paklausus dešimt klausimų galima gauti didžiausią suminį balą 50 (kai pasirinkti 5 kiekvieno punkto taškai). Taigi, nagrinėjant nuomonę naudojant klausimyną, šiuo atveju 50 reikštų didžiausią galimą nuomonių sumą, o taškai mažiau nei 50 atitiktų mažesnes nuomonių sumas.

1 pavyzdys. Teiginys. Reguliari techninė priežiūra yra automobilio saugaus eksploatavimo garantas.

1 – Visiškai nesutinku.
2 – Nuosaikiai nesutinku.
3 – Vienodai sutinku ir nesutinku.
4 – Nuosaikiai sutinku.
5 – Visiškai sutinku.

2 pavyzdys. Teiginys. Nuo plaukų spalvos priklauso žmogaus charakteris.

1 – Visiškai nesutinku.
2 – Nuosaikiai nesutinku.
3 – Vienodai sutinku ir nesutinku.
4 – Nuosaikiai sutinku.
5 – Visiškai sutinku.

Kartais mėginama sudaryti Likerto skalę kaip klausimą. Pavyzdžiui, „Ar nuo plaukų spalvos priklauso žmogaus charakteris?“ Tačiau negalima „sutikti“ ar „nesutikti“ su klausimu! Siekiant išmatuoti sutikimo ar nesutikimo intensyvumą, pateikiamame dalyke turi būti kažkas, į ką reikėtų sureaguoti, būtent, deklaratyvus teiginys.

Kyla klausimas, kiek atsakymų variantų reikia pateikti? Dažniausiai naudojama 5 ar 6 pasirinkimo alternatyvų skalė. Nėra jokios griežtos taisyklės, ar turi būti lyginis ar nelyginis pasirinkimo alternatyvų skaičius. Sprendimas priklauso nuo to, kas yra tiriama, taip pat nuo ankstesnių šios srities darbų apimties. Tai viena iš daugelio problemų, kuri gali būti patikrinta atliekant bandomąjį tyrimą.

Visada iškyla viduriniojo taško problema. Klausimas, ar taškas turi būti vidurinysis, neutralusis ar „balansuojantis“? Pirmiausia reikėtų nuspręsti, kaip šis vidurinysis taškas bus pavadintas. Galimi šie viduriniojo taško variantai:

a) neutralu;

b) nei sutinku, nei nesutinku (skatina negalvoti);

c) vienodai sutinku ir nesutinku.

Kai kas pasakytų, kad visi šie variantai yra lygiaverčiai. Tačiau, šio tipo literatūra teigia, kad jie turi subtilius, bet aiškius skirtumus, kuriuos turėtų sukelti respondentų galvose. Pavyzdžiui, De Velis (1991) teigia, kad:

variantas b) leidžia numanyti abejingą nesidomėjimą, o kiti variantai reiškia griežtą, net lygų potraukį ir sutikti, ir nesutikti;

variantai a) ir c) sukelia aktyvų mąstymą, o b) turėtų patikti tam, kas visai nenori galvoti apie pateikiamą problemą.

Viduriniojo taško problemos esmę tikriausiai būtų galima apibrėžti klausimu: „Ar viduriniojo taško pateikimas nesudarys sąlygų respondentui „išsisukti“?“ Tuo tarpu tyrėjai visada nori žinoti požiūrių intensyvumą.

Atlikus išsamius tyrimus, buvo nustatyta:

- kai pateikiamas vidurinysis taškas, jį pasirenka apie 20 procentų apklaustųjų. Tačiau, jei šio varianto nėra, jo neprašoma ir nepasigendama;
- viduriniojo taško buvimas/nebuvimas turi labai nedidelį poveikį kitų kategorijų atsakymų pasiskirstymui.

Gutmano skalė. Matuojant Gutmano skale, teiginiai išdėstomi tokia tvarka, kad respondentui sutikus su bet kuriuo iš eilėje išdėstytų teiginių, logiškai jis turėtų sutikti ir su visais prieš tai buvusiais teiginiais.

Semantinė diferencialinė skalė. Šį skalių sudarymo metodą, panašų į Likerto, sukūrė Ch. E. Osgūdas (*Osgood*) apie 1950 metus.

Semantinė diferencialinė skalė telkiasi apie tiriamą objektą (stimulą), kuriuo gali būti asmuo, reiškinys, pozicija ir panašiai.

Konstruojant semantinę diferencialinę skalę privalu:

1. Nustatyti stimulą.
2. Kitu žingsniu išvardyti kelias būdvardžių poras (inkarus). Kaip kraštinės reikšmės (inkarai) apgalvotai taikomos priešingos reikšmės (antonimai).
3. Respondentas pasirenka linijos atkarpą, kuri geriausiai atitinka jo nuomonę apie stimulą.

Semantinės diferencialinės skalės pavyzdys.

AUTOMOBILIŲ PARDAVĖJAI

Garbingi	–	–	–	–	–	–	–	–	–	Negarbingi
Ramūs	–	–	–	–	–	–	–	–	–	Triukšmingi

Šiame pavyzdyje stimulas yra automobilių pardavėjai.

Stimulas vertinamas taikant būdvardžių poras, kurios apima visą galimą nuomonių įvairovę.

Praktiškai vienos ar dviejų porų būdvardžių neužtenka, kad apibūdintume stimulą, todėl šis pavyzdys yra negalutinis.

Lieka neaišku, kiek galimų pasirinkimo linijų (brūkšnių) reikia pateikti tarp būdvardžių poros. Kaip paprastai apklausose griežtų taisyklių nėra. Dažniausiai naudojama nuo 7 iki 9 pasirinkimo linijų.

Kartais naudojama mažiau pasirinkimų ir jie numeruojami.

Šiuo atveju visiems skalės taškams iš anksto suteikiamos reikšmės: +2 +1 0 –1 –2.

Sudarant semantinę diferencialinę skalę priešingos reikšmės neturi priešdėlio „ne“, pavyzdžiui:

Naudingas				**Žalingas**
+2	+1	0	–1	–2

1.3.2. Imties sudarymo būdas ir dydis

Kasdienėje praktikoje mes susiduriame su daugybe tvirtinimų ar paneigimų, kuriuose minima tam tikra skaitinė arba kiekybinė informacija.

Pavyzdžiai:

- Daugiau kaip pusė mūsų šalies gyventojų tiki geresne Lietuvos ateitimi.
- Daugiau kaip 60 procentų Lietuvos gyventojų pasitiki spauda, o daugiau kaip 70 procentų – bažnyčia.
- X politiko reitingas per paskutiniuosius 3 mėnesius krito 10 procentų.

Tokius teiginius vadiname statistiniais teiginiais. Jie paremti duomenimis, kurie gauti statistinio stebėjimo metu. Renkant duomenis, visuomet reikia ištirti tam tikrų objektų ar individų grupę, kuri statistikoje vadinama *statistine visuma*, *generaline aibe arba populiacija*. Dažnai sudėtinga tiksliai apibūdinti populiaciją, nustatyti jos dydį, o visa tai padarius įsitikinama, kad visuma tiriamų objektų yra labai didelė. Vykdyti ištisinį tyrimą, rinkti duomenis apie kiekvieną objektą ar individą kartais yra neįmanoma arba tai reikalauja didelių laiko ir materialinių sąnaudų. Todėl dažniausiai renkami duomenys tik iš dalies visumos narių. Dalis populiacijos objektų arba individų, apie kuriuos renkame mus dominančius duomenis, vadinami *imtimi*.

Pagrindinės priežastys, dėl kurių tiriama ne visa populiacija, o atrenkama imtis:

- sutaupoma pinigų;
- sutaupoma laiko;
- kai kurie populiacijos nariai gali būti nepasiekiami;
- tiriant kai kuriuos elementus gali tekti sunaikinti;
- tikslumas – ištirti mažesnį kiekį elementų iš kruopščiai atrinktos imties galima tiksliau.

Metodas, kuriuo atrenkama tinkama imtis, turint tikslą nustatyti visai populiacijai būdingus bruožus ar ypatybes, vadinamas *atranka*.

Esminė imties sudarymo problema – imties parinkimas, kuri atstovautų visai generalinei aibei (populiacijai), būtų *reprezentatyvi*. Pagrindinė imties reprezentatyvumo užtikrinimo sąlyga yra jos *atsitiktinis* parinkimas iš populiacijos. Užtikrindami tokį imties parinkimą, mes galė-

sime, remdamiesi imties tyrimo rezultatais, daryti išvadas apie visą populiaciją.

Imtys sudaromos įvairiais būdais, tačiau pagrindiniai imčių sudarymo būdai yra du: *tikimybinis* ir *netikimybinis*. Atrenkant tikimybiniu būdu, kiekvieno elemento tikimybė patekti į imtį yra žinoma, o sudarant netikimybines imtis, daug lemia atsitiktinumas.

Trumpai apibūdinsime kiekvieną *tikimybinės imties* sudarymo atvejį.

Sudarant *paprastąją atsitiktinę* imtį (*atsitiktinės atrankos metodas*), visų populiacijos elementų galimybės priklausyti imčiai yra vienodos. Tai užtikrinama naudojant kompiuterinį atsitiktinių skaičių generatorių arba, sunumeravus visus elementus, traukiant juos burtų būdu. Šis būdas taikytinas esant nedideliam populiacijos narių skaičiui. Didelėse populiacijose priskirti kiekvienam individui žymę ir paskui atlikti atsitiktinį traukimą nėra paprasta.

Sisteminės imties sudarymo principas yra toks: atsižvelgus į populiacijos ir numatomos imties dydį, parenkamas išrinkimo žingsnis; tada populiacijos elementai išdėstomi į eilę; iš kelių pirmųjų elementų pasirenkamas vienas, o toliau pasirinktu žingsniu atrenkami visi likę elementai. Jei populiacija didelė, šis atrankos būdas artimas atsitiktinei atrankai.

Sluoksninės imties atveju visa populiacija skirstoma į sluoksnius (mažesnes aibes), kurių kiekviena yra vienalytė tiriamojo požymio atžvilgiu. Kiekvienam sluoksniui taikomas paprastosios atsitiktinės imties sudarymo metodas. Taigi galima atlikti visos populiacijos ir atskirų sluoksnių tyrimą. Teoriškai įrodoma, jog taikant sluoksninės imties sudarymo schemą, gaunami tikslesni rezultatai, nei paprastosios atsitiktinės imties atveju.

Lizdinės (*klasterinės*) imties atveju visa populiacija skirstoma pagal tam tikrą požymį į panašias grupes – lizdus (klasterius). Iš visų lizdų aibės paprastosios atsitiktinės imties būdu atrenkama dalis. Pagal tiriamąjį požymį išskirtos grupės (lizdai) turi būti kiek galima panašios vienos į kitą. Dažniausiai taikoma lizdinės imties forma – teritorinė imtis, kurioje lizdai sudaromi naudojant žemėlapius. Šiuo atveju tirti objektus yra paprasčiau ir pigiau, nes nereikia turėti informacijos apie smulkiausius imties vienetus, lengviau vykdyti duomenų rinkimą.

Visi aptartieji tikimybiniai imčių sudarymo atvejai leidžia surinktų statistinių duomenų analizei taikyti tikimybių teoriją.

Netikimybinės (*neatsitiktinės*) imtys pasižymi subjektyvumu, todėl yra mažiau reprezentatyvios. Šių imčių atveju neužtikrinamos vienodos galimybės kiekvienam populiacijos elementui patekti į imtį. Gautiems tiriamo požymio įverčiams nėra matematinio jų vertinimo pagrindimo. Pagrindinis šių imčių pranašumas – paprastas, suprantamas jų sudarymas, kuris esant ir nedideliam imties tūriui (imties elementų skaičiui), kai tyrimą vykdo gerai išmanantys tyrimo objektus asmenys, duoda gana patikimus rezultatus.

***Netikimybinių* imčių sudarymo būdai**

Ekspertinės imties atveju elementai į imtį įtraukiami atsižvelgiant į ekspertų nuomonę. Imtis pasiteisina ir duoda gerus rezultatus, kai jos apimtis nedidelė. Populiacijos vienetų savybės yra stipriai kintančios, todėl tokią imtį turi sudaryti atitinkamos srities geras specialistas.

Kvotinė atranka gaunama tada, kai populiacijos pogrupiams nustatomos jų atstovavimo kvotos. Toliau kiekvienoje kvotoje atsitiktinai pasirenkant elementus yra užtikrinamas didesnis imties reprezentatyvumas. Ši atranka panaši į sluoksninę, tačiau respondentai (kvotos) pasirenkami neatsitiktinai. Kvotos būdas naudingas, kai nereikia didelio tikslumo, o pakanka bendro vaizdo. Šis būdas gali būti taikomas greitai tiriant visuomenės nuomonę ir prekių rinką.

Proginės (*parankios*) imties atveju į imtį įtraukiami pirmieji pasitaikę populiacijos elementai. Tai mažai reprezentatyvi, pigi ir paprastai vykdoma atranka.

„*Sniego kamuolio*“ atranka yra tada, kai atsižvelgiant į gautą informaciją, keičiamas respondentų skaičius (dažniausiai didinamas).

Imtys neapima visos populiacijos, todėl jos elementams apskaičiuotas parametras skiriasi nuo visos populiacijos atitinkamo parametro. Gaunama *imties paklaida*, kurią dažniausiai lemia atrankos klaidos. Atrankos klaidų priežastys yra šios:

Atsitiktinumas (priklauso nuo imties didumo). Pavyzdžiui, išmatavus 100 moterų ūgį, visos buvo aukštesnės nei 170 cm.

Tendencingumas (sisteminės paklaidos). Jį nulemia „matavimo instrumento“ netobulumas. Pavyzdžiui, vykdant apklausą internetu, bus sužinota nuomonė tik tų respondentų, kurie turi internetą. Sistemingos paklaidos dažnai daromos numatant rinkimų rezultatus.



1.2 pav. **Imčių parinkimo atvejai**

Imties paklaidai didelį poveikį turi ir *atsakymo lygmuo* – atsakiusiųjų ir visų parinktų respondentų santykis.

Atsakymų matavimo klaidos:

Apklausos efektas. Tas pats respondentas skirtingų apklausų metu gali pateikti skirtingus atsakymus. Pavyzdžiui, kai kurie respondentai linkę „pasijauninti“, kiti nurodo daugiau metų nei turi.

Atsakančiojo (*respondento*) *efektas*. Kai kurie respondentai pateikia neteisingus atsakymus, norėdami padaryti įspūdį tyrėjams. Pavyzdžiui, ūkininkai vyras ir žmona dažnai pateikia skirtingus duomenis apie gautą derlių.

Nepasitikėjimas tyrimo tikslu. Pavyzdys: klausimas apie pajamas. Jei klausia oficialus asmuo – pajamos paprastai būna mažesnės.

Tendencinga apklausa. Ji vyksta tada, kai tyrėjai yra suinteresuoti. Pavyzdžiui, nuo trąšų gamintojų priklausantys tyrėjai gali parinkti geresnius sklypus, norėdami įrodyti trąšų poveikį derliui.

Imties dydžio nustatymas yra vienas iš svarbiausių tyrimo projekto uždavinių, kartu ir vienas iš sudėtingiausių. Pirma, reikia apibrėžti populiaciją. Antra, reikia išspręsti daug problemų: turimas tyrimo finansavimas gali turėti lemiamą įtaką imties dydžiui; imties dydis priklauso nuo numatomos analizės tipo; nuo norimo pasiekti įvertinimo tikslumo; nuo tiriamų kintamųjų skaičiaus; nuo tiriamos populiacijos vienalytiškumo ir t. t.

Matematiškai imties dydis yra funkcija, priklausanti nuo kelių argumentų: norimo pasiekti įverčių tikslumo, populiacijos sklaidos ir norimo patikimumo.

Socialiniuose ir rinkos tyrimuose paprastai taikomas 95 rečiau 99 procentų *patikimumo* (*pasikliovimo*) *lygmuo*.

Imties dydį galima skaičiuoti pagal formulę

$$n = \frac{z^2 * s^2}{\Delta^2}, \qquad (1.1)$$

čia: Δ – tyrimo tikslumas (paklaida) procentais,

z – konstanta, apibrėžiama pagal pasirinktą patikimumo lygmens reikšmę (patikimumo lygmenų konstantų z reikšmės pateiktos 1.3 lentelėje),

s – standartinis nuokrypis

$$s = \sqrt{\Delta(100 - \Delta)}\,. \qquad (1.2)$$

1.3 lentelė. **Patikimumo lygmenų konstantos**

Patikimumo lygmuo	*Konstanta z*
60 %	±0,84
70 %	±1,03
80 %	±1,29
85 %	±1,44
90 %	±1,64
95 %	±1,96
97 %	±2,18
99 %	±2,58
99,7 %	±2,58

Pavyzdys. Nustatysime imties dydį, kai norime patikrinti, ar tikrai 35 procentai asmenų padaro eismo avarijas būdami neblaivūs. Šiuo atveju dispersija:

$$s^2 = 35 * (100 - 35) = 2275. \qquad (1.3)$$

Matavimo paklaida ± 5 procentai, konstanta z = 1,96, kai patikimumo lygmuo 95 procentai. Tada

$$n = \frac{1{,}96^2 * 2275}{25} \approx 350. \qquad (1.4)$$

Šiuo atveju imties dydis yra 350 respondentų.

Imties statistinė paklaida skaičiuojama pagal formulę

$$\Delta_{x\%} = \pm z\sqrt{\frac{pq}{n}}, \qquad (1.5)$$

čia: z – statistinė konstanta atitinkamam patikimumo lygmeniui,

$p = q = 50$ % – tiriamojo požymio pasirodymo/nepasirodymo tikimybė (mūsų atveju respondento patekimas/nepatekimas į imtį; atsitiktinėms imtims tikimybė lygi ½, t. y. 50 %),

n – imties dydis – apklaustųjų skaičius.

1 pavyzdys. Apklausta 1 000 respondentų. Patikimumo lygmuo 95 procentai. Imties statistinė paklaida:

$$\Delta = \pm 1{,}96 * \sqrt{\frac{50 * 50}{1000}} = \pm 1{,}96 * \sqrt{2{,}5} = \pm 3{,}1\ \%\ . \qquad (1.6)$$

Apskaičiuota statistinė paklaida taikoma visų imties viduje naudojamų kintamųjų reikšmėms. Pavyzdžiui, jei apklausos metu nustatėme, kad 17 procentų respondentų mokosi labai gerai ir puikiai, tai ši reikšmė kinta nuo 13,9 (17 % – 3,1 %) iki 20,1 (17 % + 3,1 %) procento.

2 pavyzdys. Kai populiacijos dydis žinomas, imties dydžiui nustatyti galima taikyti formulę (Schwarze, 1993):

$$n = \frac{N * 1{,}96^2 * p * q}{\varepsilon^2 * (N - 1) + 1{,}96^2 * p * q}, \qquad (1.7)$$

čia: N – populiacijos dydis,

ε – pageidaujamas tikslumas, paprastai ε = 5 %,

$p = q = 50\ \%$ – tiriamojo požymio pasirodymo/nepasirodymo tikimybė (mūsų atveju respondento patekimas/nepatekimas į imtį; atsitiktinėms imtims tikimybė lygi ½, t. y. 50 %),

1,96 – atitinka standartizuoto normaliojo skirstinio 95 procentų patikimumo lygmenį (žr. 1.3 lentelę).

Tipinėje mokykloje mokosi apie 1 200 mokinių. Norima ištirti, kiek mokinių reikėtų apklausti, siekiant nustatyti, kiek mokinių gauna dienpinigius. Patikimumo lygmuo 95 procentai, o pageidaujamas tikslumas 5 procentai. Pagal duotą formulę randame n = 291.

3 pavyzdys. Kartais reikia nustatyti paklaidos dydį, kai imtis žinoma. Mūsų pavyzdyje reikėjo apklausti 291 mokinį, tačiau pavyko apklausti tik 200. Kokiu tikslumu pavyko apklausa?

Remiantis formule, randame

$$\varepsilon = \sqrt{\frac{N * 1{,}96^2 * p * q}{n * (N - 1) + 1{,}96^2 * p * q}} = 6{,}8\ \%. \qquad (1.8)$$

Taigi, jei nustatytume, kad 70 procentų apklaustųjų yra būdinga tam tikra savybė su 95 procentų patikimumu, galima sakyti, jog šia savybe pasižymi nuo 63,2 iki 76,8 procentų populiacijos.

Imties dydžiui skaičiuoti internete skirta daug įvairių skaičiuoklių. Jas nesunkiai galime rasti interneto naršyklėje atlikę paiešką, pavyzdžiui „imties skaičiuoklė“ arba „imties dydžio skaičiuoklė“.

1.3.3. Anketos kodavimo schemos sudarymas. Duomenų kodavimas

Apklausos anketų rezultatai tolesnei analizei atlikti turi būti įvesti į kompiuterinę duomenų bazę. Anketos klausimai ir atsakymų variantai paprastai pateikiami apibrėžiant kintamuosius duomenų bazėse (mūsų atveju SPSS programiniame pakete). Visa tai pavaizduojama anketos kodavimo schemoje, kurioje įrašyti anketos klausimai, kintamųjų vardai ir tipai, atsakymų variantai bei kodai. 1.4 lentelėje matome apklausos anketos apie kavos pirkimą kodavimo schemą.

1.4 lentelė. **Anketos kodavimo schema**

Anketos klausimas	*Kintamojo vardas duomenų bazėje*	*Kintamojo tipas*	*Atsakymo variantai*	*Atsakymo kodas*
1. Jūsų lytis?	*Lytis*	Nominalusis	Vyras Moteris	1 2
2. Jūsų amžius?	*Amžius*	Ranginis*	15–25 m. 26–35 m. 36–45 m. Daugiau kaip 45 m.	1 2 3 4
3. Jūsų mėnesinės pajamos?	*Pajamos*	Ranginis	Iki 800 Lt 801–1 000 Lt 1 001–1 500 Lt 1 501–2 000 Lt Daugiau kaip 2 000 Lt	1 2 3 4 5
4. Kokio dydžio pakuotę perkate?	*Pakuotės_dydis*	Kiekybinis	Neribojama	Papildomai koduoti nereikia

Anketos klausimas	*Kintamojo vardas duomenų bazėje*	*Kintamojo tipas*	*Atsakymo variantai*	*Atsakymo kodas*
5. Įvertinkite 10 balų skalėje šias kavos rūšis: Paulig Nescafe Jacobs Eduscho Folgers Merrild Elite Ar Mani Tshibo	*R_A_vertin* *R_B_vertin* *R_C_vertin* *R_D_vertin* *R_E_vertin* *R_F_vertin* *R_G_vertin* *R_H_vertin* *R_I_vertin*	Ranginis	Kiekvienam kintamajam nuo 1 iki 10	Papildomai koduoti nereikia
6. Kokias kavos rūšis dar žinote?	*Kitos_rūšys*	Nominalusis	Neribojama	Papildomai koduoti nereikia
7. Kur perkate kavą?	*Pirkimo_ vieta_PC* *Pirkimo_ vieta_SP* *Pirkimo_ vieta_K*	Nominalusis	Prekybos centre Specializuotoje parduotuvėje Kitur	1 2 3

Pastaba. Šioje apklausoje kintamasis *Amžius* yra ranginio tipo. Paprastai amžius vertinamas kaip santykių skalės kiekybinis kintamasis. Rangų skalė apima mažiau informacijos, nei santykių. Apklausos vykdytojai, tirdami priklausomybę nuo amžiaus grupių, turi būti atidūs, nes respondentai gali priklausyti tik vienai amžiaus grupei, pavyzdžiui, dieninių studijų pirmo kurso studentai.

Skirtingo tipo klausimai yra koduojami 3 tipų kintamaisiais:

Uždarieji klausimai su vieno varianto atsakymais koduojami vienu kintamuoju (*Lytis*, *Amžius*, *Pakuotės_dydis*). Kintamojo kitimo skalės tipas gali būti bet koks (atitinkamai pavyzdyje: vardinis, ranginis, intervalinis).

Uždarieji klausimai su daugiavariantiniais atsakymais koduojami keliais vienvariantiniais kintamaisiais (*Pirkimo_vieta_PC*, *Pirkimo_ vieta_SP*, *Pirkimo_vieta_K*). Kintamųjų skalė – tik vardinė.

Atvirieji klausimai su bet kuriuo skaičiumi atsakymų variantų koduojami vienu kintamuoju. Kintamojo skalė gali būti intervalinė (skaitmeniniams kintamiesiems *Pakuotės_dydis*) arba vardinė (nominaliniams kintamiesiems *Kitos_rūšys*).

Kiekybiniai kintamieji papildomai nekoduojami, nes jau yra išreikšti skaičiais. Tačiau, jei kintamasis yra kokybinis, prieš pradedant statistinę analizę tokį kintamąjį reikia koduoti: tekstinė informacija keičiama į skaitmeninę. Pavyzdžiui, kokybinį kintamąjį *Lytis*, kurio galimi atsakymai yra „Moteris“ arba „Vyras“ galima pakeisti skaičiais 2 ir 1, kur 2 reiškia, kad stebėjimas atitinka įrašą „Moteris“, o 1 – atitinkamai „Vyras“.

Ranginiai kintamieji išreikšti skaičiais/skaičiumi (pavyzdžiui, įvertinimas dešimtbalėje skalėje arba užimama vieta) irgi papildomai nekoduojami. Jei ranginis kintamasis turi tekstinę reikšmę („gerai“, „labai gerai“, ... arba „30–40 m.“, „41–50 m.“ ...), tai šios reikšmės koduojamos.

Kad išvengtume klaidų nustatant ranginių kintamųjų ryšį, koduojant ranginius kintamuosius, reikia, kad rangavimas būtų viena kryptimi – nuo didžiausio rango iki mažiausio, arba atvirkščiai (žr. 1.5 lentelę).

1.5 lentelė. **Ranginių kintamųjų kodavimas**

Anketos klausimas	*Kintamasis*	*Atsakymo variantai*	*Teisingas kodavimas*
Jūsų pajamos?	*Pajamos*	iki 100 Lt 101–500 Lt 501–1 000 Lt 1001–2 000 Lt	1 2 3 4
Jūsų išsilavinimas?	*Išsilavinimas*	Aukštasis Vidurinis Pradinis	3 2 1

Sunkiau teisingai atlikti kintamųjų kodavimą, kada vienam klausimui nurodoma keletas atsakymo variantų. Pavyzdžiui, „Pažymėkite svarbos tvarka (nuo 1 – svarbiausia iki 3 – mažiausiai svarbu) jogurto savybes: „Skonis“, „Kaina“ ir „Ekologiškumas“.

Šiuo atveju kiekvienas respondentas nurodo kiekvienos savybės svarbą. Atsakymai įtraukiami kaip atskiri kintamieji: *Skonis*, *Kaina*,

Ekologiškumas. Tačiau šiuo atveju rangavimo skalė atvirkštinė, nei parodyta 1.5 lentelėje. Todėl reikia atlikti perkodavimą: 1 vieta – 3, 2 vieta – 2, 3 vieta – 1.

Nominaliųjų kintamųjų kodavimo tvarka neturi reikšmės, tačiau patariama kintamuosius *taip/ne* koduoti: *taip* – 1, *ne* – 0.

Reikėtų neužmiršti, kad koduojant kokybinius kintamuosius, jiems suteikiami skaitmeniniai kodai, o ne reikšmės: 1.5 lentelėje kintamojo *Pajamos* kodai niekaip neatspindi tikrųjų pajamų reikšmių. Kodavimas reikalingas pirminės informacijos duomenų bazei (iš anketų atsakymų) sukurti, kuri įvedama į kompiuterį ir analizuojama taikant statistinius bei kognityvinius metodus.

KONTROLINIAI KLAUSIMAI

1.1. Kokie yra tyrimo projekto pagrindiniai elementai?
1.2. Kokius žinote tiriamų procesų aprašymo modelius?
1.3. Kokios dalys išskiriamos apklausos anketoje?
1.4. Nuo ko priklauso atsakymų matavimo klaidos?
1.5. Kokias kintamųjų matavimų skales žinote?
1.6. Ką vadiname populiacija ir imtimi? Ar visada populiacija ir imtis skiriasi?
1.7. Kokie pagrindiniai imčių sudarymo tikslai?
1.8. Koks patikimumo lygmuo dažniausiai taikomas socialiniuose tyrimuose?
1.9. Kaip vedant duomenis į kompiuterį koduojami kiekybiniai kintamieji?

UŽDUOTYS

1.1 užduotis
2010 m. Lietuvoje buvo užregistruota apie 370 tūkstančių bedarbių. Reikia ištirti, su kokiomis pagrindinėmis problemomis jie susiduria, norėdami gauti pašalpas. Kiek bedarbių reikėtų apklausti, siekiant turėti atsakymų 95 procentų patikimumo lygmenį ir 5 procentų tikslumą?

2. DUOMENŲ ĮVEDIMAS Į KOMPIUTERĮ

Duomenų įvedimas į kompiuterį – tai labai svarbus parengiamasis statistinio duomenų apdorojimo etapas. Jis sudarytas iš nesugadintų (užpildytų, teisingai užpildytų) anketų atrankos, sugadintų ir neatsakytų anketų dalies nustatymo, anketos kodavimo schemos sudarymo, duomenų kodavimo bei duomenų bazės (duomenų rinkmenos) kūrimo.

2.1. SPSS duomenų bazės pildymas

Duomenys į failą (rinkmeną) gali būti įvedami tiesiogiai arba įkeliami iš įvairių šaltinių. Aptarsime:

- duomenų tiesioginį įvedimą į SPSS duomenų bazę;
- duomenų įvedimą (skaitymą) iš SPSS duomenų failų;
- duomenų įvedimą (įkėlimą) iš skaičiuoklių (pavyzdžiui, Microsoft Excel) ir tekstinių failų.

SPSS duomenų redaktorius

Duomenų redaktorius sukurtas elektroninės lentelės principu: eilutėse turime stebėjimų duomenis (*Cases*), stulpeliuose – kintamuosius (*Variables*).

Spragtelėję duomenų redaktoriaus lango apačioje atitinkamą auselę galime pasirinkti rodinius:

- Duomenų peržiūra (*Date View*). Matome duomenų reikšmes arba duomenų apibūdinimo žymes.
- Kintamųjų peržiūra (*Variable View*). Matome informaciją apie kintamuosius: kintamojo ir kintamojo žymenos pavadinimą, duomenų tipą, matavimo skalę, praleistas reikšmes ir kt.

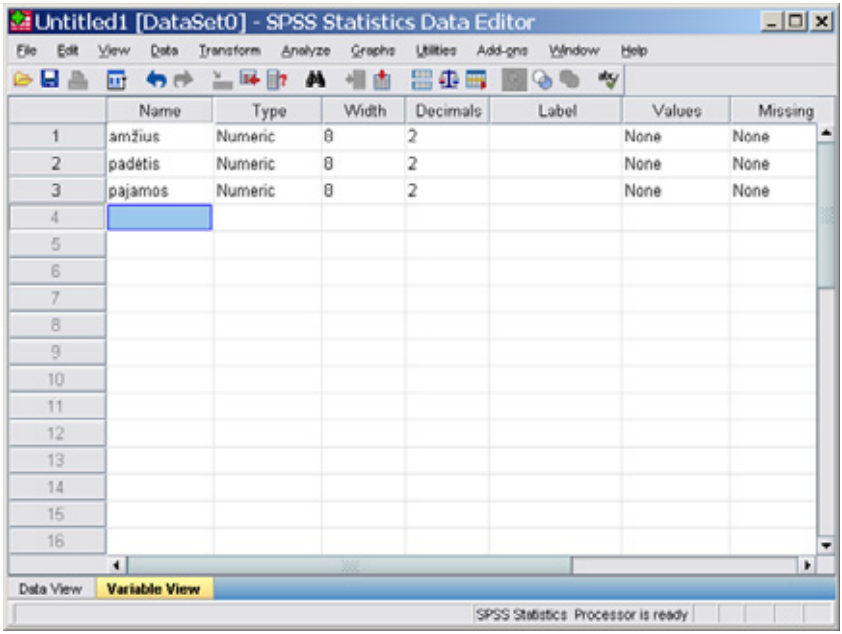
Abiejų rodinių languose rinkmenoje esančią informaciją galima papildyti, keisti arba pašalinti.

Skaitmeninių duomenų įvedimas

Nedidelius duomenų failus galima įvesti tiesiogiai į SPSS duomenų redaktorių. Taip pat SPSS duomenų redaktorius naudojamas ir didelių failų duomenims redaguoti.

- Spragtelėkite kintamųjų peržiūros *Variable View* auselę duomenų redaktoriaus lango apačioje.

Apibrėšime kintamuosius, kuriuos naudosime mūsų pavyzdyje: *amžius*, *šeiminė padėtis* ir *metinės pajamos*.



2.1 pav. **Kintamųjų vardai *Variable View* lange**

- Pirmajame stulpelyje pirmoje eilutėje įrašykite *amžius*, antroje eilutėje – *padėtis*, trečioje eilutėje – *pajamos* (žr. 2.1 pav.).

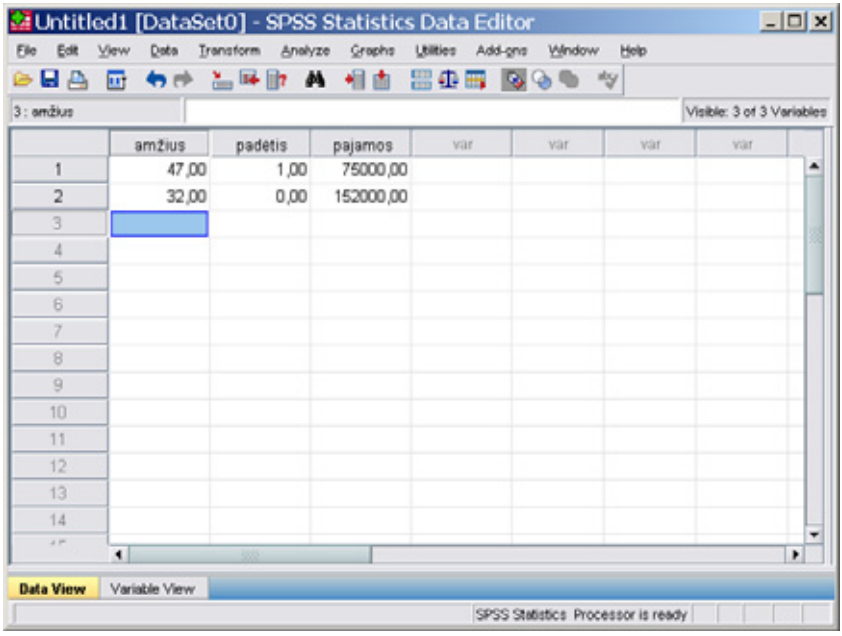
Kintamojo vardui keliami šie reikalavimai:

✓ Kintamojo vardas turi prasidėti raide. Kiti simboliai gali būti lotyniškos raidės, skaitmenys, taškas ar simboliai @, #, _, $. Neleidžiami tarpai ir specialūs simboliai, pavyzdžiui, !, ?, ‘, *.

- ✓ Paskutinis kintamojo vardo simbolis negali būti taškas arba pabraukimas.
- ✓ Didžiosios ir mažosios raidės vertinamos vienodai.
- ✓ Kintamojo vardo ilgis gali būti iki 64 simbolių.

Jei kintamiesiems vardas neįvedamas, SPSS automatiškai sukuria unikalius vardus, pavyzdžiui, *VAR00001*, *VAR00002* ir t. t. Rekomenduojama tokių kintamųjų vengti, nes jie neatspindi jokios kintamųjų prasmės.

Naujiems kintamiesiems savaime suteikiamas *Numeric* (skaitmeninių) duomenų tipas. Šis tipas gali būti keičiamas vedant, tarkim, tekstinius kintamuosius, tačiau mes vedame koduotus duomenis, todėl jų kodai turi skaitmeninį tipą.

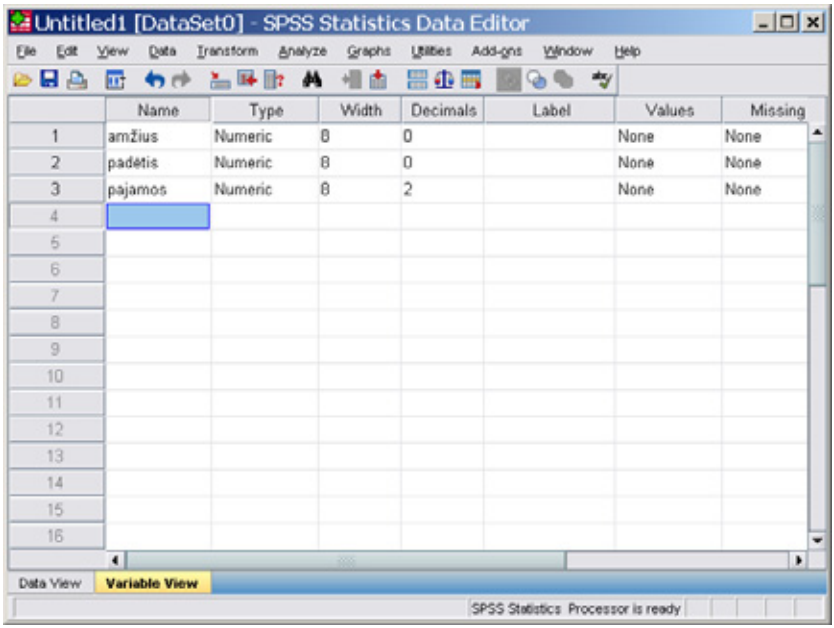


2.2 pav. **Kintamųjų reikšmės *Data View* lange**

Pradedant kurti naują duomenų bazę matavimo skalė *Measure* automatiškai pasirenkama metrinė *Scale*. Ranginiams kintamiesiems ma-

tavimo skalę reikia pakeisti į *Ordinal*, o vardiniams kintamiesiems – į *Nominal*:

- Spragtelėkite stulpelio *Measure* atitinkamame langelyje. Langelio dešinėje pamatysite juodą trikampėlį.
- Spustelėkite juodą trikampėlį.
- Atsivėrusiame sąraše pasirinkite reikiamą matavimo skalę.
- Spragtelėkite duomenų peržiūros *Data View* auselę – įvesime kintamųjų reikšmes (žr. 2.2 pav.):
  - Stulpelyje *amžius* įrašykite 47.
  - Stulpelyje *padėtis* įrašykite 1.
  - Stulpelyje *pajamos* įrašykite 75 000.
  - Perkelkite teksto žymeklį į antrą eilutę ir atitinkamuose langeliuose įrašykite 32, 0 ir 152 000.



2.3 pav. **Koduoti kintamieji *amžius* ir *padėtis* rodomi be trupmeninės dalies**

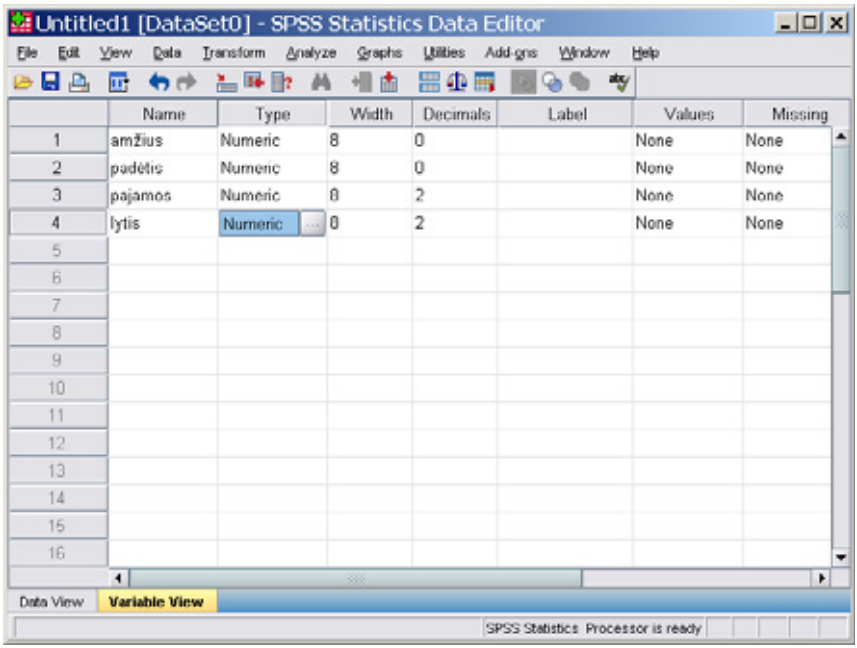
Kaip matome, stulpeliuose *amžius* ir *padėtis* duomenys vaizduojami dešimtainiu formatu. Jei reikia, galima trupmeninę šių kintamųjų dalį paslėpti (žr. 2.3 pav.):

- Spragtelėkite kintamųjų peržiūros *Variable View* auselę duomenų redaktoriaus lango apačioje.
- Eilučių *amžius* ir *padėtis* stulpelyje *Decimals* įrašykite 0.

Tekstinių duomenų įvedimas

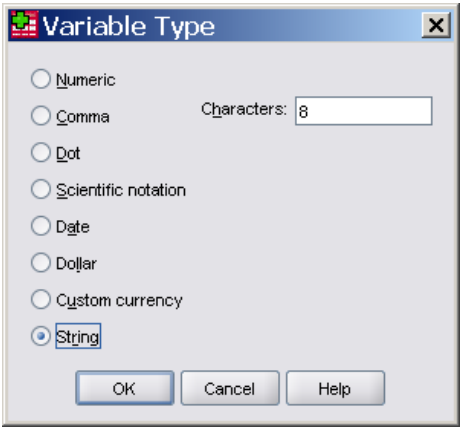
Šalia skaitmeninių duomenų į SPSS duomenų redaktorių įveskime ir tekstinius – teksto eilutes.

- Spragtelėkite kintamųjų peržiūros *Variable View* auselę duomenų redaktoriaus lango apačioje.
- Pirmos tuščios eilutės pirmajame langelyje įrašykite naujo kintamojo vardą *lytis*.
- Spragtelėkite naujojo kintamojo langelyje *Type*. Langelio *Type* dešinėje atsiranda mygtukas su trimis taškeliais (žr. 2.4 pav.).



2.4 pav. **Mygtukas kintamojo *lytis* tipui pasirinkti**

- Spustelėkite mygtuką su trimis taškeliais langelio *Type* dešinėje. Pamatysite kintamojo tipo pasirinkimo dialogo langą (žr. 2.5 pav.):



2.5 pav. **Kintamojo tipo pasirinkimo langas**

- Pasirinkite kintamojo tipą *String* ir spragtelėkite mygtuką *OK*.

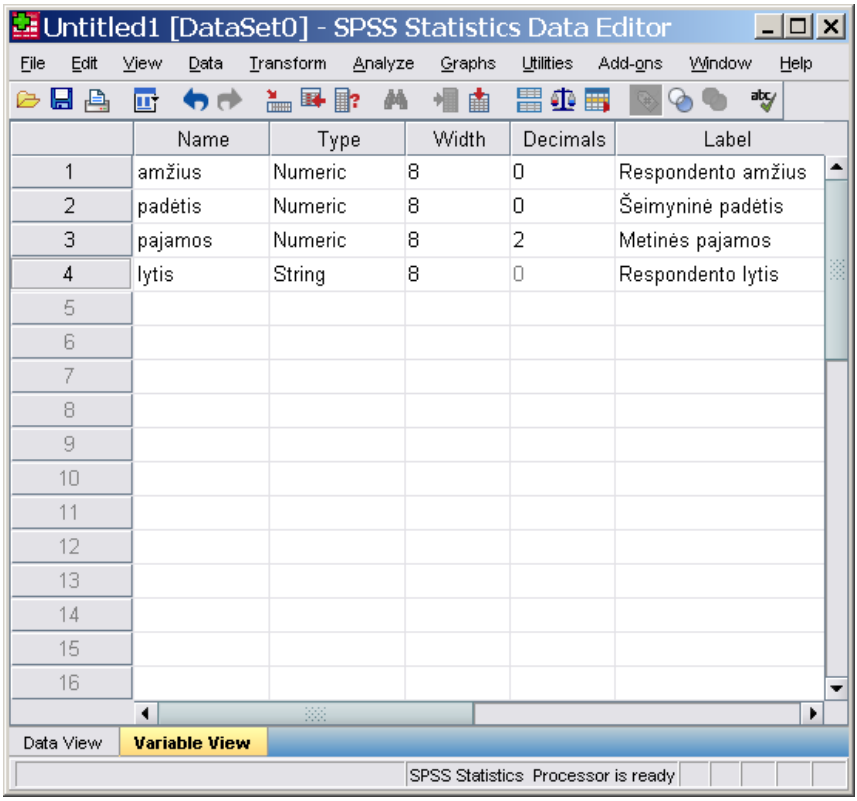
Duomenų aprašymas

Kartu su kintamųjų tipų aprašymu, galima papildomai aprašyti kintamųjų vardus bei duomenų reikšmes. Šie aprašymai yra naudojami statistikos ataskaitose ir diagramose.

Kintamųjų žymenos *Labels*

Kintamiesiems aprašyti naudojamos žymenos *Labels* (žr. 2.6 pav.). Paprastai tai platesnis kintamojo vardo variantas. Žymenos gali būti iki 256 simbolių ilgio.

- Spragtelėkite kintamųjų peržiūros *Variable View* auselę duomenų redaktoriaus apačioje.
- Eilutės *amžius* stulpelyje *Label* įrašykite „Respondento amžius“.
- Eilutės *padėtis* stulpelyje *Label* įrašykite „Šeiminė padėtis“.
- Eilutės *pajamos* stulpelyje *Label* įrašykite „Metinės pajamos“.
- Eilutės *lytis* stulpelyje *Label* įrašykite „Respondento lytis“.

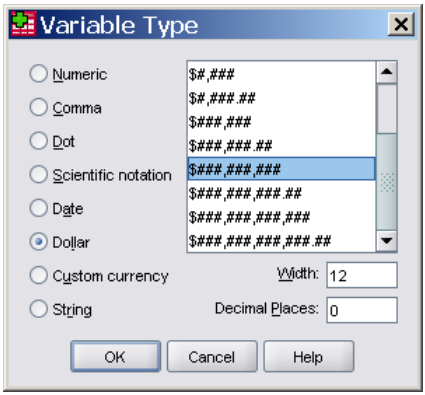


2.6 pav. **Kintamųjų žymenos**

Kintamųjų tipo ir formato keitimas

Stulpelyje *Type* matome kiekvieno kintamojo tipą einamuoju momentu. Dažniausiai tai skaitmeniniai ar tekstiniai kintamieji. SPSS palaiko ir kitokius kintamųjų formatus. Kaip rodo mūsų pavyzdys, kintamasis *pajamos* yra skaitmeninio tipo. Pakeisime jį į valiutos tipą. SPSS turi specialiai išskirtą valiutos tipą *Dollar*.

- Spragtelėkite eilutės *pajamos* langelį *Type*, tada paspauskite šiame langelyje atsiradusį mygtuką su trimis taškeliais. Bus atvertas kintamojo tipo *Variable Type* dialogo langas (žr. 2.7 pav.):



2.7 pav. **Kintamųjų tipo pasirinkimo langas**

- Spragtelėkite *Dollar*.
- Pasirinkite valiutos formatą $###,###,###.
- Paspauskite mygtuką *OK*.

Skaitmeninių kintamųjų žymenos

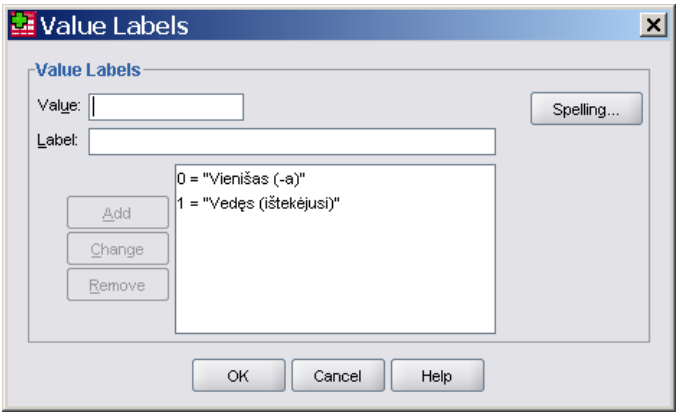
Reikšmių žymenos leidžia aprašyti skaitmeninių kintamųjų tam tikras reikšmes teksto eilutėmis. Mūsų pavyzdys rodo, jog kintamasis *padėtis* turi dvi reikšmes: 0 – „subjektas vienišas (nevedęs, netekėjusi, našlys, našlė, išsituokęs, išsituokusi)“ ir 1 – „subjektas vedęs (ištekėjusi) oficialioje ar neoficialioje santuokoje“.

- Spragtelėkite eilutės *padėtis* langelį *Values*, tada šio langelio dešinėje esantį mygtuką. Atsivers reikšmių žymenų *Value Labels* dialogo langas (žr. 2.8 pav.).

Value – skaitmeninio kintamojo reikšmė.

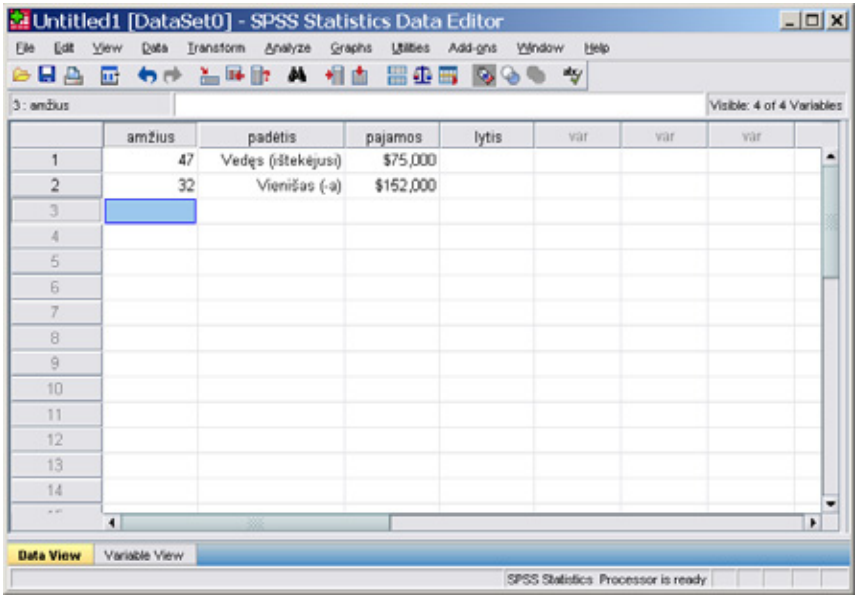
Label – tekstinė žymena, priskiriama pažymėtai skaitmeninio kintamojo reikšmei.

- Įrašykite „0“ į laukelį *Value*.
- Įrašykite „Vienišas (-a)“ į laukelį *Label*.
- Spragtelėkite *Add* norėdami šią žymeną priskirti sąrašui.
- Įrašykite „1“ į laukelį *Value*.
- Įrašykite „Vedęs (ištekėjusi)“ į laukelį *Label*.
- Spragtelėkite *Add* norėdami šią žymeną priskirti sąrašui.
- Paspauskite mygtuką *OK*.



2.8 pav. **Reikšmių žymenų pasirinkimo langas**

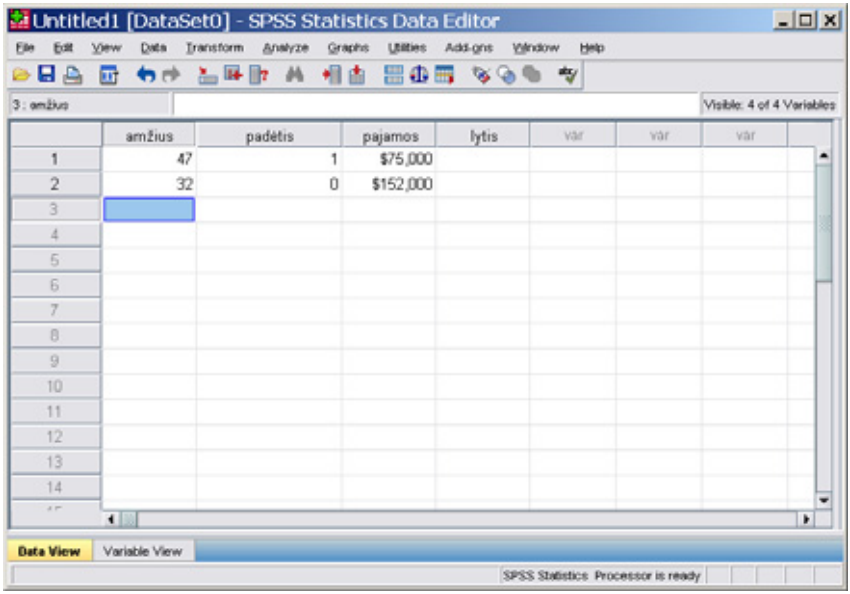
SPSS duomenų redaktoriuje mūsų pavyzdys atrodo taip (žr. 2.9 pav.):



2.9 pav. **Duomenų reikšmių peržiūros *Data View* langas su kintamojo *padėtis* reikšmių žymenomis**

Mūsų atveju reikšmių žymenos *Value Labels* yra aktyvios.

Per meniu *View* → *Value Labels* arba mygtuku  galima reikšmių žymenas *Value Labels* padaryti neaktyvias (žr. 2.10 pav.):



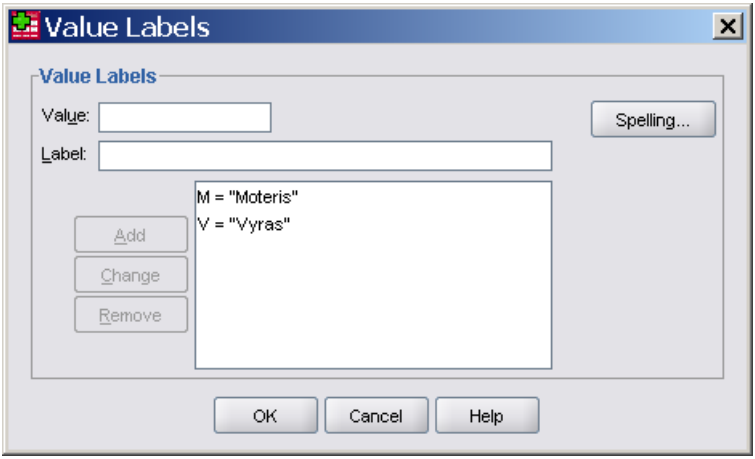
2.10 pav. **Duomenų reikšmių peržiūros *Data View* langas be kintamojo *padėtis* reikšmių žymenų**

Tekstinių kintamųjų žymenos

Kartais tekstinius kintamuosius patogu papildomai aprašyti. Pavyzdžiui, mūsų duomenys rodo, jog *lytis* gali būti pažymėta raidėmis V ir M. Galėtume pasirinkti žymenas Vyras ir Moteris, atitinkamai žyminčias kintamuosius V ir M (žr. 2.11 pav.).

- Spragtelėkite kintamųjų peržiūros *Variable View* auselę duomenų redaktoriaus lango apačioje.
- Paspauskite eilutės *lytis* langelį *Values*, tada spragtelėkite langelio dešinėje atsiradusį mygtuką su trimis taškeliais. Atsivers reikšmių žymenų *Value Labels* dialogo langas.
- Įrašykite „V“ į laukelį *Value* ir „Vyras“ į laukelį *Label*.

- Spragtelėkite mygtuką *Add*.
- Įrašykite „M“ į laukelį *Value* ir „Moteris“ į laukelį *Label*.
- Spragtelėkite mygtuką *Add*.



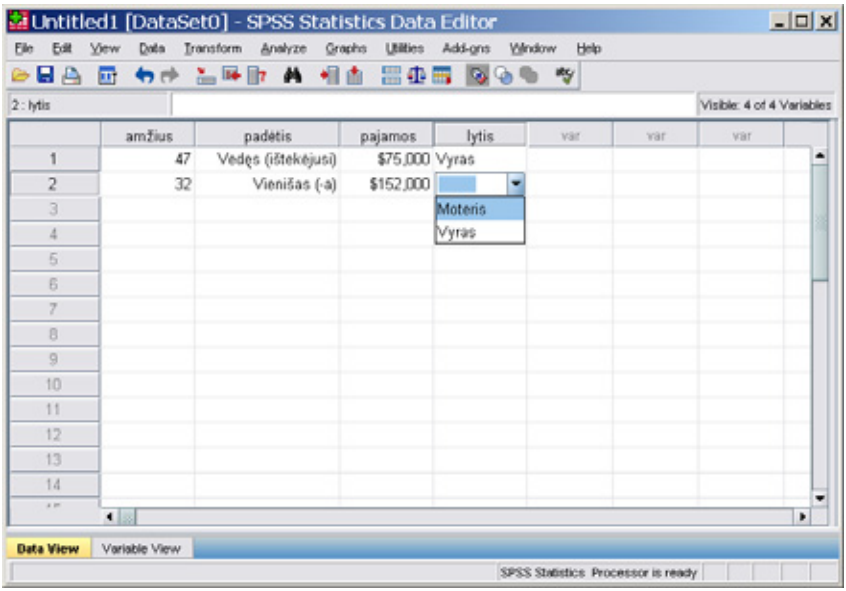
2.11 pav. **Reikšmių žymenų pasirinkimo langas**

- Paspauskite mygtuką *OK*. Išsaugosite pakeitimus ir grįšite į duomenų redaktorių.

Dėmesio. Tekstiniuose kintamuosiuose didžiosios ir mažosios raidės suvokiamos kaip skirtingos, todėl „m“ yra ne tas pats, kaip „M“.

Reikšmių žymenų naudojimas duomenims įvesti (žr. 2.12 pav.)

- Spustelėkite duomenų peržiūros *Data View* auselę duomenų redaktoriaus lango apačioje.
- Pirmoje eilutėje pažymėkite langelį *lytis*.
- Spragtelėkite šio langelio dešinėje ir atsiveriančiame sąraše pasirinkite *Vyras*.
- Analogiškai kitoje eilutėje pasirinkite *Moteris*.



2.12 pav. **Reikšmių žymenų naudojimas duomenims įvesti**

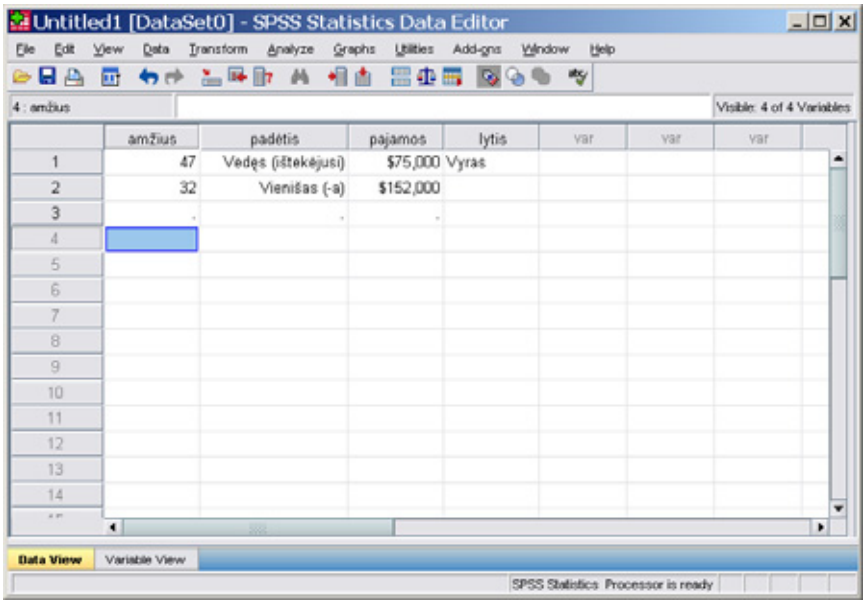
Kaip matome, galima pasirinkti tik tas reikšmes, kurios yra apibrėžtos atsiveriančiame sąraše.

Trūkstamų (*missing*) duomenų tvarkymas

Trūkstami ar neteisingai įvesti duomenys pasitaiko pernelyg dažnai, kad į juos galėtume nekreipti dėmesio. Kaip atsiranda šie duomenys? Apklausos respondentai gali į tam tikrus klausimus neatsižvelgti, gali nežinoti atsakymų, gali atsakyti netinkama forma. Jei tokių atvejų nenustatysime ar neatmesime, mūsų apklausos rezultatai gali būti neteisingi.

Skaitmeniniams duomenims tušti ar neteisingo formato langeliai yra paverčiami sisteminiais trūkstamais (*system-missing*), pažymint juos vienu tašku (žr. 2.13 pav.).

Analizuojant duomenis, nustatyti, kodėl reikšmė yra praleista (trūkstama), gali būti labai svarbu. Pavyzdžiui, gali būti svarbu nustatyti, kurie respondentai atmetė klausimus sąmoningai, o kurie tiesiog neatsakė į netinkamus klausimus.



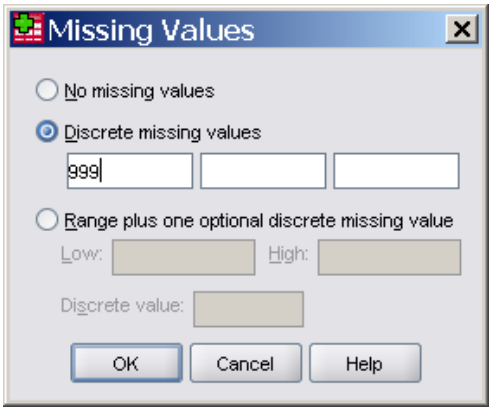
2.13 pav. **Trūkstamos reikšmės pažymėtos taškais**

Trūkstamos skaitmeninių kintamųjų reikšmės

- Spragtelėkite kintamųjų peržiūros *Variable View* auselę duomenų redaktoriaus lango apačioje.
- Paspauskite eilutės *amžius* langelį *Missing* ir langelio dešinėje esantį mygtuką. Atsivers trūkstamų reikšmių *Missing Values* dialogo langas (žr. 2.14 pav.).

Čia galite pažymėti iki 3 skirtingų trūkstamų reikšmių arba reikšmių intervalą pridedant vieną papildomą atskirą reikšmę.

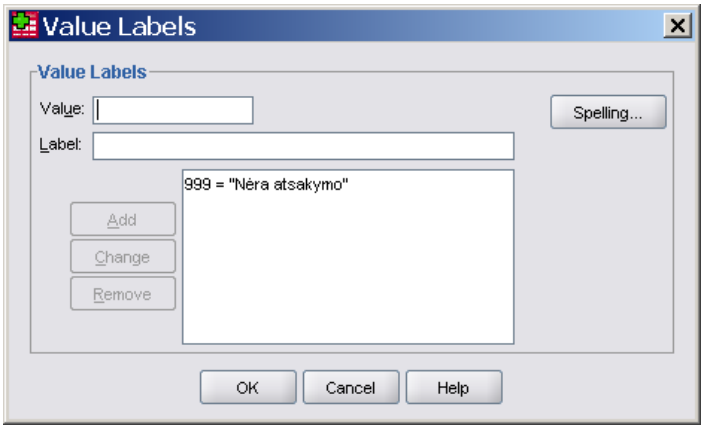
- Pasirinkite *Discrete missing values*.
- Įrašykite „999“ į pirmąjį langelį. Kitus palikite tuščius. „999“ pasirenkame tik todėl, kad užtikrintai nesutaptų su kitais duomenimis (galėtų būti ir bet koks kitas skaičius).
- Paspauskite mygtuką *OK* ir grįžkite į SPSS duomenų redaktorių.



2.14 pav. **Trūkstamų reikšmių *Missing Values* dialogo langas**

Dabar, kai pažymėjome trūkstamų duomenų reikšmes, galime jas pritaikyti trūkstamų duomenų žymenose (žr. 2.15 pav.).

- Spragtelėkite eilutės *amžius* langelį *Values*, tada paspauskite langelio dešinėje esantį mygtuką. Atsivers reikšmių žymenų *Value Labels* dialogo langas.
- Įrašykite „999“ į laukelį *Value*.
- Įrašykite „Nėra atsakymo“ į laukelį *Label*.



2.15 pav. **Reikšmių žymenų *Value Labels* dialogo langas**

- Spragtelėkite mygtuką *Add*.
- Spustelėkite mygtuką *OK*, išsaugokite pakeitimus ir grįžkite į SPSS duomenų redaktorių.

Trūkstamos tekstinių kintamųjų reikšmės

Trūkstamos tekstinių kintamųjų reikšmės tvarkomos panašiai kaip ir skaitmeninių kintamųjų, tik tušti tekstinių kintamųjų laukai nelaikomi sisteminiais trūkstamais (*system-missing*) – jie dažniau interpretuojami kaip tuščia eilutė.

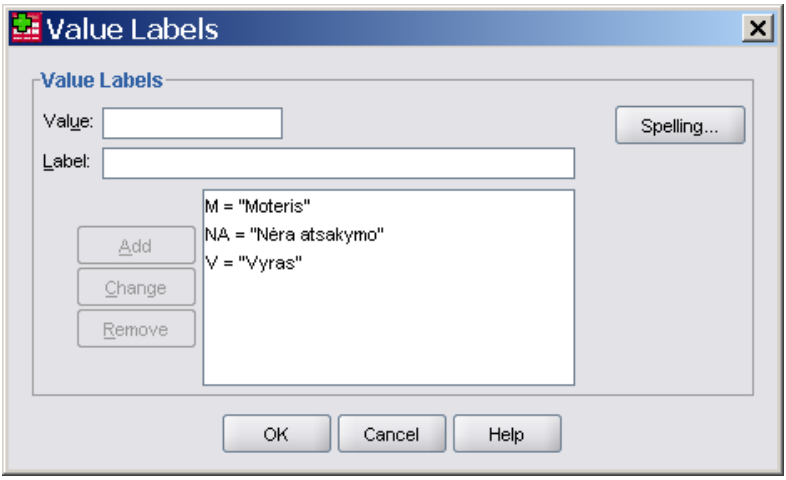
- Spragtelėkite *Variable View* auselę duomenų redaktoriaus lango apačioje.
- Paspauskite eilutės *lytis* langelį *Missing*, paskui spustelėkite langelio dešinėje atsiradusį mygtuką su trimis taškeliais. Pasirodys trūkstamų reikšmių *Missing Values* dialogo langas.
- Pasirinkite *Discrete missing values*.
- Įrašykite „*NA*“ į pirmąjį teksto langelį.

Dėmesio. Tekstiniuose kintamuosiuose didžiosios ir mažosios raidės suvokiamos kaip skirtingos, todėl „*na*“ yra ne tas pats, kaip „*NA*“.

- Paspauskite mygtuką *OK*, pakeitimus išsaugokite ir grįžkite į SPSS duomenų redaktorių.

Dabar galima pridėti trūkstamoms reikšmėms žymenas (žr. 2.16 pav.).

- Spragtelėkite eilutės *lytis* langelį *Values*. Spustelėkite langelio dešinėje esantį mygtuką su trimis taškeliais. Atsivers reikšmių žymenų *Value Labels* dialogo langas.
- Įrašykite „NA“ į laukelį *Value*.
- Įrašykite „Nėra atsakymo“ į laukelį *Label*.
- Spragtelėkite mygtuką *Add*.
- Paspauskite mygtuką *OK*, išsaugokite pakeitimus ir grįžkite į SPSS duomenų redaktorių.
- Išsaugokite mūsų pavyzdį faile vardu **duomenys.sav.**



2.16 pav. **Reikšmių žymenų *Value Labels* dialogo langas**

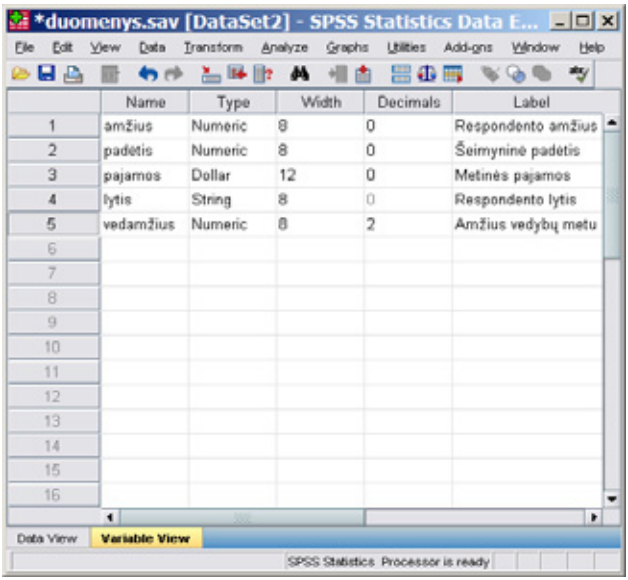
Kintamųjų atributų kopijavimas ir pritaikymas kitiems kintamiesiems

Vienų kintamųjų atributus galima pritaikyti kitiems kintamiesiems (žr. 2.17 pav.) ir taip sutaupyti daug pastangų ir laiko.

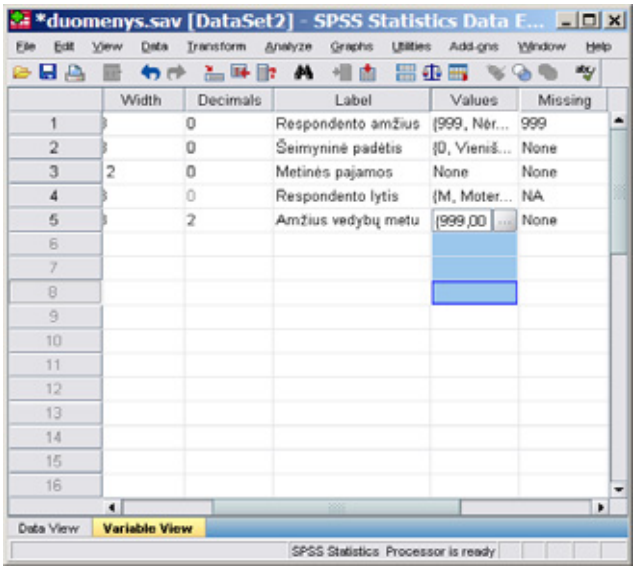
- Lange *Variable View* įveskite naująjį kintamąjį *vedamžius*.
- Stulpelyje *Label* įrašykite „Amžius vedybų metu“.
- Spragtelėkite dešiniuoju pelės klavišu langelį *Values* eilutėje *amžius.*
- Pasirinkite komandą *Copy*.
- Paspauskite dešiniuoju pelės klavišu langelį *Values* eilutėje *vedamžius.*
- Pasirinkite komandą *Paste*.

Atitinkama kintamojo *amžius* reikšmė yra pritaikyta kintamajam *vedamžius*.

Jei reikia pasirinktą atributą pritaikyti keliems kintamiesiems, pažymime atitinkamus langelius (nuspaudę kairįjį pelės klavišą per juos perbraukiame) ir tada pasirenkame komandą *Paste* (žr. 2.18 pav.).



2.17 pav. **Naujojo kintamojo *vedamžius* įvedimas**

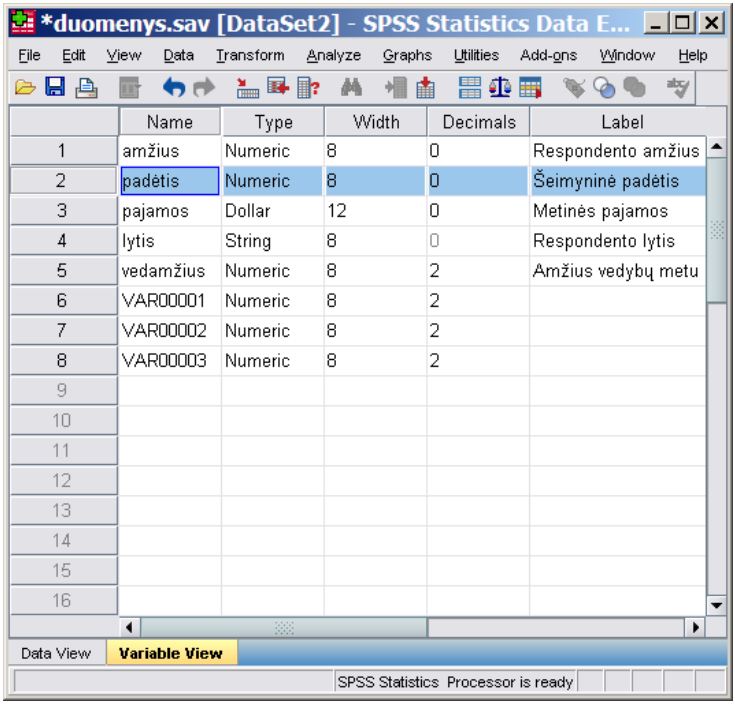


2.18 pav. **Atributo pritaikymas pasirinktiems langeliams**

Jei sukurtume tuščiose eilutėse naujus kintamuosius, jų reikšmės stulpelyje *Values* jau būtų apibrėžtos.

Galima nukopijuoti visus vieno kintamojo atributus kitam kintamajam.

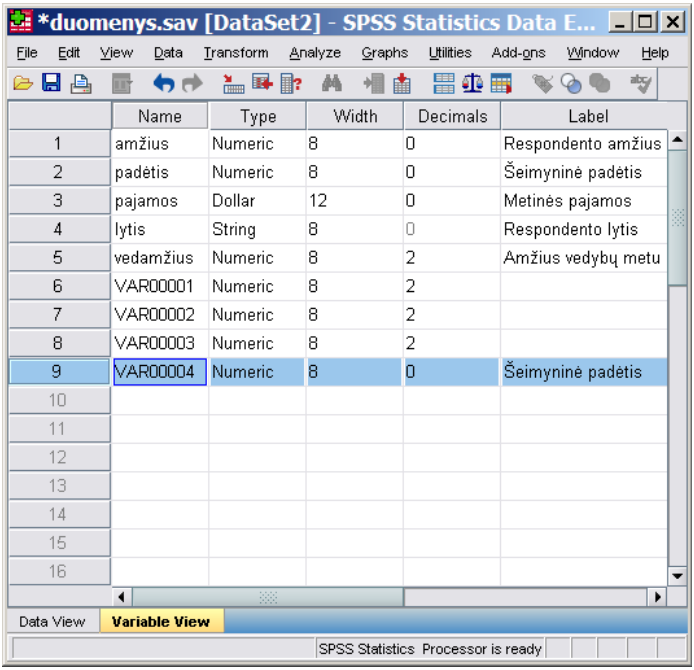
- Spustelėkite eilutės *padėtis* numerį (žr. 2.19 pav.).



2.19 pav. **Pasirenkama eilutė**

- Spragtelėkite dešiniuoju pelės klavišu pažymėtoje eilutėje.
- Pasirinkite komandą *Copy*.
- Paspauskite pirmos laisvos eilutės numerį.
- Spragtelėkite dešiniuoju pelės klavišu pažymėtoje eilutėje.
- Pasirinkite komandą *Paste*.

Visi kintamojo *padėtis* atributai bus pritaikyti naujam kintamajam (žr. 2.20 pav.).



2.20 pav. **Visi kintamojo *padėtis* atributai pritaikyti kintamajam VAR00004**

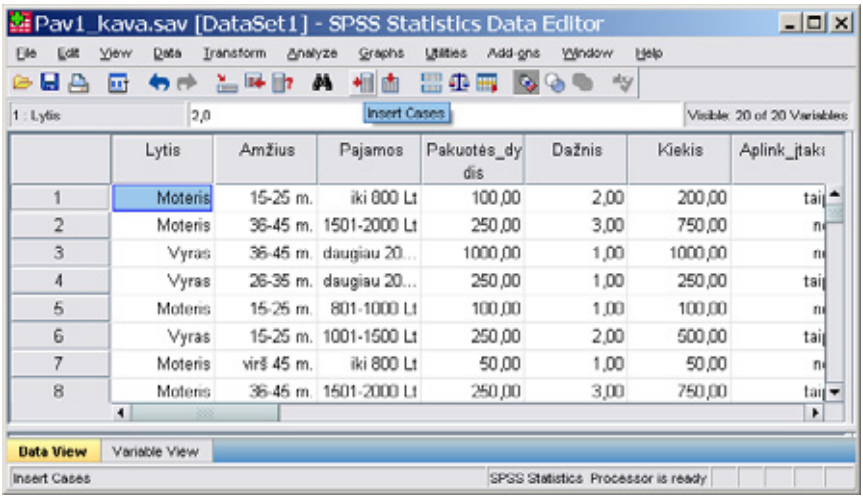
Duomenų skaitymas iš SPSS duomenų failų

Atverti SPSS duomenų failą galima meniu komandomis *File* → *Open* → *Data*… arba failų atvėrimo (*Open File*) mygtuku (žr. 2.21 pav.):



2.21 pav. **Failo atvėrimo (*Open File*) mygtukas**

Abiem atvejais pateikiamas failo atvėrimo dialogas.
Numatytieji SPSS duomenų failai turi plėtinį **.sav*.
2.22 pav. turime atvertą failą *Pav1_kava.sav*.



2.22 pav. **Failas *Pav1_kava.sav* SPSS duomenų redaktoriuje**

SPSS duomenų failai sudaryti iš kintamųjų vardų (stulpeliai) ir kintamųjų reikšmių (eilutės). Kaip rodo mūsų pavyzdys, eilutėse turime respondentų atsakymus į klausimus, surašytus stulpeliuose.

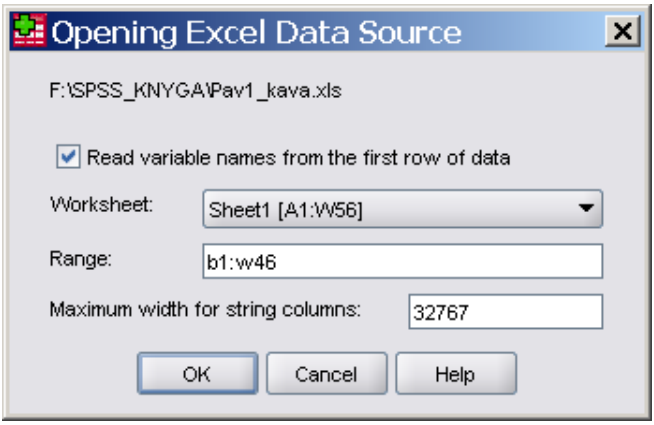
2.2. Duomenų įkėlimas iš kitų šaltinių

Dažnai patogiau vietoj duomenų vedimo tiesiogiai į SPSS duomenų redaktorių juos skaityti iš taikomųjų programų, pavyzdžiui, Microsoft Excel, Microsoft Access ir panašiai.

Duomenų skaitymas iš Excel failo

Atlikite komandas:

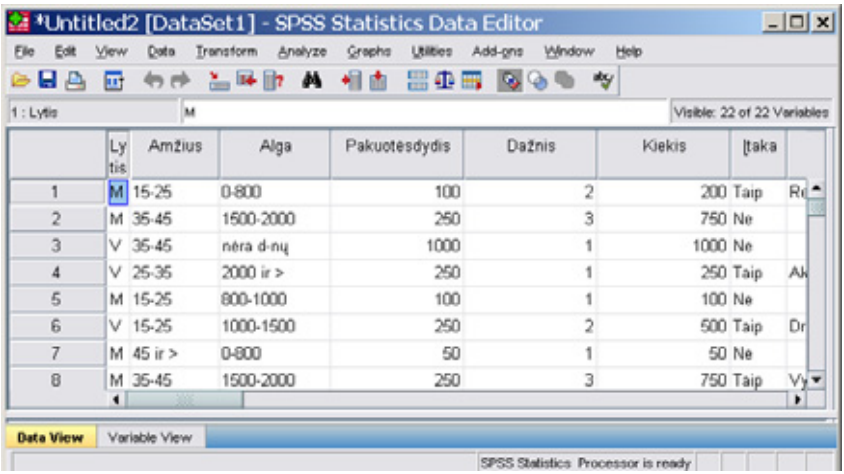
- *File* → *Open* → *Data…*
- Pasirinkite failo tipą Excel (**xls*, **xlsx* arba **xlsm*).
- Pasirinkite hipotetinį failą *Pav1_kava.xls.*
- Paspauskite mygtuką *Open*. Pateikiamas dialogo langas (žr. 2.23 pav.):



2.23 pav. **Excel duomenų šaltinio atvėrimas**

- Įsitikinkite, kad kintamųjų vardai bus skaitomi iš pirmos eilutės (parinktis ☑).
- Jei Excel stulpelių vardai netenkina SPSS kintamųjų vardų sudarymo taisyklių, jie pertvarkomi pagal reikalavimus ir išsaugomi kaip SPSS kintamųjų vardai. Jei norime įkelti dalį skaičiuoklės duomenų, langelyje *Range* pažymėkime reikiamų Excel langelių intervalą. Kadangi Excel failo pirmajame stulpelyje yra respondentų numeriai, o jų mums nereikia, langelyje *Range* pasirinksime intervalą *b1:w46*.
- Spragtelėkite mygtuką *Continue*. Bus perskaitytas Excel failas.

Duomenis dabar matome SPSS duomenų redaktoriuje (žr. 2.24 pav.). Stulpelių vardai tapo kintamųjų vardais. Kadangi originale (Excel faile) kai kurių stulpelių varduose buvo tarpelių, tai SPSS faile jie pašalinti. Pavyzdžiui, Excel faile yra *Pakuotės dydis*, o SPSS faile – *Pakuotėsdydis*.



2.24 pav. **SPSS duomenų redaktoriuje atvertas Excel failas**

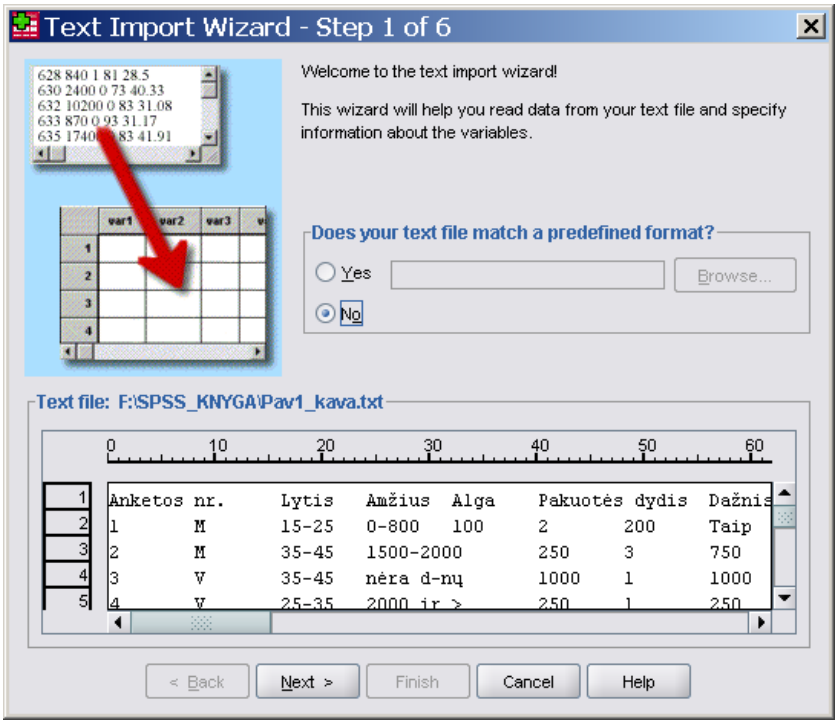
Duomenų skaitymas iš Tekstinio failo

Tekstinis failas yra vienas iš plačiai vartojamų duomenų saugojimo formatų. Dažnai skaičiuoklės ar duomenų bazės leidžia duomenis išsaugoti tekstiniu formatu. Kableliu ar tabuliacijos ženklu atskirtos duomenų eilutės atitinka kintamuosius, kurių vardai paprastai būna failo viršuje (pirmoje eilutėje).

- Atlikite komandas *File* → *Read Text Data…*
- Pasirinkite tekstinį failą *Text* (**txt* arba **dat*).
- Atverkite hipotetinį failą *Pav1_kava.txt.*

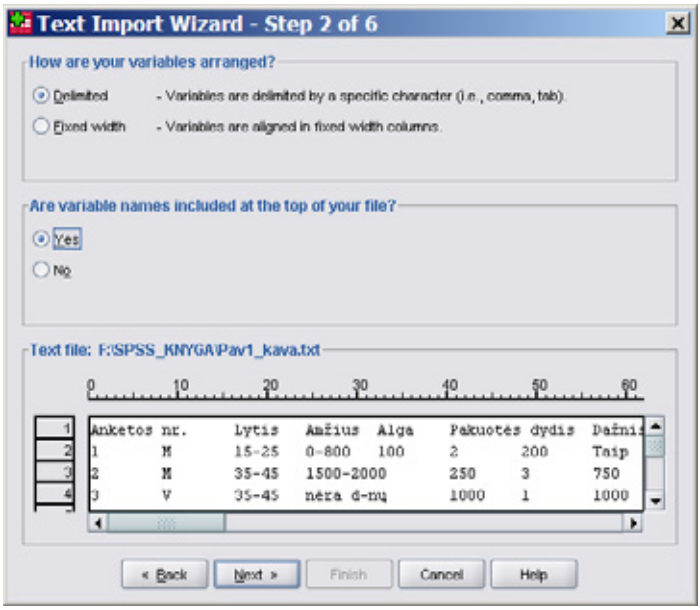
Pasirodo Pagalbos langas, padedantis pasirinktą tekstą interpretuoti (žr. 2.25 pav.):

- Pirmas žingsnis (*Step 1*): galite pasirinkti iš anksto nustatytą dokumento formatą. Paspauskite *No*, kad toks formatas nebūtų pasirinktas.
- Spragtelėkite mygtuką *Next*. Šiame faile kintamųjų duomenys atskirti tabuliacijos ženklu. Kintamųjų vardai yra failo pradžioje – pirmoje eilutėje (žr. 2.26 pav.).

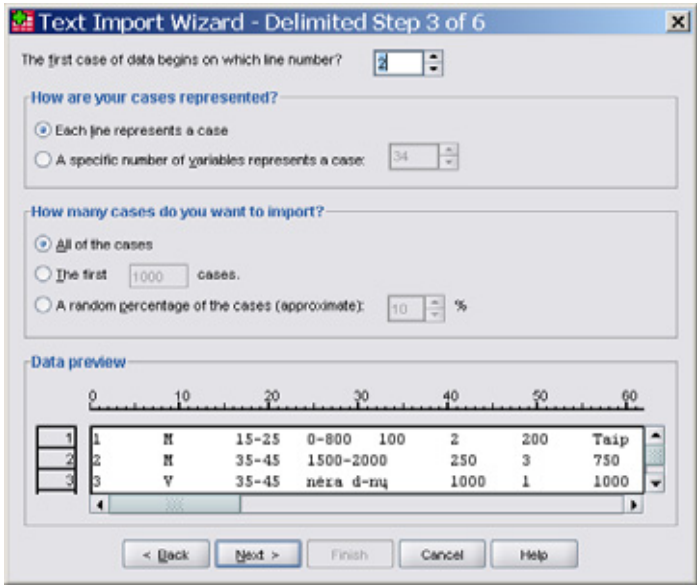


2.25 pav. **Teksto importavimo pagalba. Pirmas žingsnis (*Step 1*)**

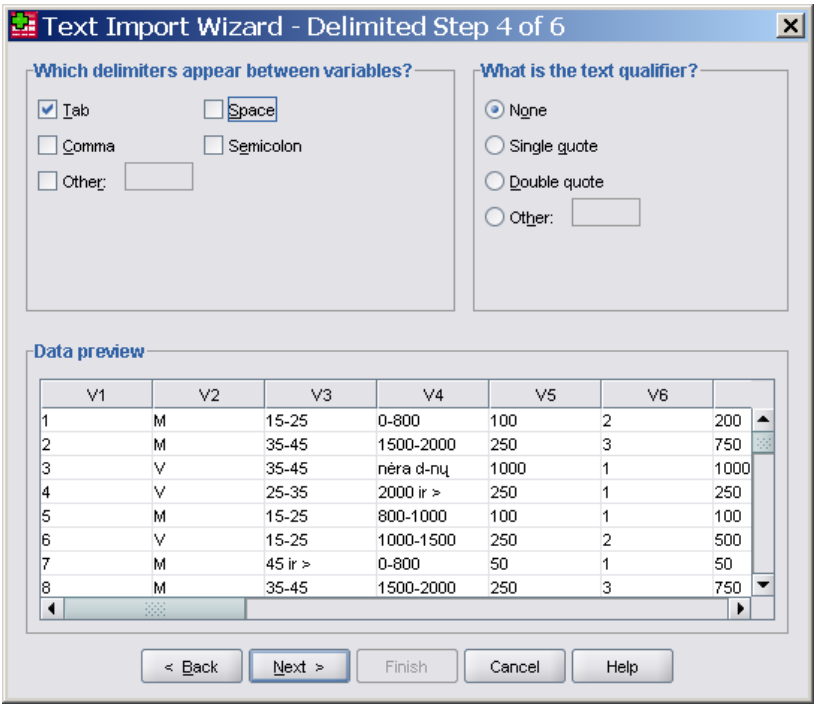
- Pasirinkite *Delimited*, nes failo duomenys atskirti tabuliacijos ženklu.
- Pasirinkite *Yes*, nes kintamųjų vardai yra failo pirmoje eilutėje.
- Spragtelėkite mygtuką *Next.*
- Įrašykite 2, taip pažymėdami, kad duomenys prasideda 2 tekstinio failo eilutėje (žr. 2.27 pav.).
- Kitus parametrus palikite nepakeistus. Spragtelėkite mygtuką *Next.*



2.26 pav. **Teksto importavimo pagalba. Antras žingsnis (*Step 2*)**



2.27 pav. **Teksto importavimo pagalba. Trečias žingsnis (*Step 3*)**



2.28 pav. **Teksto importavimo pagalba. Ketvirtas žingsnis (*Step 4*)**

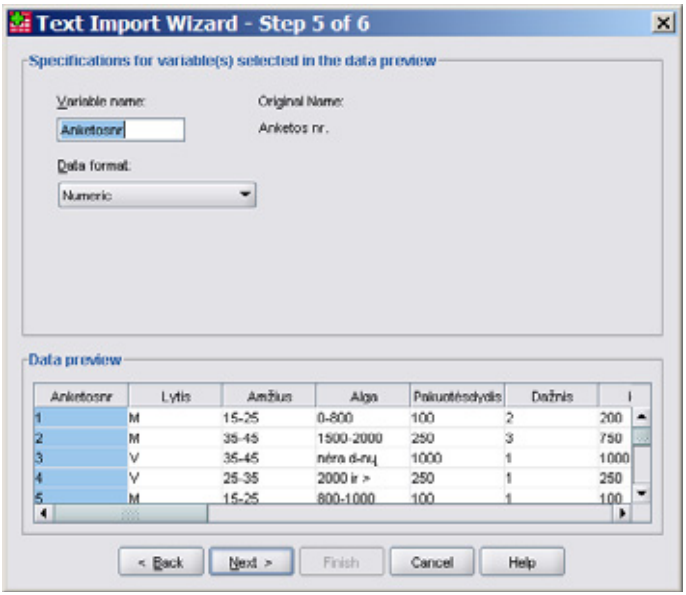
Šiame lange (žr. 2.28 pav.) galėsite įsitikinti, kad duomenys perskaityti teisingai.

- Pasirinkite *Tab*, o kitas parinktis palikite nepažymėtas.
- Spragtelėkite mygtuką *Next.*

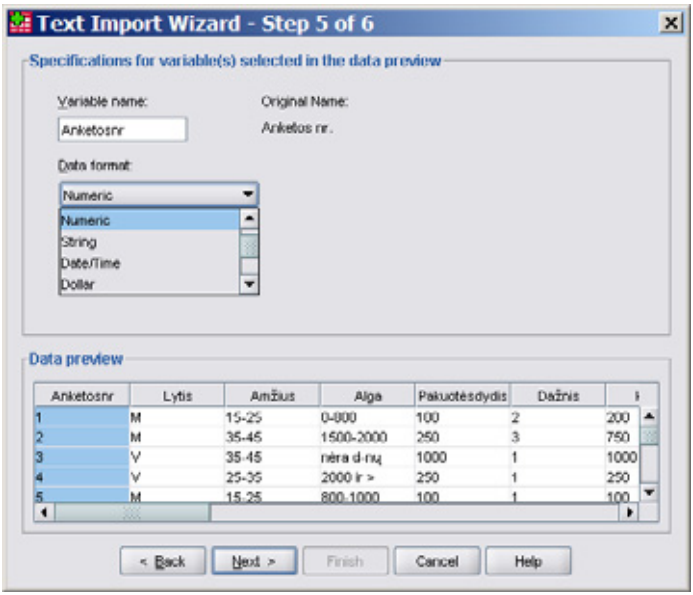
Kadangi kintamųjų vardai formatuojant gali būti sutrumpinti, šiame dialogo lange kintamųjų vardus galima paredaguoti (žr. 2.29 pav.).

Taip pat galima parinkti kintamiesiems reikiamą duomenų formatą.

- Atverkite *Data format* atveriamąjį meniu ir pasirinkite reikalingą duomenų formatą (žr. 2.30 pav.):

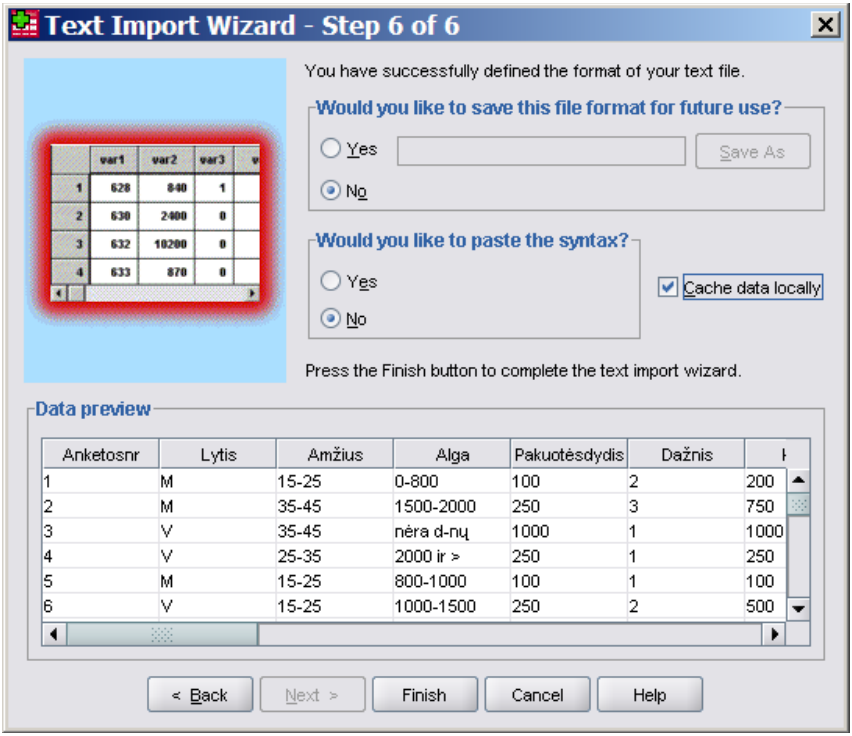


2.29 pav. **Teksto importavimo pagalba. Penktas žingsnis (*Step 5*)**



2.30 pav. **Teksto importavimo pagalba. Penktas žingsnis (*Step 5*)**

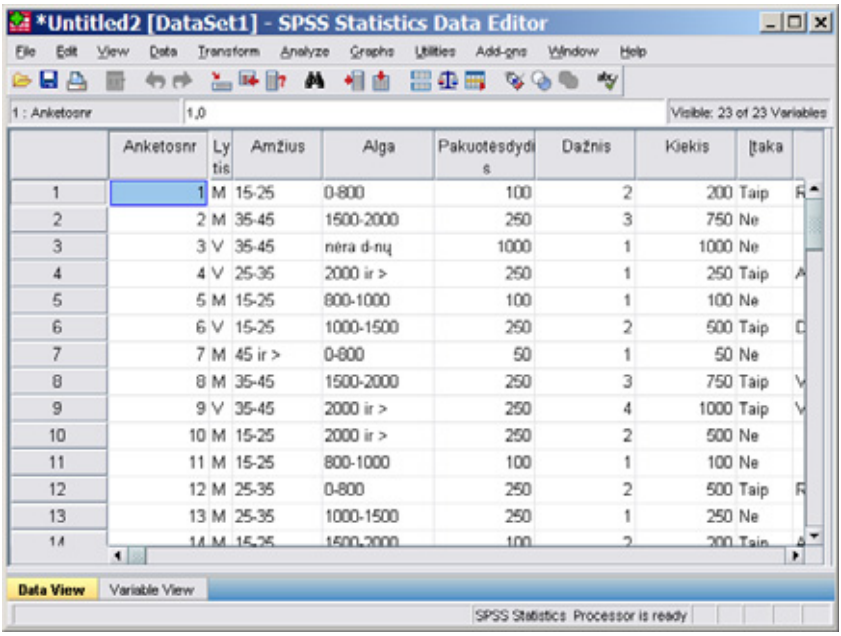
- Spragtelėkite mygtuką *Next* (žr. 2.31 pav.).



2.31 pav. **Teksto importavimo pagalba. Šeštas žingsnis (*Step 6*)**

- Palikite numatytąsias nuostatas ir spragtelėkite mygtuką *Finish* duomenims įkelti.

SPSS duomenų redaktoriuje turime įkeltą tekstinį failą (žr. 2.32 pav.):



2.32 pav. **Įkeltas tekstinis failas**

KONTROLINIAI KLAUSIMAI

2.1. Kokiais būdais į SPSS duomenų bazę galima įvesti tyrimo duomenis?
2.2. Kokie reikalavimai keliami kintamojo vardui SPSS programoje?
2.3. Ar įkeliant duomenis iš skaičiuoklės į SPSS kintamųjų vardai pakinta?

3. PIRMINĖ STATISTINĖ DUOMENŲ ANALIZĖ

Statistinio tyrimo metu surinkti duomenys apie kiekvieną objekto vienetą – labai didelis kiekis medžiagos, kuri dažnai neteikia išsamesnės informacijos apie patį objektą. Medžiagą būtina susisteminti, apibendrinti ir apdoroti. Tik tuomet bus galima išsiaiškinti būdingus statistinės visumos bruožus ir nustatyti dėsningumus. Statistinė analizė, kai kiekvienas kintamasis (požymis) nagrinėjamas atskirai, vadinama *vienmate*. Šiai analizei atlikti taikomi *aprašomosios statistikos* metodai.

Aprašomosios statistikos metodus sudaro:

- duomenų grupavimas;
- dažnių lentelės;
- statistinių charakteristikų (duomenų padėties ir sklaidos rodiklių) skaičiavimas;
- grafinis stebėjimų vaizdavimas.

Sudaromos kintamųjų dažnių lentelės, kuriose pateikiami nominalinėje arba (esant nedideliam kategorijų skaičiui) rangų skalėje išmatuotų kintamųjų dažniai.

Be vienmačių dažnių lentelių sudaromos ir porinės (gali būti ir su daugiau kintamųjų) dažnių lentelės. Jos parodo tarp kintamųjų esančius ryšius. Porinės dažnių lentelės sudaromos nominaliniams ir ranginiams kintamiesiems. Kiekybiniai kintamieji porinėms lentelėms sudaryti nenaudojami.

Aprašomosios statistikos metodų taikymo priklausomybė nuo kintamųjų tipo pateikta 3.4 skyriuje (žr. 3.1 lentelę).

3.1. Duomenų keitimas ir atranka

Pradiniai duomenys ne visada yra tokie, kokius mes norime matyti ataskaitose ar toliau analizuoti. Mums gali prireikti:

- sukurti kokybinius (*categorical*) kintamuosius iš kiekybinių (*scale*) kintamųjų;
- sujungti keletą atsakymų į vieną;
- sukurti naująjį kintamąjį, kuris gaunamas iš jau esančių kintamųjų taikant įvairias matematines funkcijas.

Naudosime hipotetinį duomenų failą *Apklausa_men_pajamos_ neto.sav*.

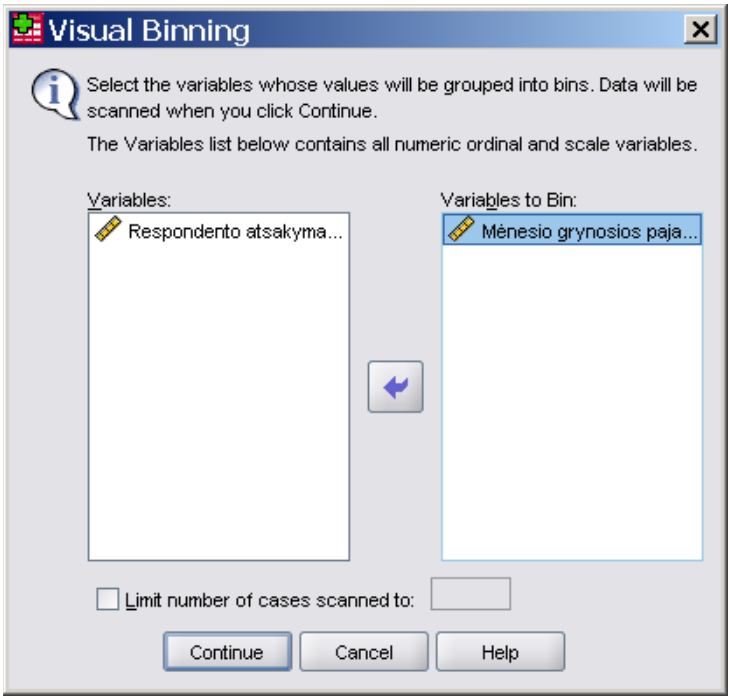
Kokybinių (*categorical*) kintamųjų sukūrimas iš kiekybinių (*scale*) kintamųjų

Faile *Apklausa_men_pajamos_neto.sav* kintamasis *Men_pajamos* yra kiekybinis (*scale*). Esant dideliam atsakymų kiekiui patogu pajamų dydžius sugrupuoti į laisvai apibrėžtus intervalus – kategorijas. Taip iš kiekybinio kintamojo *Men_pajamos* gausime kokybinį (*categorical*) kintamąjį *Men_pajamos_kat*. Šis kintamasis sveikaisiais skaičiais nuo 1 iki 10 pavaizduos pajamų kategorijas: mažiau už 300, 300–599; 600–999; 1 000–1 299; 1 300–1 599; 1 600–1 999; 2 000–2 299; 2 300–2 599; 2 600–2 999; 3 000 ir daugiau.

Sukursime kokybinį kintamąjį *Men_pajamos_kat*:

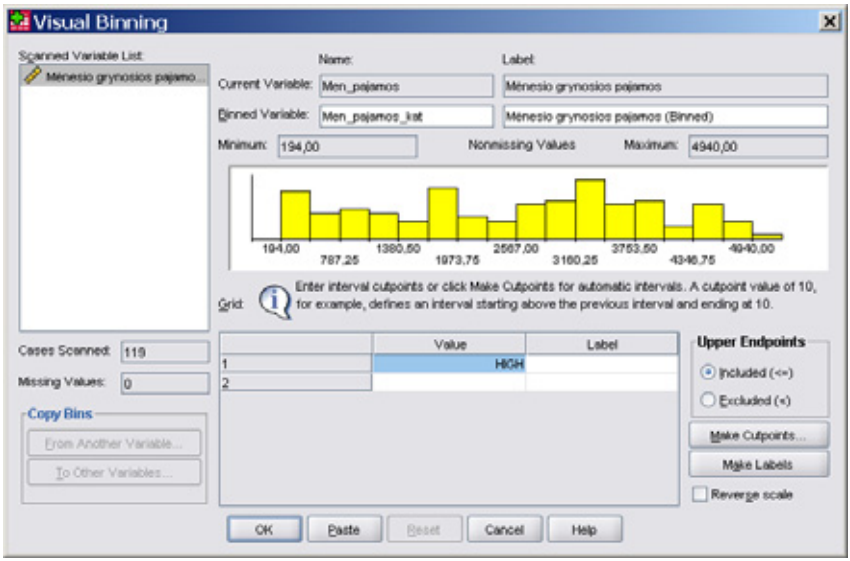
- Duomenų redaktoriuje pasirinkite: *Transform* → *Visual Binning…*

Pradiniame *Visual Binning* dialogo lange pasirenkame kiekybinius (*scale*) ir/ar kokybinius (*ordinal*) kintamuosius, kuriems sukursime naujus, sugrupuotus kintamuosius. Procedūra leidžia dviems ar daugiau pasirinktų gretimų reikšmių sukurti atskiras kategorijas – reikšmių intervalus. Kitais žodžiais tariant, kintamojo reikšmės yra skaitomos (skenuojamos), analizuojamos ir priskiriamos tam tikram intervalui – kategorijai.

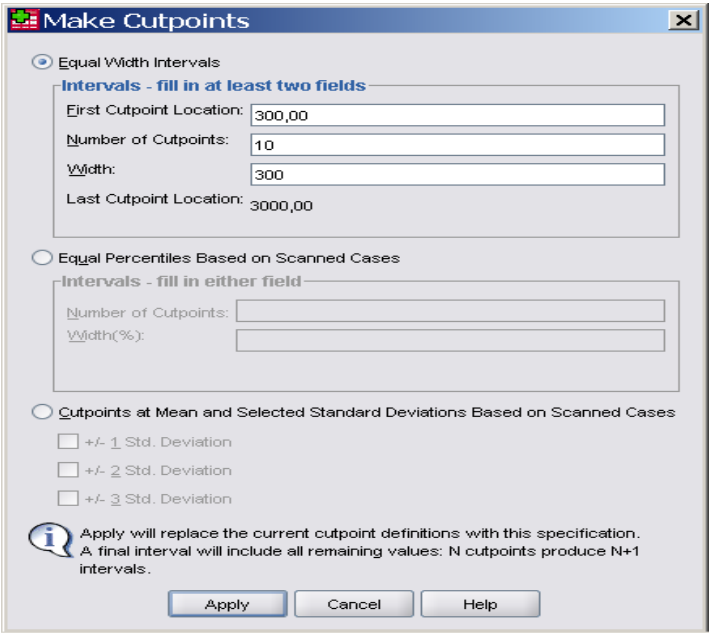


3.1 pav. **Pradinis *Visual Binning* dialogo langas**

- Pele perkelkite kiekybinį kintamąjį *Mėnesio grynosios pajamos* iš sąrašo *Variables* į sąrašą *Variables to Bin* ir spragtelėkite mygtuką *Continue* (žr. 3.1 pav.).
- Sąraše *Scanned Variable List* pažymėkite kintamąjį *Mėnesio grynosios pajamos.* Histogramoje pamatysime pasirinkto kintamojo pasiskirstymą (žr. 3.2 pav.).
- Įveskite naujajam grupuotam kintamajam vardą *Men_pajamos_kat*. Galima pakeisti ir naujojo kintamojo vardą žymenoje *Label*.
- Spragtelėkite mygtuką *Make Cutpoints...* (dalijimo taškai).
- Pasirinkite vienodus intervalus *Equal Width Intervals* (žr. 3.3 pav.).
- Įveskite 300 pirmam dalijimo taškui, 10 dalijimo intervalų skaičiui ir 300 intervalo pločiui pažymėti.



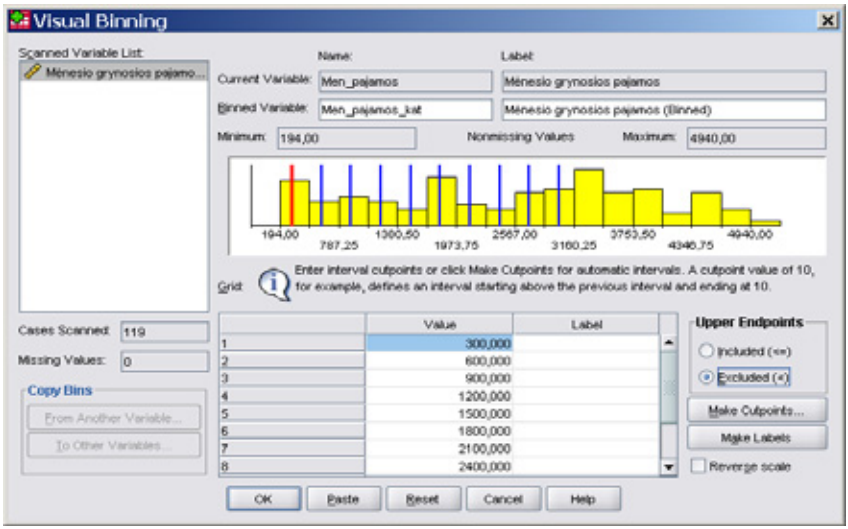
3.2 pav. **Pagrindinis *Visual Binning* dialogo langas**



3.3 pav. ***Visual Binning Make Cutpoints* langas**

Naujųjų kategorijų skaičius yra vienu didesnis, nei dalijimo taškų skaičius.

- Spragtelėkite mygtuką *Apply*.



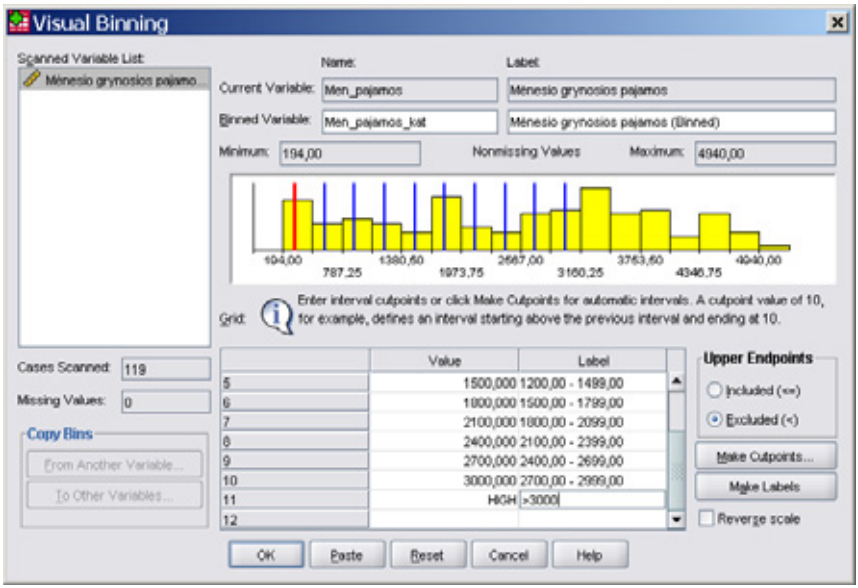
3.4 pav. ***Visual Binning* langas. Pasirinkti dalijimo taškai**

Lentelėje matome dalijimo taškus (žr. 3.4 pav.), kurie vaizduoja kiekvienos kategorijos viršutines ribas. Vertikalios linijos histogramoje taip pat vaizduoja dalijimo taškų padėtį.

Numatytieji dalijimo taškai yra prijungiami prie atitinkamų kategorijų. Mūsų atveju pirmoji kategorija apima visas mažesnes reikšmes ir lygias 300. Tačiau mes planavome, kad pirmoji kategorija apimtų reikšmes mažesnes už 300, 300–599 , … , 3 000 ir daugiau. Todėl kiekvienos kategorijos viršutinių ribų prie kategorijos intervalo prijungti nereikia.

- *Upper Endpoints* grupėje pasirinkite *Excluded* (<).
- Spragtelėkite mygtuką *Make Labels* (žr. 3.5 pav.).

Kiekviena kategorija bus žymima sveikuoju skaičiumi, t. y. naujojo kintamojo *Men_pajamos_kat* reikšmės bus nuo 1 iki 11.

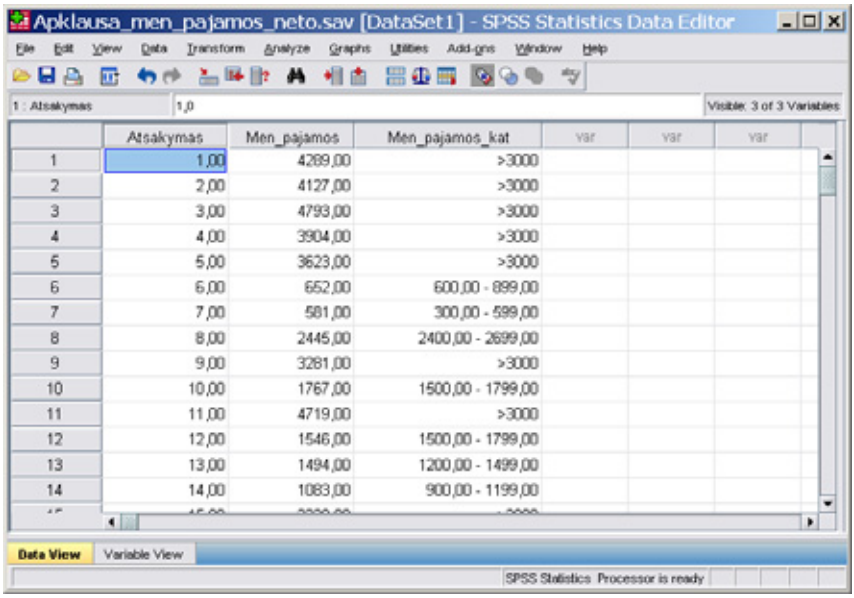


3.5 pav. **Automatiškai sukurtos reikšmių žymenos *Labels***

Lentelėje rankiniu būdu galima įvesti ar pakeisti dalijimo taškus, kategorijų žymenas (mes pakeitėme 3 000 + į >3 000). Histogramoje pertempiant dalijimo linijas galima pakeisti dalijimo taškų padėtį ar juos visai panaikinti atitinkamas dalijimo linijas nutempiant už histogramos ribų.

- Spragtelėkite mygtuką *OK*. Bus sukurtas naujasis sugrupuotas kintamasis.

Naujasis kintamasis pridedamas failo gale (žr. 3.6 pav.), todėl *Data View* lange jis matomas tolimiausiame dešiniajame stulpelyje, o *Variable View* lange – apatinėje eilutėje.



3.6 pav. **Naujasis kintamasis *Men_pajamos_kat* duomenų redaktoriuje**

Naujųjų kintamųjų apskaičiavimas

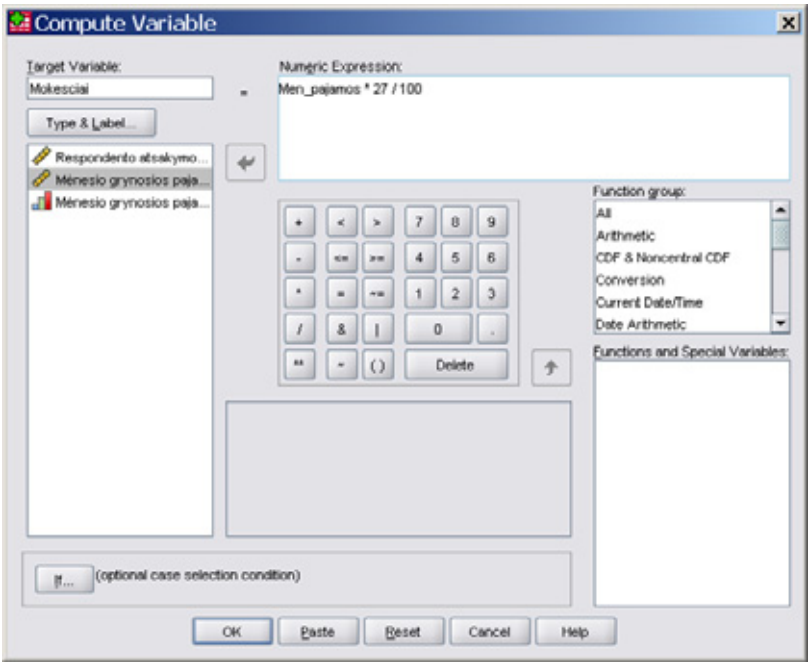
Taikydami matematines funkcijas, iš turimų kintamųjų galime apskaičiuoti naujus. Galima taikyti daugiau kaip 70 funkcijų, suskirstytų į grupes:

- Aritmetinės (*arithmetic*) funkcijos.
- Statistikinės (*statistical*) funkcijos.
- Pasiskirstymo (*distribution*) funkcijos.
- Loginės (*logical*) funkcijos.
- Datos ir laiko (*date and time*) funkcijos.
- Trūkstamų (praleistų) reikšmių (*missing value*) funkcijos.
- Tekstinės (*string*) funkcijos.

Naujojo kintamojo apskaičiavimas taikant aritmetinę išraišką

Sukursime naująjį kintamąjį *Mokesciai*, kuris nuo gaunamų mėnesinių pajamų apskaičiuos 27 procentų dydžio mokestį (žr. 3.7 pav.).

- Duomenų redaktoriuje pasirinkite komandą *Transform* → *Compute Variable…*
- Langelyje *Target Variable* įveskite „Mokesciai“.
- Pasirinkite kintamųjų sąraše *Men_pajamos* ir spragtelėję rodyklę perkelkite kintamąjį į langelį *Numeric Expression*.
- Lange *Compute Variable* esančiu skaičiuotuvu surinkite *27/100.

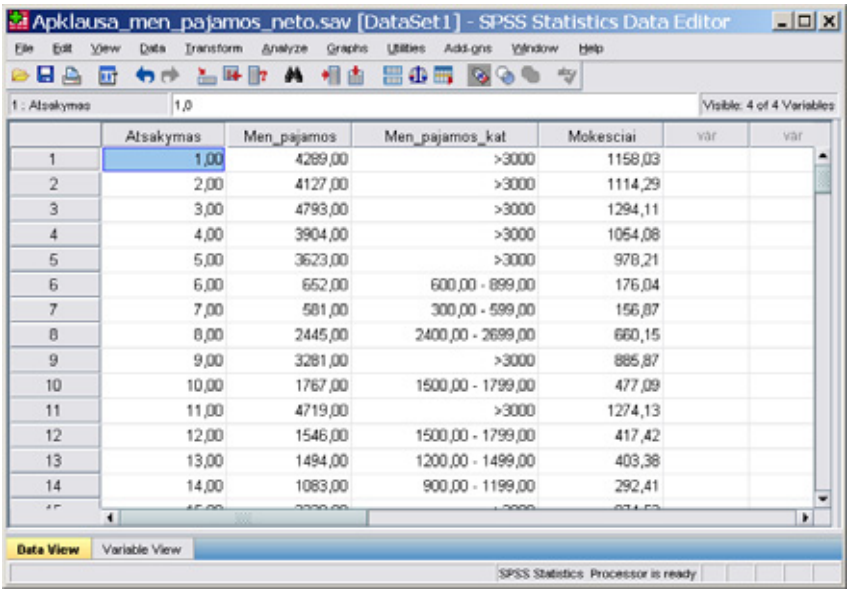


3.7 pav. ***Compute Variable*** **langas**

- Spragtelėkite mygtuką *OK*.

Duomenų redaktoriuje matysite naująjį kintamąjį *Mokesciai*.

Naujasis kintamasis pridedamas failo gale (žr. 3.8 pav.), todėl *Data View* lange jis matomas tolimiausiame dešiniajame stulpelyje, o *Variable View* lange – apatinėje eilutėje.



	Atsakymas	Men_pajamos	Men_pajamos_kat	Mokesciai	var	var
1	1,00	4289,00	>3000	1158,03		
2	2,00	4127,00	>3000	1114,29		
3	3,00	4793,00	>3000	1294,11		
4	4,00	3904,00	>3000	1054,08		
5	5,00	3623,00	>3000	978,21		
6	6,00	652,00	600,00 - 899,00	176,04		
7	7,00	581,00	300,00 - 599,00	156,87		
8	8,00	2445,00	2400,00 - 2699,00	660,15		
9	9,00	3281,00	>3000	885,87		
10	10,00	1767,00	1500,00 - 1799,00	477,09		
11	11,00	4719,00	>3000	1274,13		
12	12,00	1546,00	1500,00 - 1799,00	417,42		
13	13,00	1494,00	1200,00 - 1499,00	403,38		
14	14,00	1083,00	900,00 - 1199,00	292,41		

3.8 pav. **Naujasis kintamasis *Mokesciai* duomenų redaktoriuje**

Naujojo kintamojo apskaičiavimas taikant loginę formulę

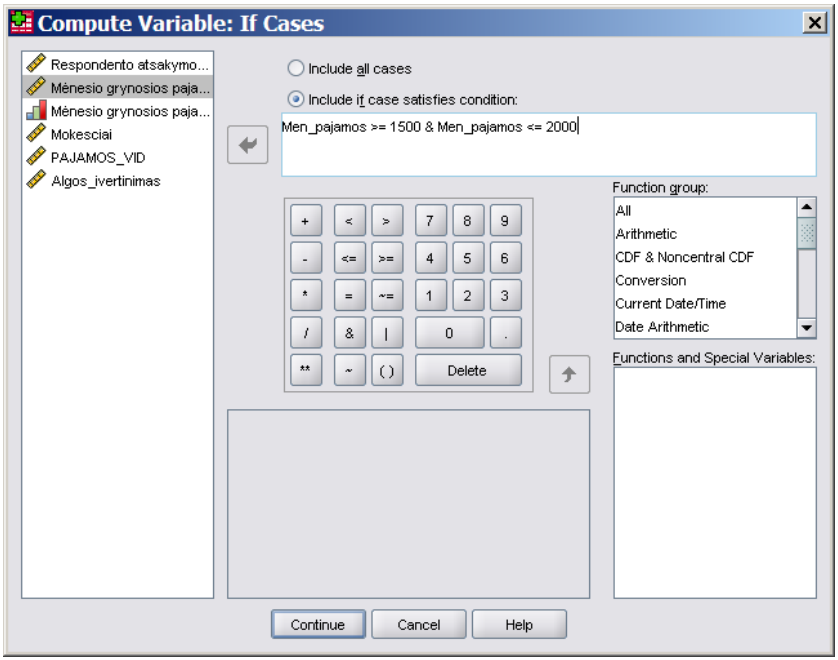
SPSS leidžia apskaičiuoti naujuosius kintamuosius iš esančių stebėjimo duomenų (*cases*) taikant logines sąlygas. Jei loginės sąlygos yra tenkinamos, kiekvienai kintamojo reikšmei yra suteikiama loginė reikšmė *true*, jei sąlygos netenkinamos – *false*, jei reikšmė praleista – *missing*.

Jei loginės formulės rezultatas yra *true*, naujasis kintamasis sukuriamas. Jei loginės formulės rezultatas yra *false* arba *missing*, naujasis kintamasis nesukuriamas.

Loginei (sąlygos) formulei sudaryti galima taikyti kintamųjų vardus, konstantas, aritmetinius operatorius, skaitmenines ir kitas funkcijas, dialogo langelyje esančius santykio operatorius: < (mažiau), > (daugiau), <= (mažiau arba lygu), >= (daugiau arba lygu), ~= (nelygu) ir loginius operatorius: & (loginis IR), | (loginis ARBA), ~ (loginis NE).

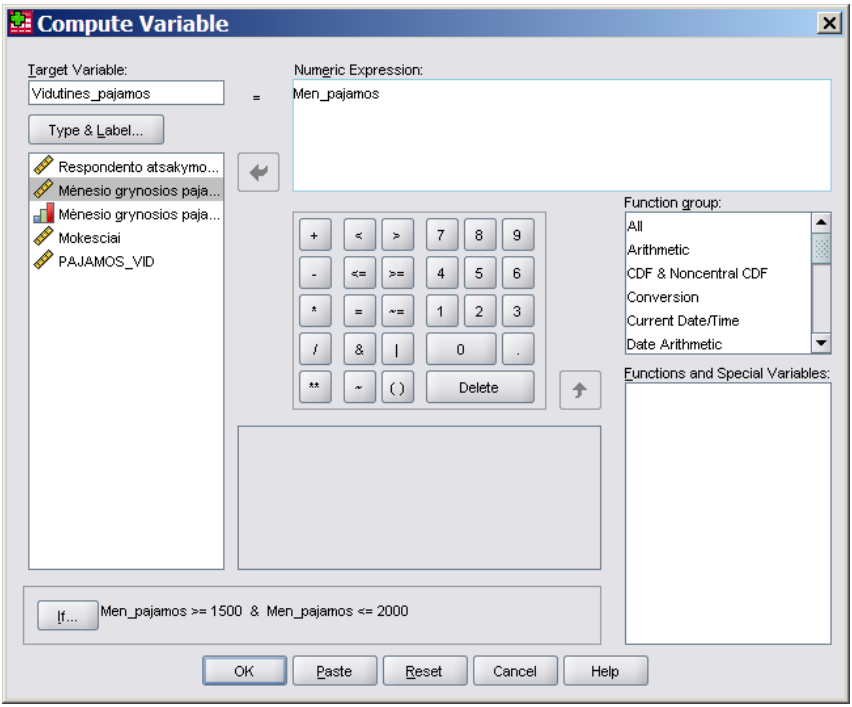
Sukursime naująjį kintamąjį *Vidutines_pajamos*. Šiam kintamajam, taikant loginę formulę, priskirsime kintamojo *Men_pajamos* reikšmes nuo 1 500 iki 2 000 Lt imtinai.

- Duomenų redaktoriuje pasirinkite komandą *Transform → Compute Variable…*
- Spragtelėkite *If...*
- Atsivėrusiame dialogo lange *Compute Variable: If Cases* pasirinkite *Include if case satisfies condition.*
- Pasirinkite kintamųjų sąraše *Men_pajamos* ir spragtelėję rodyklę perkelkite kintamąjį į loginių sąlygų užrašymo langelį.
- Įveskite loginę formulę *Men_pajamos* >= 1500 & *Men_ pajamos*<=2000 (žr. 3.9 pav.).



3.9 pav. **Dialogo langas *Compute Variable: If Cases***

- Spragtelėkite mygtuką *Continue* (žr. 3.10 pav.).

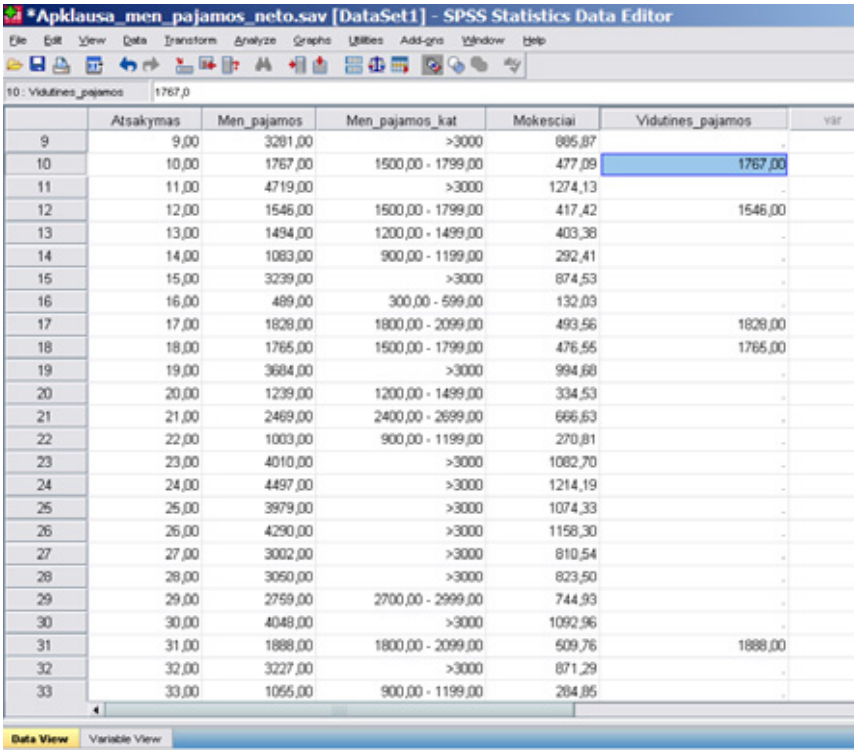


3.10 pav. **Dialogo langas *Compute Variable***

- Langelyje *Target Variable* įveskite naujojo kintamojo vardą *Vidutines_pajamos.*
- Pasirinkite kintamųjų sąraše *Men_pajamos* ir spragtelėję rodyklę perkelkite kintamąjį į langelį *Numeric Expression*.
- Spragtelėkite mygtuką *OK*.

Duomenų redaktoriuje matysite naująjį kintamąjį *Vidutines_ pajamos* (žr. 3.11 pav.).

Naujasis kintamasis pridedamas failo gale, todėl *Data View* lange jis matomas tolimiausiame dešiniajame stulpelyje, o *Variable View* lange – apatinėje eilutėje.



	Atsakymas	Men_pajamos	Men_pajamos_kat	Mokesciai	Vidutines_pajamos
9	9,00	3281,00	>3000	885,87	.
10	10,00	1767,00	1500,00 - 1799,00	477,09	1767,00
11	11,00	4719,00	>3000	1274,13	.
12	12,00	1546,00	1500,00 - 1799,00	417,42	1546,00
13	13,00	1494,00	1200,00 - 1499,00	403,38	.
14	14,00	1083,00	900,00 - 1199,00	292,41	.
15	15,00	3239,00	>3000	874,53	.
16	16,00	489,00	300,00 - 599,00	132,03	.
17	17,00	1828,00	1800,00 - 2099,00	493,56	1828,00
18	18,00	1765,00	1500,00 - 1799,00	476,55	1765,00
19	19,00	3684,00	>3000	994,68	.
20	20,00	1239,00	1200,00 - 1499,00	334,53	.
21	21,00	2469,00	2400,00 - 2699,00	666,63	.
22	22,00	1003,00	900,00 - 1199,00	270,81	.
23	23,00	4010,00	>3000	1082,70	.
24	24,00	4497,00	>3000	1214,19	.
25	25,00	3979,00	>3000	1074,33	.
26	26,00	4290,00	>3000	1158,30	.
27	27,00	3002,00	>3000	810,54	.
28	28,00	3050,00	>3000	823,50	.
29	29,00	2759,00	2700,00 - 2999,00	744,93	.
30	30,00	4048,00	>3000	1092,96	.
31	31,00	1888,00	1800,00 - 2099,00	509,76	1888,00
32	32,00	3227,00	>3000	871,29	.
33	33,00	1055,00	900,00 - 1199,00	284,85	.

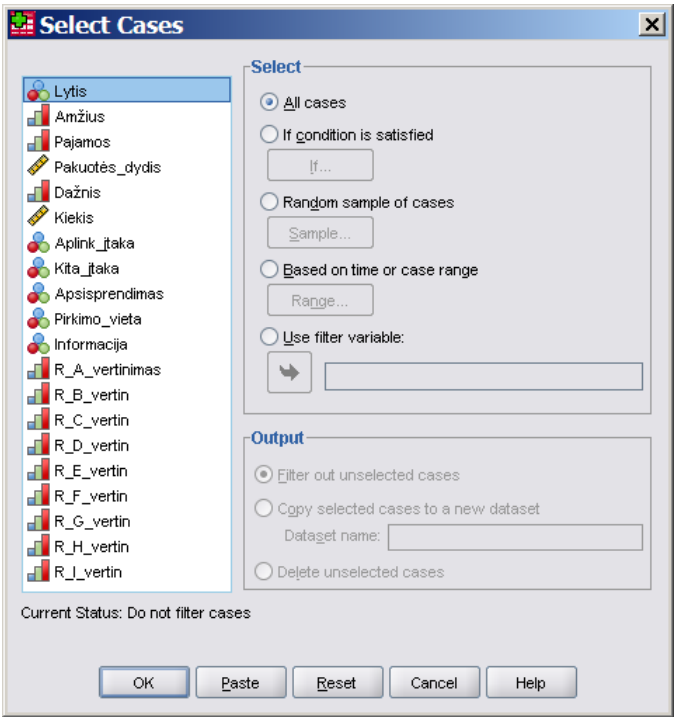
3.11 pav. **Naujasis kintamasis *Vidutines_pajamos* duomenų redaktoriuje**

3.2. Duomenų filtravimas

Dažnai analizei reikalingi tik tie stebėjimų duomenys, kurie tenkina nustatytus kriterijus. Įrašus, kurie netenkina atrankos kriterijų, galima filtruoti arba pašalinti. Filtruoti įrašai lieka duomenų rinkmenoje, tačiau neanalizuojami.

Jei norite atrinkti stebėjimų duomenis:

- Atverkite bylą su analizuojamais duomenimis.
- Nurodykite komandas *Data* → *Select Cases...*

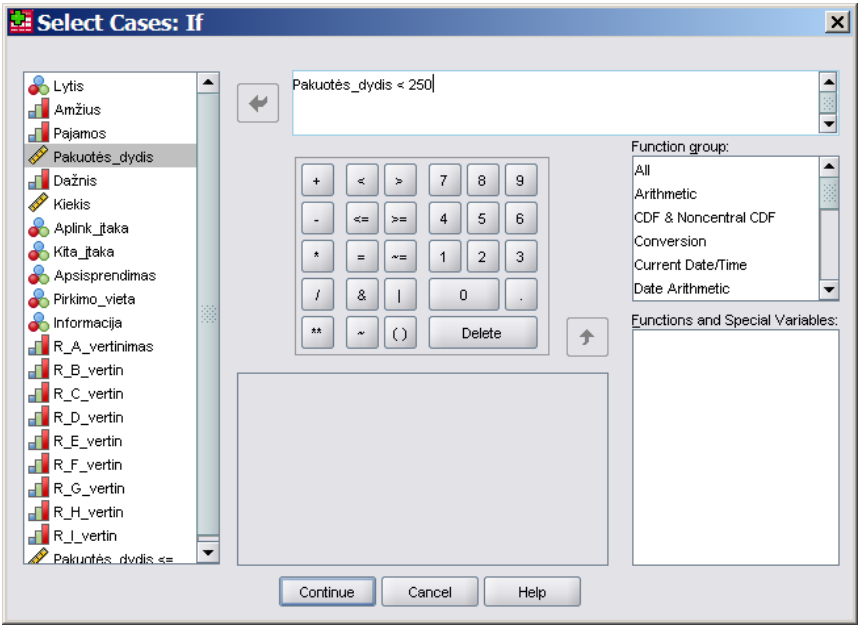


3.12 pav. **Dialogo langas *Select Cases***

- Dialogo lange *Select Cases* (žr. 3.12 pav.) pasirinkite reikiamą duomenų atrankos metodą:
  1. *All cases* (visos reikšmės) – tai variantas pagal nustatymą.
  2. *If condition is satisfied* (jeigu sąlyga yra tenkinama).
  3. *Random sample of cases* (atsitiktinė duomenų atranka).
  4. *Based on time or case range* (pagal diapazoną).
  5. *Use filter variable* (pagal filtruojantį kintamąjį).

Duomenų atranka pagal loginę sąlygą

- Pasirinkite duomenų atrankos metodą *If condition is satisfied.*
- Paspauskite mygtuką *If...* Atsivers dialogo langas *Select Cases: If*.
- Pažymėkite kintamąjį *Pakuotės_dydis* ir spragtelėję rodyklę įkelkite jį į sąlygos nustatymo laukelį. Naudodamiesi dialogo lango klaviatūra, užrašykite duomenų atrankos sąlygą (žr. 3.13 pav.).

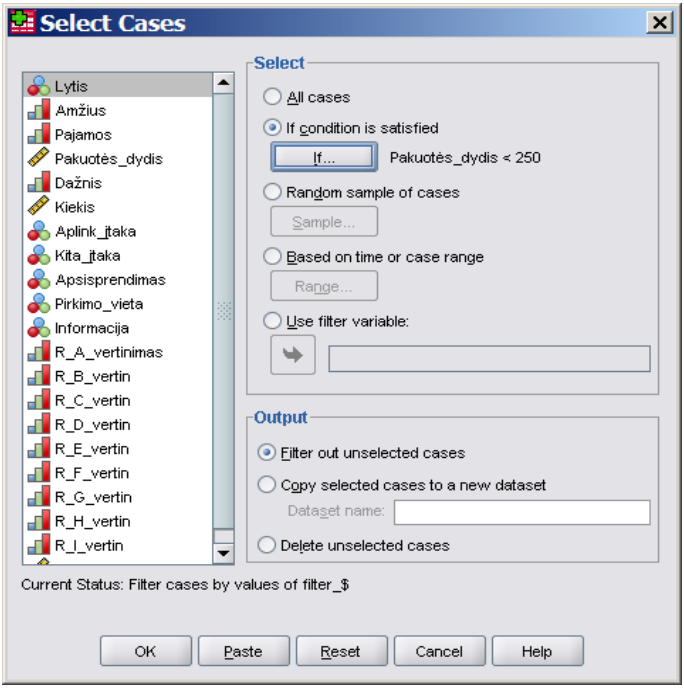


3.13 pav. **Dialogo langas *Select Cases: If***

- Spragtelėkite mygtuką *Continue*. Grįšite į dialogo langą *Select Cases*. Galima pasirinkti 3 rezultatų išvesties (*output*) variantus:
  a) filtruoti sąlygos netenkinančius rezultatus (*Data Editor* lange šie įrašai bus pažymėti įžambiuoju brūkšniu);
  b) nukopijuoti loginę sąlygą tenkinančius duomenis į naująją duomenų bazę;

c) pašalinti loginės sąlygos netenkinančius duomenis (jei rinkmena po tokio filtravimo bus išsaugota – pašalintų duomenų atkurti negalima!).

- Pasirinkite *Filter out unselected cases* išvesties sąlygą (žr. 3.14 pav.).

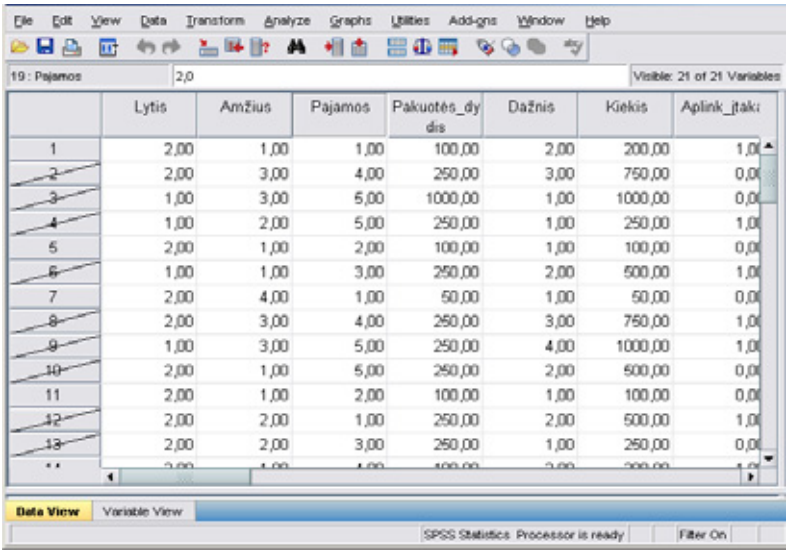


3.14 pav. **Dialogo langas *Select Cases.* Išvestis *Filter out...***

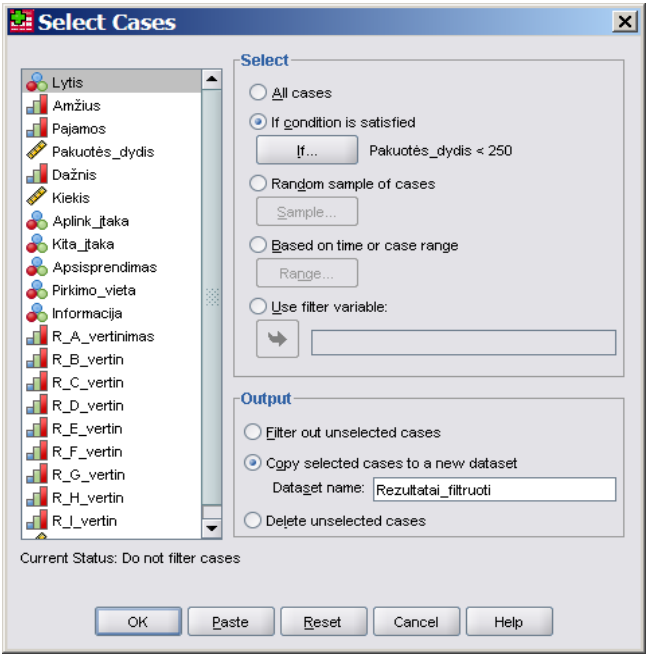
- Spragtelėkite mygtuką *OK* (žr. 3.15 pav.).

Norėdami atsisakyti filtravimo, nurodykite komandą *Data* → *Select Cases...* → *Reset* → *OK*.

Jei filtravimo rezultatus norite išsaugoti naujoje rinkmenoje, dialogo lango *Select Cases* srityje *Output* (išvestis) pasirinkite *Copy selected cases to a new dataset* ir nurodykite naujo failo vardą (žr. 3.16 pav.).



3.15 pav. **Loginio filtravimo rezultatas**

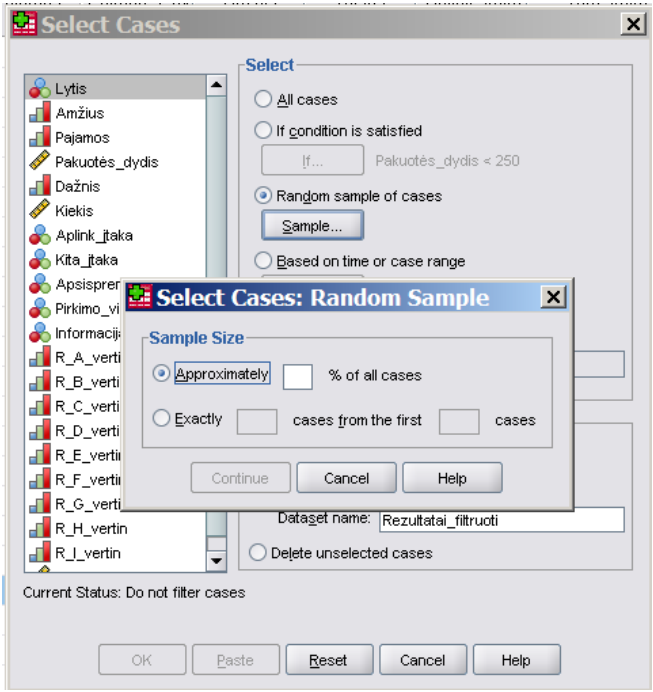


3.16 pav. **Dialogo langas *Select Cases.* Išvestis *Copy selected...***

Atsitiktinė duomenų atranka

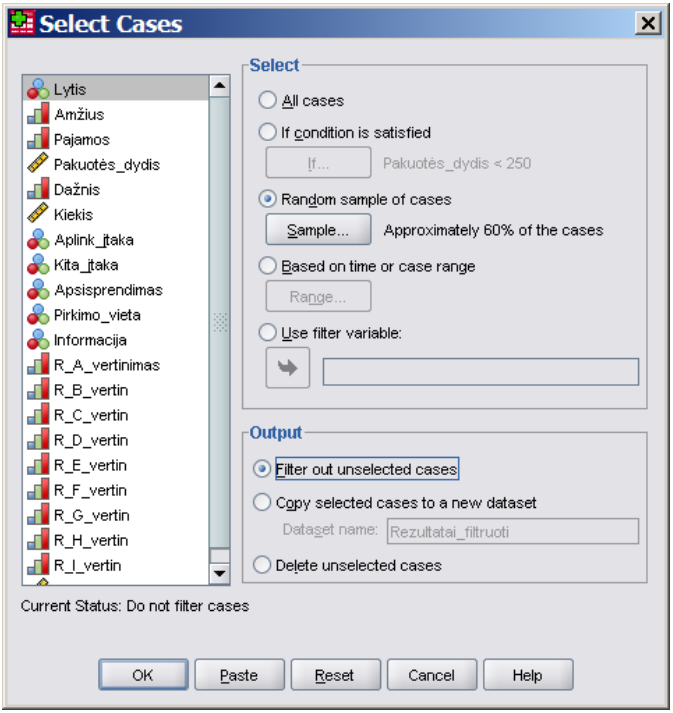
Kai stebėjimų skaičius yra didelis, pirminei hipotezei patikrinti gali būti tikslinga taikyti nedidelę atsitiktinę imtį.

- Atsitiktinei duomenų atrankai iš stebėjimų visumos dialogo lange *Select Cases* pažymėkite variantą *Random sample of cases* ir spragtelėkite mygtuką *Sample*... Atsivers dialogo langelis *Select Cases: Random Sample* (žr. 3.17 pav.), kuriame Jūs galite pasirinkti šiuos imties dydžio nurodymo būdus:
  a) *Approximately* (apytiksliai). Įrašykite į laukelį procentinę atsitiktinės imties dydžio reikšmę – SPSS programa sukurs atsitiktinę imtį, savo dydžiu apytiksliai atitinkančią nurodytą visų stebėjimų dalį.
  b) *Exactly* (tiksliai). Šiuo atveju nurodykite tikslų stebėjimų skaičių atsitiktinėje imtyje. Taip pat nurodykite skaičių stebėjimų, iš kurių bus paimta atsitiktinė imtis. Šis skaičius neturi viršyti visų stebėjimų skaičiaus duomenų rinkmenoje.



3.17 pav. **Dialogo langas *Select Cases: Random Sample***

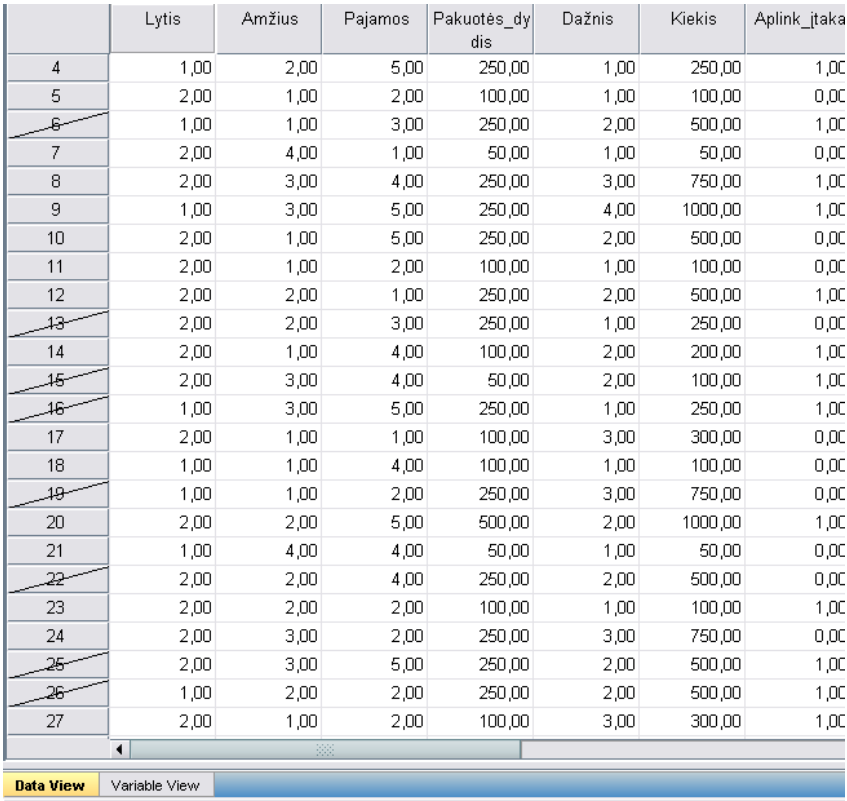
- Dialogo lange *Select Cases: Random Sample* pasirinkite *Approximately* ir įrašykite procentinę atsitiktinės imties reikšmę 60. Spragtelėkite mygtuką *Continue* (žr. 3.18 pav.).



3.18 pav. **Dialogo langas *Select Cases.* Filtravimo būdas *Random Sample***

- Paspauskite mygtuką *OK*.

SPSS programa palieka apytikriai 60 procentų visų stebėjimų dydžio imtį (žr. 3.19 pav.):



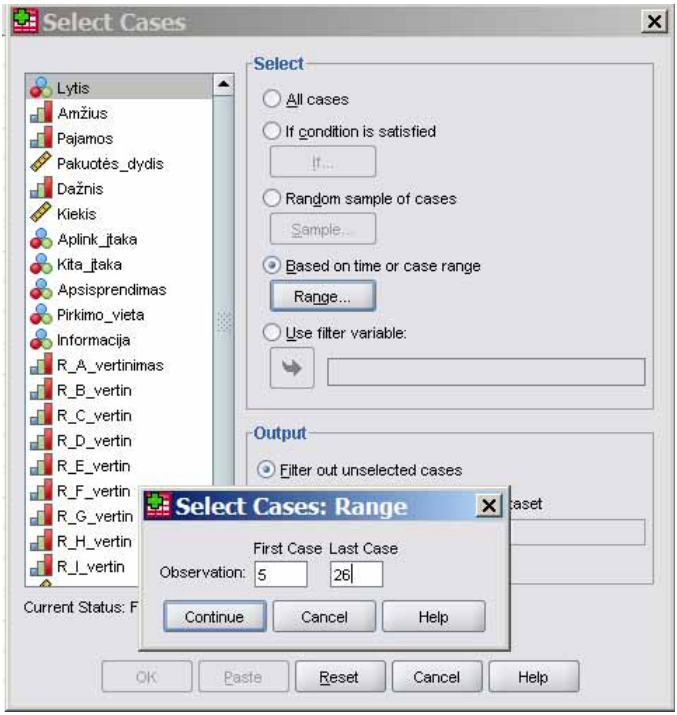
	Lytis	Amžius	Pajamos	Pakuotės_dydis	Dažnis	Kiekis	Aplink_įtaka
4	1,00	2,00	5,00	250,00	1,00	250,00	1,00
5	2,00	1,00	2,00	100,00	1,00	100,00	0,00
6	1,00	1,00	3,00	250,00	2,00	500,00	1,00
7	2,00	4,00	1,00	50,00	1,00	50,00	0,00
8	2,00	3,00	4,00	250,00	3,00	750,00	1,00
9	1,00	3,00	5,00	250,00	4,00	1000,00	1,00
10	2,00	1,00	5,00	250,00	2,00	500,00	0,00
11	2,00	1,00	2,00	100,00	1,00	100,00	0,00
12	2,00	2,00	1,00	250,00	2,00	500,00	1,00
13	2,00	2,00	3,00	250,00	1,00	250,00	0,00
14	2,00	1,00	4,00	100,00	2,00	200,00	1,00
15	2,00	3,00	4,00	50,00	2,00	100,00	1,00
16	1,00	3,00	5,00	250,00	1,00	250,00	1,00
17	2,00	1,00	1,00	100,00	3,00	300,00	0,00
18	1,00	1,00	4,00	100,00	1,00	100,00	0,00
19	1,00	1,00	2,00	250,00	3,00	750,00	0,00
20	2,00	2,00	5,00	500,00	2,00	1000,00	1,00
21	1,00	4,00	4,00	50,00	1,00	50,00	0,00
22	2,00	2,00	4,00	250,00	2,00	500,00	0,00
23	2,00	2,00	2,00	100,00	1,00	100,00	1,00
24	2,00	3,00	2,00	250,00	3,00	750,00	0,00
25	2,00	3,00	5,00	250,00	2,00	500,00	1,00
26	1,00	2,00	2,00	250,00	2,00	500,00	1,00
27	2,00	1,00	2,00	100,00	3,00	300,00	1,00

Data View | Variable View

3.19 pav. **Atsitiktinio filtravimo rezultatas**

Diapazoninė duomenų atranka

- Dialogo lange *Select Cases* pažymėkite variantą *Based on time or case range* ir spragtelėkite mygtuką *Range*...
- Atsivėrusiame dialogo langelyje *Select Cases: Range* laukelyje *First Case* nurodykite diapazono pradžią (duomenų įrašo eilutės numerį), o laukelyje *Last Case* – diapazono pabaigą. Datos ir laiko diapazonai gali būti nurodyti tik tam tikro formato duomenims (žr. 3.20 pav.).



3.20 pav. **Diapazoninė duomenų atranka. Diapazono pasirinkimas**

- Spragtelėkite mygtuką *Continue*, paskui dialogo langelyje *Select Cases* paspauskite mygtuką *OK*.

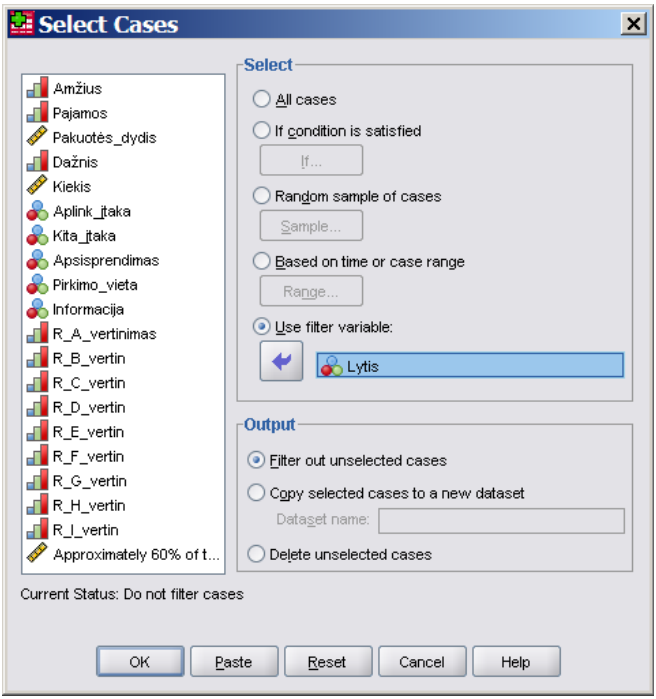
Duomenų atranka pagal filtruojantį kintamąjį

Šią atranką galima taikyti tada, kai norime turėti duomenų bazę be praleistų duomenų. Tarkime, kad mūsų duomenų bazėje 4 ir 6 įrašuose kintamasis *Lytis* turi praleistas (*missing*) reikšmes (žr. 3.21 pav.). Norėdami jas nufiltruoti, taikysime duomenų atranką pagal filtruojantį kintamąjį *Lytis*.

- Dialogo langelyje *Select Cases* pažymėkite variantą *Use filter variable* ir į kintamojo langelį įkelkite kintamąjį *Lytis* (žr. 3.22 pav.).

	Lytis	Amžius	Pajamos	Pakuotės_dydis	Dažnis
1	2,00	1,00	1,00	100,00	2,00
2	2,00	3,00	4,00	250,00	3,00
3	1,00	3,00	5,00	1000,00	1,00
4	.	2,00	5,00	250,00	1,00
5	2,00	1,00	2,00	100,00	1,00
6	.	1,00	3,00	250,00	2,00
7	2,00	4,00	1,00	50,00	1,00
8	2,00	3,00	4,00	250,00	3,00
9	1,00	3,00	5,00	250,00	4,00
10	2,00	1,00	5,00	250,00	2,00
11	2,00	1,00	2,00	100,00	1,00
12	2,00	2,00	1,00	250,00	2,00
13	2,00	2,00	3,00	250,00	1,00
14	2,00	1,00	4,00	100,00	2,00
15	2,00	3,00	4,00	50,00	2,00

3.21 pav. **Duomenų bazė su praleistomis reikšmėmis**



3.22 pav. **Duomenų atranka pagal filtruojantį kintamąjį**

- Spragtelėkite mygtuką *OK* (žr. 3.23 pav.).

	Lytis	Amžius	Pajamos	Pakuotės_dydis	Dažnis
1	2,00	1,00	1,00	100,00	2,00
2	2,00	3,00	4,00	250,00	3,00
3	1,00	3,00	5,00	1000,00	1,00
4	.	2,00	5,00	250,00	1,00
5	2,00	1,00	2,00	100,00	1,00
6	.	1,00	3,00	250,00	2,00
7	2,00	4,00	1,00	50,00	1,00
8	2,00	3,00	4,00	250,00	3,00
9	1,00	3,00	5,00	250,00	4,00
10	2,00	1,00	5,00	250,00	2,00
11	2,00	1,00	2,00	100,00	1,00
12	2,00	2,00	1,00	250,00	2,00
13	2,00	2,00	3,00	250,00	1,00
14	2,00	1,00	4,00	100,00	2,00
15	2,00	3,00	4,00	50,00	2,00

3.23 pav. **Filtravimo pagal filtruojantį kintamąjį rezultatas**

3.3. SPSS paketo sintaksės redaktorius

SPSS sintaksės redaktorius leidžia atlikti skaičiavimus nenaudojant valdymo skydelio mygtukų (meniu komandų) ir dialogo langų, bet taikant specialią SPSS veiksmų komandų kalbą. Komandos užrašomos vadinamosiose *Syntax* rinkmenose, kurios iš esmės yra programos, užrašytos SPSS programavimo kalba.

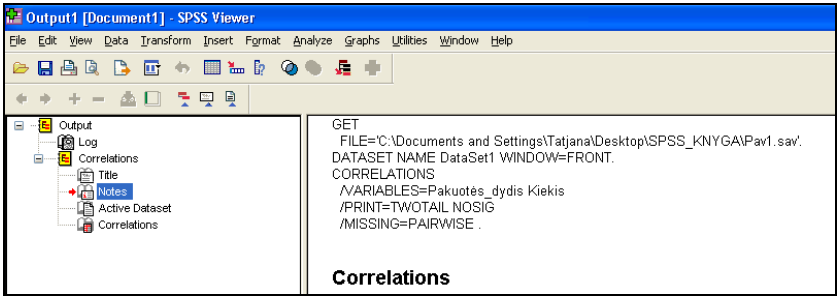
Visos sintaksės komandos skirstomos į kategorijas:

- SPSS komanda;
- pagalbinė SPSS komanda (patikslina SPSS komandos veiksmą, į vieną komandą gali būti įtraukta keletas pagalbinių komandų);

- specifikacijos (duomenys, kurie papildo SPSS komandą arba pagalbinę SPSS komandą). Specifikacijos gali būti sudarytos iš reikšminių žodžių, skaičių, aritmetinių operacijų, kintamųjų vardų ir specialių skiriamųjų ženklų;
- reišminiai žodžiai.

Taikant *Syntax* programas, galima automatizuoti dažnai kartojamus veiksmus. Gerai išmanant SPSS programavimo kalbą, galima kurti papildomų statistinės analizės metodų, kurių neįmanoma atlikti per meniu komandas, programas.

Paprasčiausias būdas sukurti *Syntax* programą – nukopijuoti ją iš *Output-SPSS Viewer* lango, kur ji savaime įterpiama atlikus skaičiavimus meniu. Šią programą vėliau galima redaguoti keičiant komandų parametrus. Visų SPSS analizių rezultatų (*Output*) failuose automatiškai įterpiamos pastabos (*Notes*). Juose surašoma visa informacija, apibūdinanti konkrečią analizę ir leidžianti ją pakartoti. Šios pastabos būna paslėptos, atsiveria du kartus spragtelėjus punktą *Notes* kairėje lango pusėje (žr. 3.24 pav.). SPSS skaičiavimo pastabų (*Notes*) pavyzdys pateiktas 3.25 pav.



3.24 pav. ***Output-SPSS Viewer* langas**

SPSS sintaksės komandų kalba yra sudėtinga, todėl šiame vadovėlyje išsamiai nenagrinėjama, tačiau kiekvienas tiriamas analizės pavyzdys papildytas SPSS sintaksės programa. Komandų sintaksės aprašymas pateikiamas surinkus komandą *Help → Command Syntax Referente.* Yra SPSS komandų sintaksės taisyklės:

Correlations

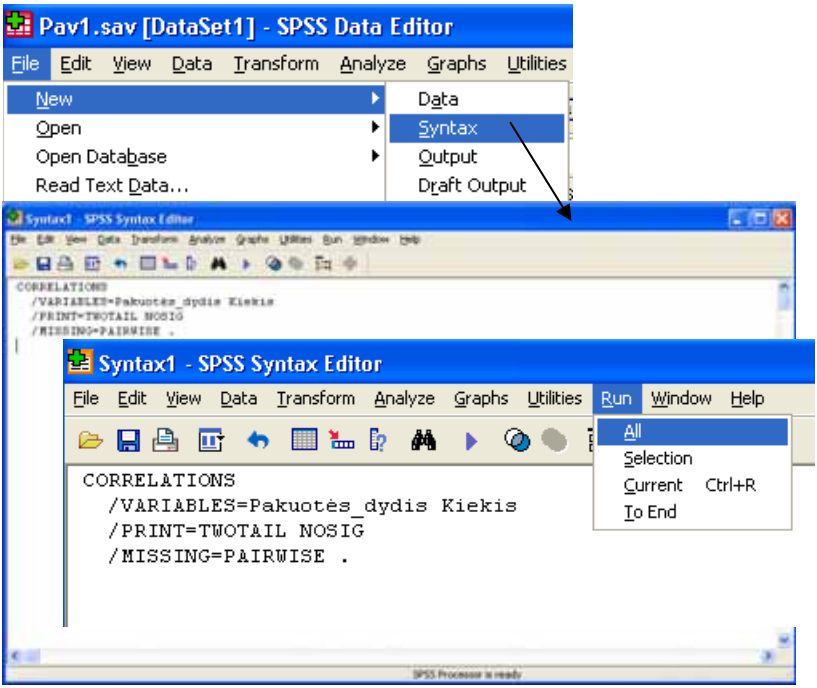
Notes

Output Created		07-AUG-2010 23:12:20
Comments		
Input	Data	C:\Documents and Settings\Tatjana\Desktop\SPSS_KNYGA\Pav1.sav
	Active Dataset	DataSet1
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	60
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing.
	Cases Used	Statistics for each pair of variables are based on all the cases with valid data for that pair.
Syntax		CORRELATIONS /VARIABLES=Pakuotės_dydis Kiekis /PRINT=TWOTAIL NOSIG /MISSING=PAIRWISE .
Resources	Elapsed Time	0:00:00,05
	Processor Time	0:00:00,05

3.25 pav. **SPSS skaičiavimo pastabų (*Notes*) pavyzdys (Pirsono koreliacijos koeficiento skaičiavimas)**

- Kiekviena komanda pradedama iš naujos eilutės ir baigiama tašku.
- Komanda gali užimti keletą eilučių. Kiekviena vienos komandos eilutė pradedama nuo tarpo.
- Pagalbinės komandos atskiriamos ženklu „/“.
- Eilutės ilgis ne daugiau kaip 80 ženklų.
- Dešimtainė trupmena atskiriama tašku.

Sintaksės redaktoriaus dialogo langas *SPSS Syntax Editor* iškviečiamas komanda *File → New → Syntax* (žr. 3.26 pav.). Šiame lange galima rinkti ir vykdyti SPSS komandas.



3.26 pav. **SPSS sintaksės redaktoriaus dialogo langas**

SPSS sintaksės lange (žr. 3.26 pav.) įrašyta Pirsono koreliacijos koeficiento skaičiavimo komanda:

```
CORRELATIONS
    /VARIABLES=Pakuotės_dydis Kiekis
    /PRINT=TWOTAIL NOSIG
    /MISSING=PAIRWISE .
```

Čia: CORRELATIONS – komanda,
VARIABLES, PRINT ir MISSING – pagalbinės komandos,
TWOTAIL NOSIG ir PAIRWISE – reikšminiai žodžiai,
Pakuotės_dydis, Kiekis – kintamųjų vardai.

Programinis failas vykdomas komanda *Run.* Komandos meniu leidžia pasirinkti viso failo vykdymą (*All*) arba jo dalies.

Jei SPSS duomenų failas yra aktyvus, SPSS sintaksės lange galima iš karto nurodyti vykdomas komandas. Priešingu atveju programą reikia pradėti nurodant duomenų failo iškvietimo komandą:

```
GET
  FILE='C:\Documents and Settings\Desktop\SPSS_KNYGA\Pav1.
sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
```

SPSS sintaksės programos failą galima išsaugoti.

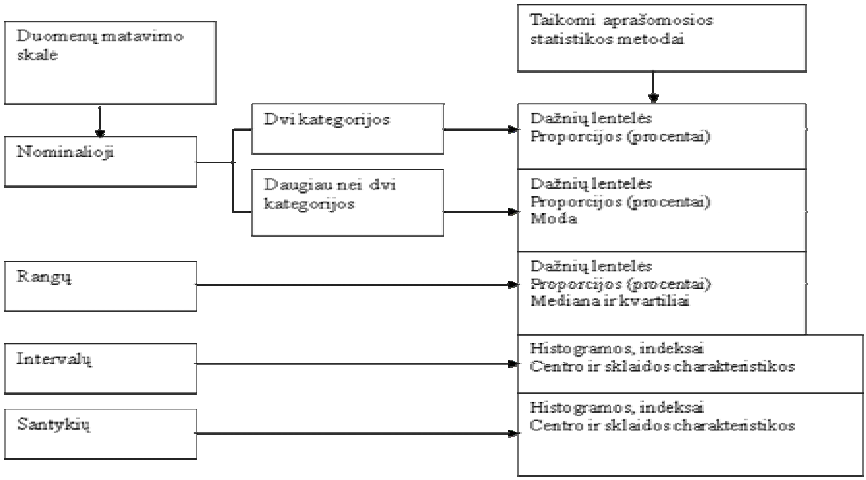
3.4. Aprašomoji statistika

Prieš atliekant nuodugnesnę surinktų duomenų statistinę analizę, pirmiausia reikia panagrinėti kiekvieną kintamąjį (požymį) atskirai. Statistinė analizė, kai kiekvienas kintamasis nagrinėjamas atskirai, vadinama *vienmate*. Šiai analizei atlikti taikomi *aprašomosios statistikos* metodai.

Dažnių lentelės sudaromos kintamiesiems, kurie matuojami nominaliaja arba, esant nedideliam kategorijų skaičiui, rangų skale.

Pradinei duomenų analizei, be grupuotų dažnių lentelių (vienmačių), dažnai dar sudaromos porinės (ir daugiau kintamųjų) dažnių lentelės. Porinės dažnių lentelės skirtos tarp kintamųjų esamiems ryšiams pavaizduoti. Tokios dažnių lentelės paprastai sudaromos nominaliniams ir ranginiams kintamiesiems su palyginus mažu stebimų reikšmių (kategorijų) skaičiumi. Porinėms lentelėms sudaryti kiekybiniai kintamieji nenaudojami. Jeigu į tokią lentelę būtina įtraukti kiekybinį kintamąjį (pavyzdžiui, pajamas), tai pirmiausia šį kintamąjį reikia perkoduoti taip, kad jo reikšmės apimtų keletą intervalų (pavyzdžiui, mažas pajamas, vidutines, dideles).

3.1 lentelė. **Aprašomosios statistikos metodų taikymo priklausomybė nuo kintamųjų tipo**



Aprašomoji statistika, be dažnių lentelių ir diagramų, nominaliniams kintamiesiems jokių kitokių metodų nesiūlo. Ranginiams kintamiesiems ir kiekybiniams kintamiesiems, išmatuotiems intervalų ar santykių skalėmis, bet neturintiems normaliojo skirstinio, šalia dažnių lentelių ir diagramų dažniausiai dar skaičiuojama vidurinė (centrinė) atsitiktinio dydžio reikšmė (mediana) ir abu kvartiliai. Kvartiliais vadinami skaičiai, dalijantys variacinę eilutę į keturias lygias dalis (variacinė eilutė – duomenų eilutė, išdėstyta nemažėjimo tvarka. Jei duomenys išdėstyti didėjimo ar mažėjimo tvarka, turime ranžiruotą eilutę). Kintamiesiems, išmatuotiems intervalų ar santykių skalėmis ir turintiems normalųjį skirstinį, skaičiuojamos dažnių lentelės, centro ir sklaidos charakteristikos, brėžiamos diagramos (histogramos, linijinės diagramos ir panašiai).

Kadangi dauguma metodų, taikomų intervalų ir santykių skalės kintamųjų analizei, remiasi prielaida, kad šie kintamieji turi normalųjį skirstinį, pradinei duomenų analizei priskiriami paprasčiausi normališkumo tikrinimo kriterijai.

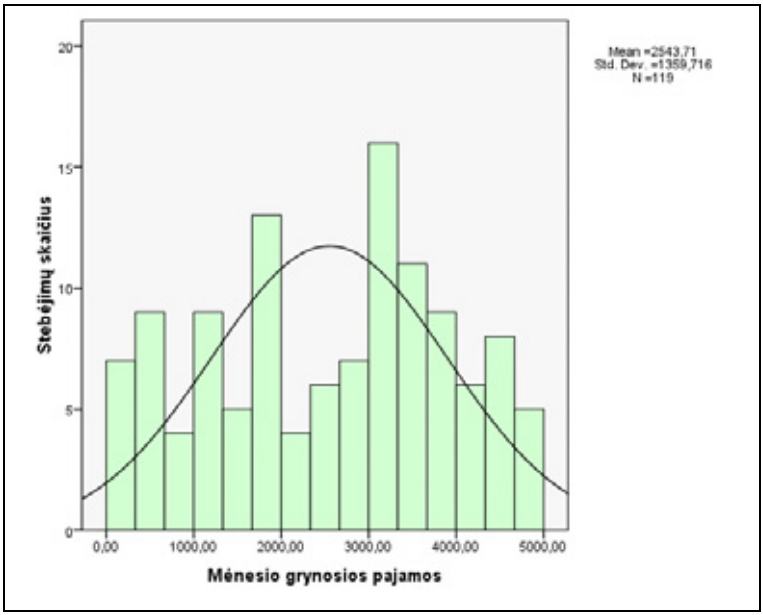
Esant normaliajam skirstiniui dauguma reikšmių grupuojasi apie vidurkį, o tolstant nuo jo į abi puses stebėjimų tankis tolygiai mažėja.

Siekiant įsitikinti, ar imtis turi normalųjį skirstinį, taikomi du paprasti grafiniai būdai: histograma ir normalumo tiesė.

Histograma braižoma pasirinkus šį diagramų tipą ir papildomai brėžiant normaliojo skirstinio kreivę.

Pavyzdys. Apklausiant gauti respondentų atsakymai apie jų gaunamas pajamas – intervalinis kintamasis *Mėnesio grynosios pajamos*. SPSS nubrėšime jo histogramą (žr. 3.27 pav.).

- Atverkite bylą su analizuojamais duomenimis.
- Nurodykite komandas *Graphs* → *Legacy Dialogs* → *Histogram*…
- Įkelkite kintamąjį *Mėnesio grynosios pajamos* ir paspauskite mygtuką *OK*.



3.27 pav. **Dažnių histograma**

Brėžiant histogramą, svarbu parinkti tinkamą grupių skaičių. Kompiuterinės programos dažniausiai automatiškai parenka grupių skaičių, kurį galima keisti.

Grupių skaičius nustatomas pagal bendras taisykles. Beveik į kiekvieną grupę turi patekti vidutiniškai ne mažiau kaip penki duomenys, o grupių turėtų būti ne mažiau kaip 5 ir ne daugiau nei 15.

Grupės turi būti vienodo ilgio, kad jas būtų galima lyginti tarpusavyje. Kai duomenų daug (iki 100), grupių skaičių n galima nustatyti pagal Sterdžeso formulę

$$n = 1 + 3{,}322 * \log N, \qquad (3.1)$$

čia: log – dešimtainis logaritmas,
N – visų duomenų skaičius.

Brėžiant normalumo tiesę (žr. 3.28 pav.) (statistinėse programose vadinamą *Normal probability plot* arba *Normal Q-Q plot* (SPSS)) duomenys keičiami taip, kad tuo atveju, kai duomenys turi normalųjį skirstinį, jie sudarytų tiesę. SPSS programoje:

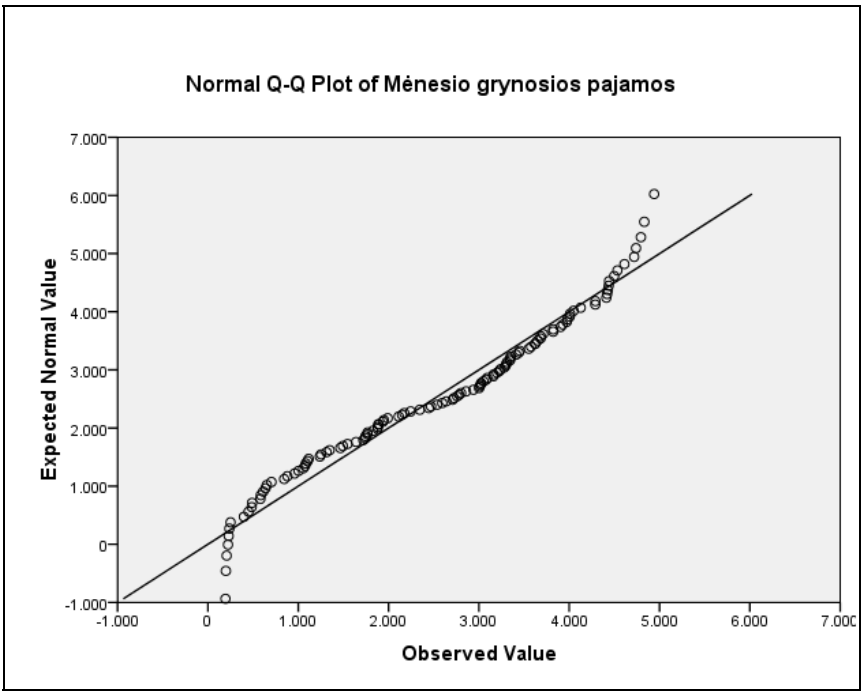
- Atverkite bylą su analizuojamais duomenimis.
- Nurodykite komandas *Analyze* → *Descriptive Statistics* → *Q-Q Plots*…
- Įkelkite kintamąjį *Mėnesio grynosios pajamos* ir paspauskite mygtuką *OK*.

Jei duomenų skirstinys nėra normalusis, taškai nuo tiesės skirsis. Iš vaizdo, pateikto 3.28 pav., galima spręsti, ar duomenys turi normalųjį skirstinį. Netgi nedidelė darbo su realiais duomenimis patirtis leidžia žmogui gana užtikrintai atskirti duomenų nukrypimą nuo normaliojo skirstinio. Mūsų atveju taškai nėra tiesėje – skirstinys nėra normalusis.

Jei netenkina vizuali kontrolė, patikrinti galima taikant Kolmogorovo–Smirnovo ir Šapiro–Vilko (kai duomenų yra mažiau nei 50) kriterijus.

- Ativerkite bylą su analizuojamais duomenimis.
- Nurodykite komandas *Analyze* → *Descriptive Statistics* → *Explore*…
- Įkelkite kintamąjį *Mėnesio grynosios pajamos.*
- Paspauskite mygtuką *Plots*…
- Įsitikinkite, kad pasirinkta *Normality plots with tests*. Paspauskite mygtukus *Continue* ir *OK*.

Rezultatą žiūrėkite 3.2 lentelėje.



3.28 pav. **Normalumo tiesė**

3.2 lentelė. **Skirstinio normalumo tikrinimo rezultatai**

Tests of Normality						
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Men_pajamos	,094	119	,011	,955	119	,001
a. Lilliefors Significance Correction						

Kadangi stebimasis reikšmingumo lygmuo *Sig*. (angl. *Significance* – reikšmingumas) yra mažesnis už reikšmingumo lygmenį 0,05, galima daryti išvadą, kad skirstinys nėra normalusis.

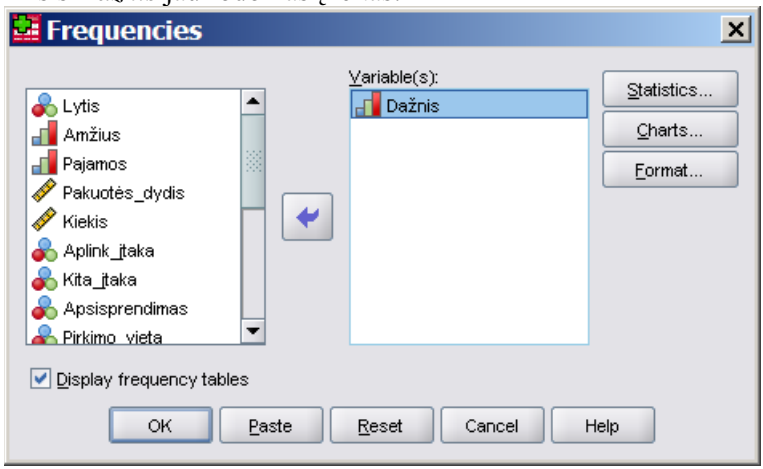
3.4.1. Vienmatė dažnių analizė

Dažnių analizės rezultatai ar grafinė medžiaga yra pakankami daugeliui kintamųjų apibūdinti. Tai pirmasis duomenų statistinės analizės etapas.

Vienmatė dažnių lentelė sudaroma nagrinėjant pavadinimų (nominaliuosius) ir ranginius kintamuosius. Šia lentele naudojamasi tiriant, kaip atskiro kintamojo reikšmės pasiskirsčiusios imtyje.

Pavyzdys. Apklausoje buvo tirta, kaip dažnai per mėnesį respondentai perka kavą. Kintamasis *Dažnis*. SPSS gausime šio kintamojo dažnių lentelę:

- Atverkite bylą su analizuojamais duomenimis.
- Nurodykite komandas *Analyze* → *Descriptive Statistics* → *Frequencies...*
- Dialogo langelyje *Frequencies* pažymėkite analizuojamą kintamąjį *Dažnis* ir spragtelėję mygtuką su rodyklės ženklu įkelkite jį į analizuojamų kintamųjų sąrašą *Variable*(*s*). 3.29 pav. kintamasis *Dažnis* jau rodomas įkeltas:



3.29 pav. **Dialogo langelis *Frequencies***

- Įsitikinę, kad pažymėtas laukelis *Display frequency tables* (rodyti dažnių lenteles), spragtelėkite mygtuką *OK*. Rezultatus matome rodinio (*Viewer*) lange (žr. 3.30 pav.).

Statistics

Dažnis

N	Valid	45
	Missing	0

Dažnis

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1,00	14	31,1	31,1	31,1
	2,00	18	40,0	40,0	71,1
	3,00	12	26,7	26,7	97,8
	4,00	1	2,2	2,2	100,0
	Total	45	100,0	100,0	

3.30 pav. **Dažnių lentelės išvestis**

Kintamasis *Dažnis* vaizduoja respondentų atsakymus į klausimą, kiek kartų per mėnesį jie perka kavą. Kintamojo reikšmės 1, 2, 3, 4.

Kiekviena dažnių lentelės eilutė aprašo vieną kintamojo reikšmę. Pirmajame stulpelyje pateikiamos visos galiojančios kintamojo reikšmės. Antrajame stulpelyje *Frequency* (dažnis) pateiktas kiekvieno atsakymo į anketos klausimą varianto dažnis, t. y. kiek respondentų pirko kavą per mėnesį vieną kartą (1), kiek du (2) ir t. t. Trečiajame stulpelyje *Percent* pateiktas kiekvieno atsakymo procentinis dažnis, kuris gaunamas kiekvieno atsakymo variantų skaičių dalijant iš stebėjimų skaičiaus, t. y. bendro respondentų skaičiaus. Ketvirtajame stulpelyje *Valid Percent* pateikiamas atsakiusiųjų procentinis dažnis. Kai nėra praleistų duomenų trečiojo ir ketvirtojo stulpelių reikšmės sutampa. Penktajame stulpelyje *Cumulative Percent* pateikiamas sukauptasis dažnis, kuris gaunamas sudedant prieš tai buvusias reikšmes.

Pavyzdžiui, bendras respondentų, atsakiusiųjų, kad per mėnesį perka kavą vieną ar du kartus, procentas yra 71,1 procento.

Atliktos procedūros sintaksė:

```
GET
FILE='C:\...\Pav1_kava.sav'.
FREQUENCIES VARIABLES=Amžius
/ORDER=ANALYSIS.
```

3.4.2. Porinės dažnių lentelės

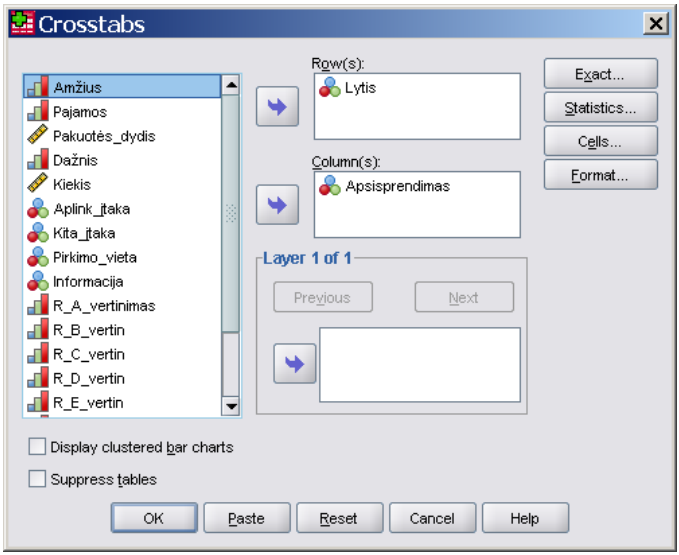
Porinės dažnių lentelės atskleidžia tarp kintamųjų esančius ryšius. Tokios dažnių lentelės paprastai sudaromos nominaliniams ir ranginiams kintamiesiems su palyginus mažu stebimų reikšmių (kategorijų) skaičiumi. Porinėms lentelėms sudaryti kiekybiniai kintamieji nenaudojami. SPSS tuo tikslu naudojamos vadinamosios požymių dažnių lentelės (*Crosstabs*) ir didelis testų pasirinkimas priklausomybės laipsniui tarp nagrinėjamų kintamųjų įvertinti.

Jeigu į tokią lentelę būtina įtraukti kiekybinį kintamąjį (pajamas ir panašiai), tai pirmiausia šio kintamojo reikšmes reikia suskirstyti į keletą intervalų (mažos pajamos, vidutinės, didelės, labai didelės).

Paprasčiausia porinės dažnių lentelės forma yra 2 × 2 lentelė, gaunama tuo atveju, kai tiriami du kintamieji, kurių kiekvienas gali įgyti tik dvi reikšmes (dichotominiai kintamieji).

Pavyzdys. Apklausoje buvo tirta, kokiais kriterijais remiantis vyrai ir moterys perka kavą. Buvo galima rinktis iš penkių kriterijų: „rūšis“, „žema kaina“, „patraukli pakuotė“, „kokybė“, „kainos ir kokybės santykis“. Pirmasis kintamasis (respondentų lytis) pavadintas *Lytis* ir jam suteiktas skaitmeninis formatas (kintamojo reikšmės 1 ir 2), o antrasis kintamasis pavadintas *Apsisprendimas.* Jam taip pat suteiktas skaitmeninis formatas. Šio kintamojo reikšmės 1, 2, 3, 4, 5.

- Atverkite bylą su analizuojamais duomenimis.
- Nurodykite komandas *Analyze* → *Descriptive Statistics* → *Crosstabs...*
- Dialogo lange *Crosstabs* pažymėkite kintamąjį *Lytis* ir spragtelėję mygtuką su rodyklės ženklu įkelkite jį į laukelį *Row*(*s*). Šis kintamasis bus dažnių lentelės eilutėse. Atitinkamai kintamąjį *Apsisprendimas* įkelkite į laukelį *Column*(*s*). Šis kintamasis bus dažnių lentelės stulpeliuose (žr. 3.31 pav.).



3.31 pav. **Dialogo langas *Crosstabs***

- Spragtelėkite mygtuką *OK*.

Sąraše *Row*(*s*) yra 2 kintamieji, *Column*(*s*) – 5 kintamieji, gausime 2 * 5 = 10 dažnių lentelę.

3.3 lentelė. **Standartinio formato požymių dažnių lentelė**

Lytis * Apsisprendimas Crosstabulation							
Count							
		Apsisprendimas					**Total**
		1,00	**2,00**	**3,00**	**4,00**	**5,00**	
Lytis	**1,00**	**3**	**2**	**3**	**3**	**4**	**15**
	2,00	**8**	**7**	**5**	**6**	**4**	**30**
Total		**11**	**9**	**8**	**9**	**8**	**45**

Ši lentelė gaunama pagal SPSS standartinius nustatymus: rodomos absoliučios stebėjimų skaičiaus reikšmės. Dialogo lange *Crosstabs*

esančiais mygtukais *Exact…*, *Statistics…*, *Cells…* arba *Format…* galima lentelėje parodyti daugiau informacijos: procentinę dažnių išraišką, tikėtinas dažnių reikšmes ir kt.

Palyginti tikėtinas dažnių reikšmes su stebėtomis (eksperimentinėmis) galima daryti tam tikras išvadas apie vieno kintamojo reikšmių priklausomybę nuo kito kintamojo. Tai bus nagrinėjama vėlesniuose skyriuose.

Atliktos procedūros sintaksė:

```
GET
FILE='C:\...\Pav1_kava.sav'.
CROSSTABS
/TABLES=Lytis BY Apsisprendimas
/FORMAT=AVALUE TABLES
/CELLS=COUNT
/COUNT ROUND CELL.
```

3.4.3. Pagrindiniai grafikai: histogramos, stulpelinės ir skritulinės diagramos

Statistinės programos siūlo įvairiausių tipų diagramas. Histogramos (stulpelinės diagramos) rodo duomenų pokyčius per tam tikrą laikotarpį arba vaizduoja lyginamus elementus. Linijinės diagramos gali vaizduoti tęstinius duomenis per tam tikrą laikotarpį, nustatytą įprastoje skalėje, todėl puikiai tinka norint vaizduoti duomenų kitimo tendencijas lygiais intervalais. Skritulinės diagramos rodo atskirų duomenų sekos elementų dydžius procentais nuo visų elementų sumos.

Standartiniai grafikai sudaromi naudojantis vartotojo aplinkos meniu procedūromis. Toks diagramų sudarymo būdas yra labai lengvas. Pasirinkus norimą diagramos tipą, nurodomas kintamojo vardas ir pagal iš anksto nustatytą schemą iš kintamojo reikšmių yra sudaroma diagrama. Sudarytą diagramą lengva redaguoti. Redagavimo principas daugelyje šiuolaikinių statistinių programų yra vienodas. Norint pakeisti kurį nors užrašą, liniją ar spalvą tereikia du kartus spustelėti kairįjį pelės klavišą ant norimo redaguoti elemento.

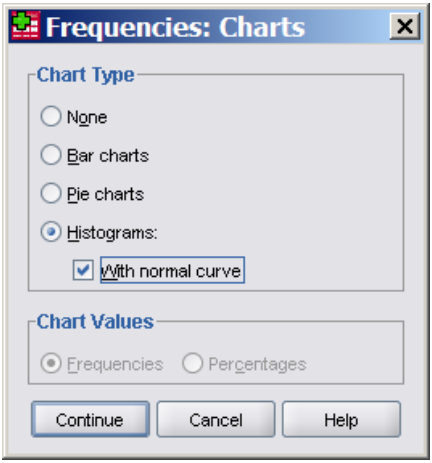
Dažnių lentelių rezultatus galima pavaizduoti grafiškai. Dažniai vaizduojami stulpeliais, kurių viršuje galima užrašyti apskaičiuotas daž-

nių reikšmes. Tokia dažnių diagrama vadinama stulpeline diagrama arba histograma.

Histogramomis vaizduojamas intervalų skalės kintamųjų pasiskirstymas. Kintamojo reikšmės yra suskirstomos į intervalus, nustatomas į šiuos intervalus patenkančių reikšmių skaičius (dažnis) ir rodikliai pavaizduojami stačiakampių stulpelių, kurių kiekvieno aukštis yra proporcingas atitinkamo intervalo duomenų dažniui, seka. Pagal pradinius nustatymus intervalų skaičių ir plotį programa parenka automatiškai, tačiau šiuos parametrus gali nustatyti ir vartotojas.

Pavyzdys. Apklausoje buvo tirta, kaip dažnai per mėnesį respondentai perka kavą. Kintamasis *Dažnis*. SPSS pavaizduosime šį kintamąjį grafiškai.

- Atveriame bylą su analizuojamais duomenimis.
- Nurodome komandas *Analyze* → *Descriptive Statistics* → *Frequencies...*
- Dialogo langelyje *Frequencies* pažymime analizuojamą kintamąjį *Dažnis* ir paspaudę mygtuką su rodyklės ženklu įkeliame jį į analizuojamų kintamųjų sąrašą *Variable*(*s*).
- Spragtelėkite mygtuką *Charts...*(diagramos). Atsivers dialogo langelis *Frequencies: Charts* (žr. 3.32 pav.)



3.32 pav. **Dialogo langelis *Frequencies: Charts***

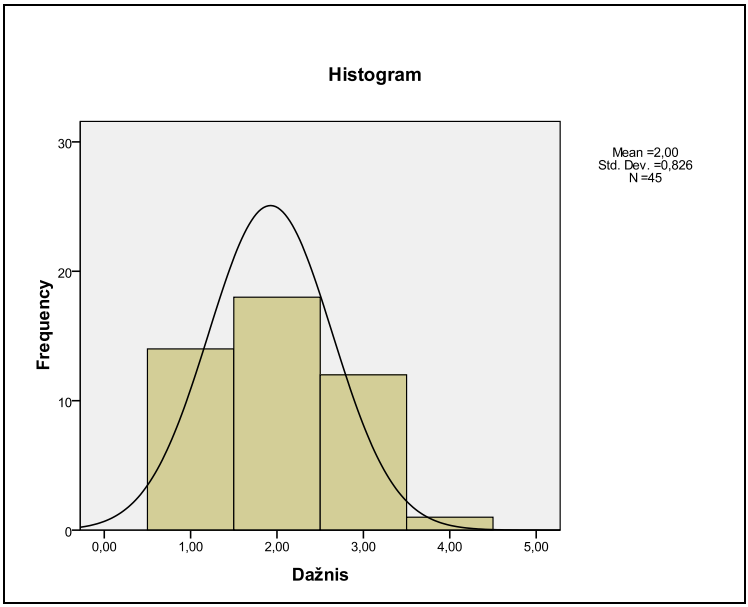
Statistics

Dažnis

N	Valid	45
	Missing	0

Dažnis

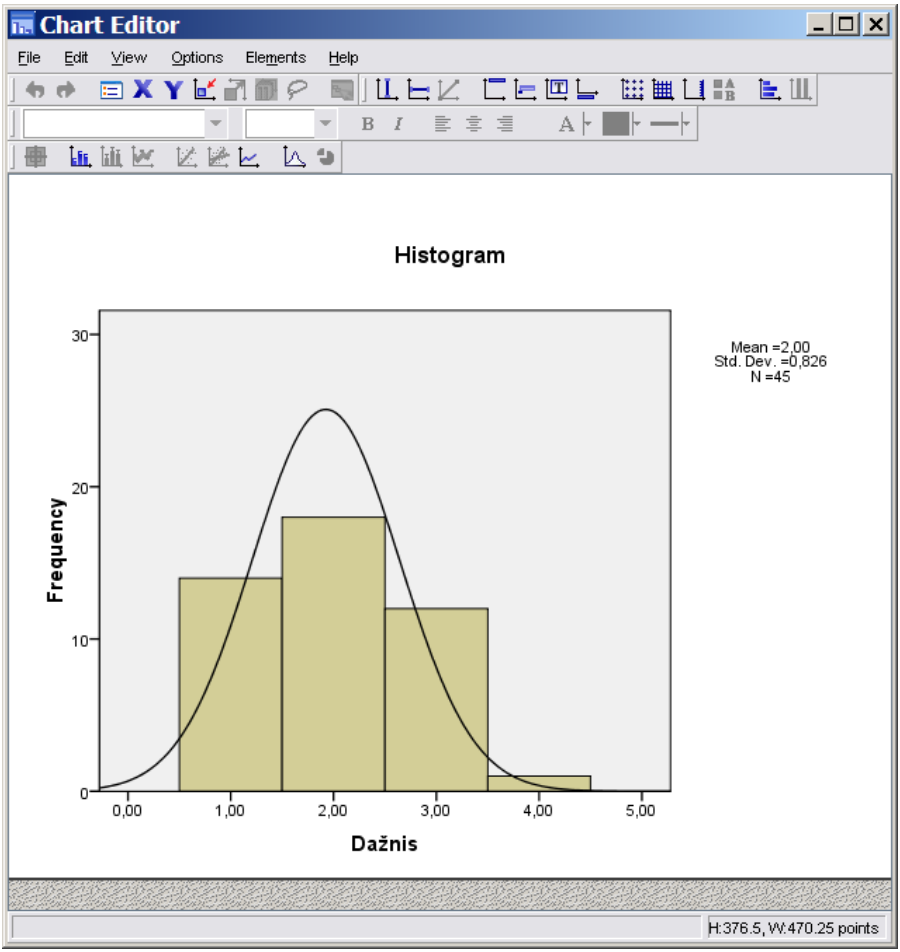
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1,00	14	31,1	31,1	31,1
	2,00	18	40,0	40,0	71,1
	3,00	12	26,7	26,7	97,8
	4,00	1	2,2	2,2	100,0
	Total	45	100,0	100,0	



3.33 pav. **Histograma *Viewer* lange**

Grupėje *Chart Type* (diagramos tipas) galite pasirinkti diagramos tipą:

- *Bar charts* (stulpelinės diagramos).
- *Pie charts* (skritulinės diagramos).
- *Histograms* (histogramos).



3.34 pav. **Diagramų redaktoriaus *Chart Editor* langas**

Pasirinkę stulpelinę diagramą, galite nurodyti, kokie dydžiai turi būti atidėti y ašyje: *Frequences* (dažniai) ar *Percentages* (procentai). Pasirinkę histogramą, galite pažymėti laukelį *With normal curve* (su normaliojo skirstinio kreive).

- Pasirinkite diagramos tipą *Histograms*, pažymėkite laukelį *With normal curve*, spragtelėkite dialogo langelio *Frequencies: Charts* mygtuką *Continue*, tada paspauskite dialogo langelio *Frequencies* mygtuką *OK*.

Rodinio (*Viewer*) lange pasirodys Jūsų pasirinkta diagrama (žr. 3.33 pav.).

Jeigu norite redaguoti diagramą pagal savo poreikius, dukart spragtelėkite pele bet kurioje diagramos vietoje. Diagrama atsidurs diagramų redaktoriaus lange (žr. 3.34 pav.).

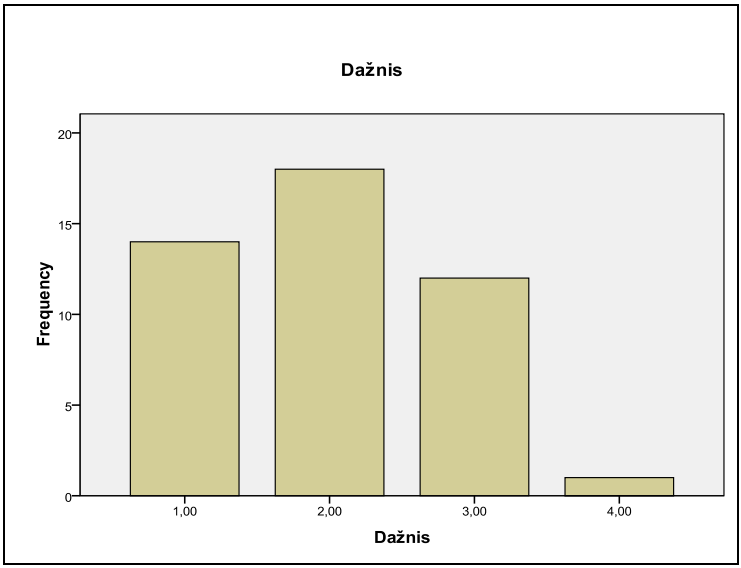
Kaip matome, pasikeičia meniu ir įrankių juostos – vyrauja diagramų redagavimo komandos ir įrankiai. Jūs galite keisti kiekvieną diagramos elementą ir pasirinkti kitą diagramos tipą. Dažniausiai redaguojami:

- Diagramos ir ašių pavadinimai.
- Legenda.
- Diagramos tinklelis ir ašių formatai.
- Duomenų sekų formatai.
- Pastabų tekstas.
- Vidiniai ir išoriniai rėmeliai.

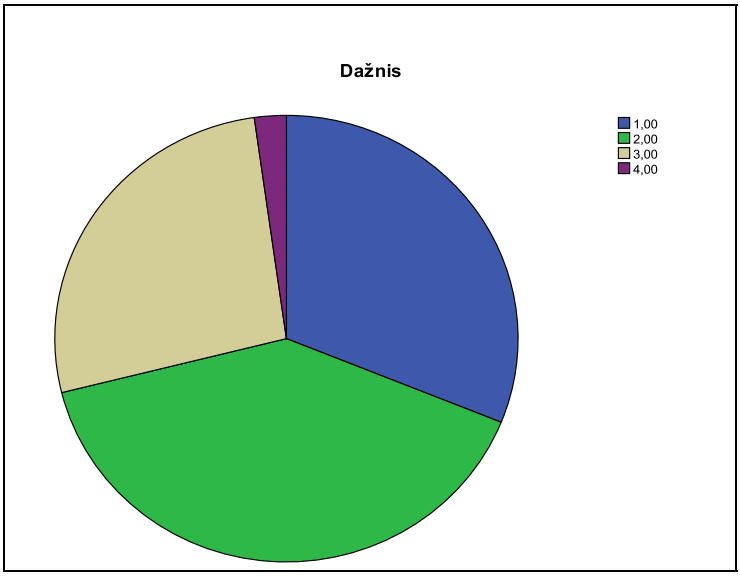
Padaryti diagramų pakeitimai išsaugomi, kai uždarote diagramų redaktoriaus langą.

Duomenys, matuojami nominalioje skalėje, dažniausiai vaizduojami stulpelinėmis (*bar charts*) ir skritulinėmis (*pie charts*) diagramomis. Pastarosios ypač tinka, kai požymio reikšmių skaičius yra nedidelis.

Tą patį kintamąjį *Dažnis* pavaizduosime stulpelinėje (žr. 3.35 pav.) ir skritulinėje (žr. 3.36 pav.) diagramose. Dialogo langelyje *Frequencies: Charts* grupėje *Chart Type* atitinkamai pasirenkame *Bar charts* arba *Pie charts*



3.35 pav. **Stulpelinė diagrama**



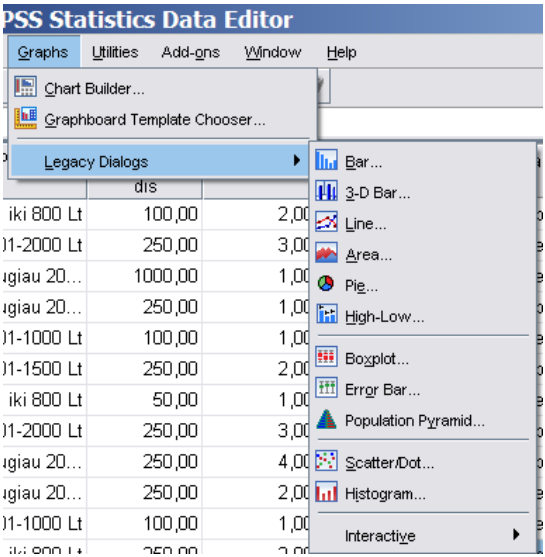
3.36 pav. **Skritulinė diagrama**

Atliktos procedūros sintaksė:

```
GET
FILE='C:\...\Pav1_kava.sav'.
FREQUENCIES VARIABLES=Dažnis
  /HISTOGRAM NORMAL
  /ORDER=ANALYSIS.
```

Vienas iš SPSS paketo pranašumų – didelė grafikų ir diagramų įvairovė. Grafikai gali būti sudaromi tiek taikant grafikų meniu, tiek statistinių metodų meniu komandas. Tokius pat grafikus (histogramą, stulpelinę ar skritulinę diagramą) pavyzdyje analizuojamam kintamajam *Dažnis* galime nubrėžti ir naudodamiesi grafikų meniu *Graphs* (žr. 3.37 pav.):

- Atverkite bylą su analizuojamais duomenimis.
- Nurodykite komandas *Graphs* → *Legacy Dialogs* → *Histogram...*



3.37 pav. **Dialogo langelis *Legacy Dialogs***

- Pažymėkite analizuojamą kintamąjį *Dažnis* ir spragtelėję mygtuką su rodyklės ženklu įkelkite jį į analizuojamų kintamųjų sąrašą *Variable*.
- Paspauskite mygtuką *OK*.

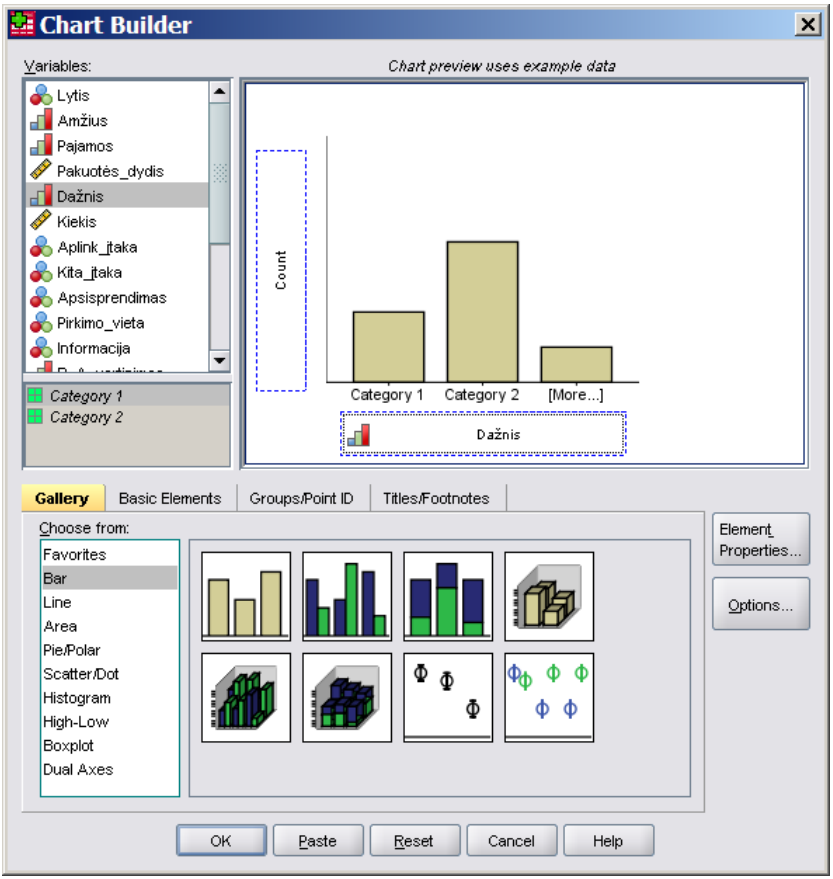
Bus nubrėžta kintamojo *Dažnis* histograma. Analogiškai, pasirinkus dialogo langelyje *Legacy Dialogs* diagramos tipus *Bar...* ar *Pie...*, bus nubrėžta stulpelinė arba skritulinė diagramos.

Bendruoju atveju diagramos brėžiamos naudojantis meniu *Graphs*. Nubrėšime kintamojo *Dažnis* stulpelinę diagramą (žr. 3.38 pav.):

- Atverkite bylą su analizuojamais duomenimis.
- Nurodykite komandas *Graphs → Chart Builder...*
- Pasirinkite srityje *Gallery* diagramos tipą *Bar.*
- Nuvilkite stulpelinės diagramos norimą tipą (*Simple Bar*) į diagramos peržiūros langą.
- Pažymėkite srityje *Variable*s analizuojamą kintamąjį *Dažnis* ir nuvilkite jį į x ašies laukelį diagramos peržiūros lange. Pagal nustatymą y ašies laukelyje figūruoja kintamasis *Count* (mūsų analizuojamo kintamojo pasirodymo dažnis).
- Spragtelėkite mygtuką *OK*.

Atliktos procedūros sintaksė:

```
* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset"
VARIABLES=Dažnis COUNT()[name="COUNT"]
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
  BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: Dažnis=col(source(s), name("Dažnis"), unit.category())
  DATA: COUNT=col(source(s), name("COUNT"))
  GUIDE: axis(dim(1), label("Dažnis"))
  GUIDE: axis(dim(2), label("Count"))
  SCALE: linear(dim(2), include(0))
  ELEMENT: interval(position(Dažnis*COUNT), sha-
pe.interior(shape.square))
  END GPL.
```

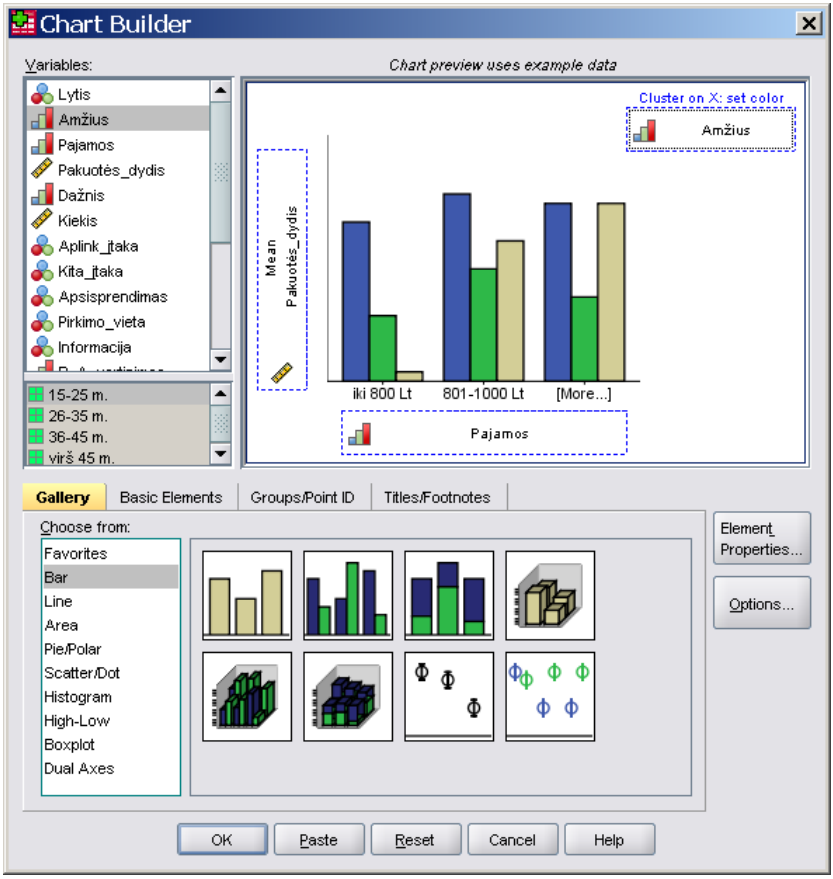


3.38 pav. **Dialogo langas *Chart Builder***

3.4.4. Sudėtingesnių grafikų braižymas. Trimatės (pirminės) ir sumuojančios diagramos

Pateiksime keletą kiek sudėtingesnių grafikų braižymo pavyzdžių. Juos galima nubrėžti naudojantis meniu *Graphs* komandomis *Chart Builder…* arba *Legacy Dialogs*.

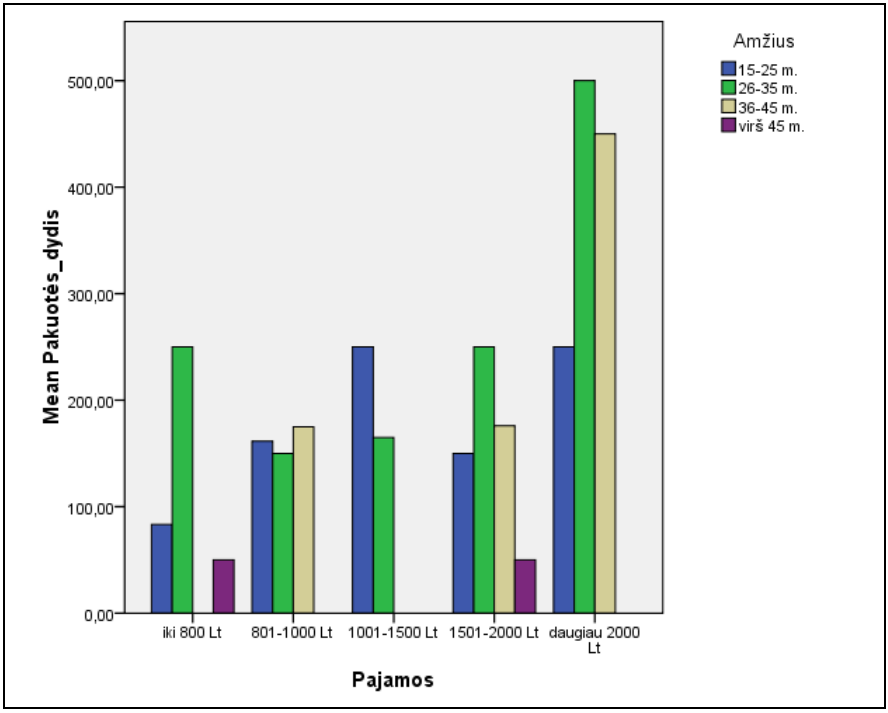
Nubrėšime stulpelinę diagramą, kurioje vaizduojama vidutinio perkamos kavos pakuotės dydžio priklausomybė nuo pajamų atskirose pirkėjų amžiaus grupėse (žr. 3.39 pav.).



3.39 pav. **Dialogo langas *Chart Builder***

- Atverkite bylą su analizuojamais duomenimis.
- Nurodykite komandas *Graphs* → *Chart Builder...*
- Pasirinkite srityje *Gallery* stulpelinę grupuotą diagramą (*Clustered Bar*) ir nuvilkite ją į diagramos peržiūros langą.

- Lauke *Statistic* nustatytas y ašyje esantis kintamasis *Count* (mūsų analizuojamo kintamojo pasirodymo dažnis). Pakeiskite jį į *Mean* (vidurkis) ir spragtelėkite mygtuką *Apply*.
- Pažymėkite srityje *Variables* kintamąjį *Pakuotės_dydis* ir nuvilkite jį y ašies laukelį diagramos peržiūros lange; kintamąjį *Pajamos* nuvilkite į x ašies laukelį; kintamąjį *Amžius* – į laukelį *Cluster on X: set color.*
- Spragtelėkite mygtuką *OK* (žr. 3.40 pav.).



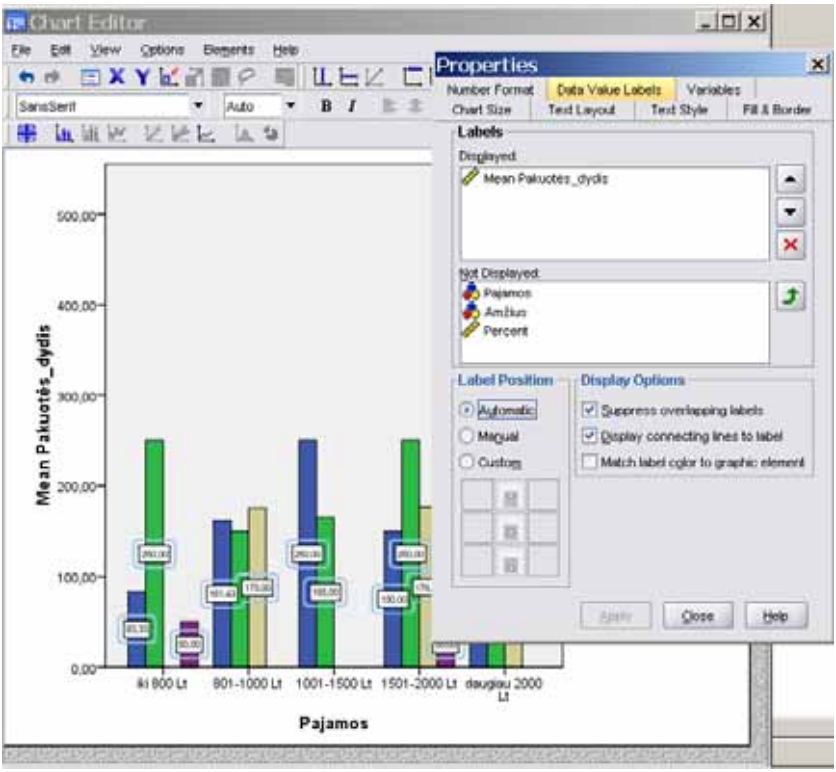
3.40 pav. **Sugrupuota stulpelinė diagrama**

Atliktos procedūros sintaksė:

```
* Chart Builder.
GGRAPH
/GRAPHDATASET NAME="graphdataset"
VARIABLES=Pajamos
MEAN(Pakuotės_dydis)[name="Pakuotės_dydis"] Amžius
MISSING=LISTWISE
REPORTMISSING=NO
/GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
SOURCE: s=userSource(id("graphdataset"))
DATA: Pajamos=col(source(s), name("Pajamos"), unit.category())
DATA: MEAN_Pakuotės_dydis=col(source(s), name("Pakuotės_dydis"))
DATA: Amžius=col(source(s), name("Amžius"), unit.category())
COORD: rect(dim(1,2), cluster(3,0))
GUIDE: axis(dim(3), label("Pajamos"))
GUIDE: axis(dim(2), label("Mean Pakuotės_dydis"))
GUIDE: legend(aesthetic(aesthetic.color.interior), label("Amžius"))
SCALE: cat(dim(3), include("1.00", "2.00", "3.00", "4.00", "5.00"))
SCALE: linear(dim(2), include(0))
SCALE: cat(aesthetic(aesthetic.color.interior), include("1.00", "2.00", "3.00", "4.00"))
SCALE: cat(dim(1), include("1.00", "2.00", "3.00", "4.00"))
ELEMENT: interval(position(Amžius*MEAN_Pakuotės_dydis*Pajamos), color.interior(Amžius), shape.interior(shape.square))
END GPL.
```

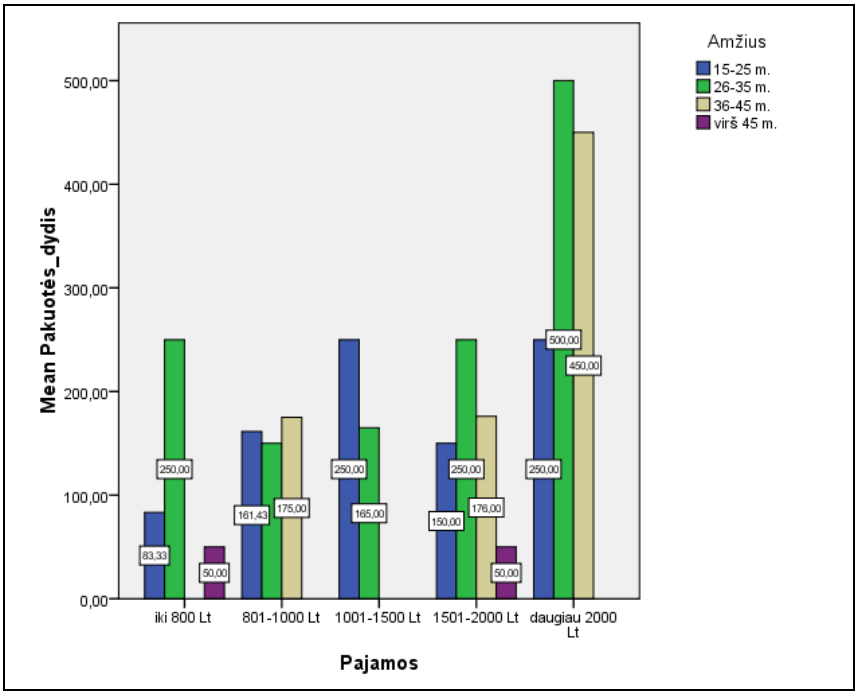
3.40 pav. diagramoje stulpeliams galima pridėti norimas žymenas.

- Dukart spragtelėkite pele bet kurioje diagramos vietoje. Diagrama atsidurs diagramų redaktoriaus lange.
- Nurodykite komandas *Elements* → *Show Data Labels* (arba spragtelėkite mygtuką *Show Data Labels*). Atsivers lango *Properties* duomenų žymenų kortelė (žr. 3.41 pav.).



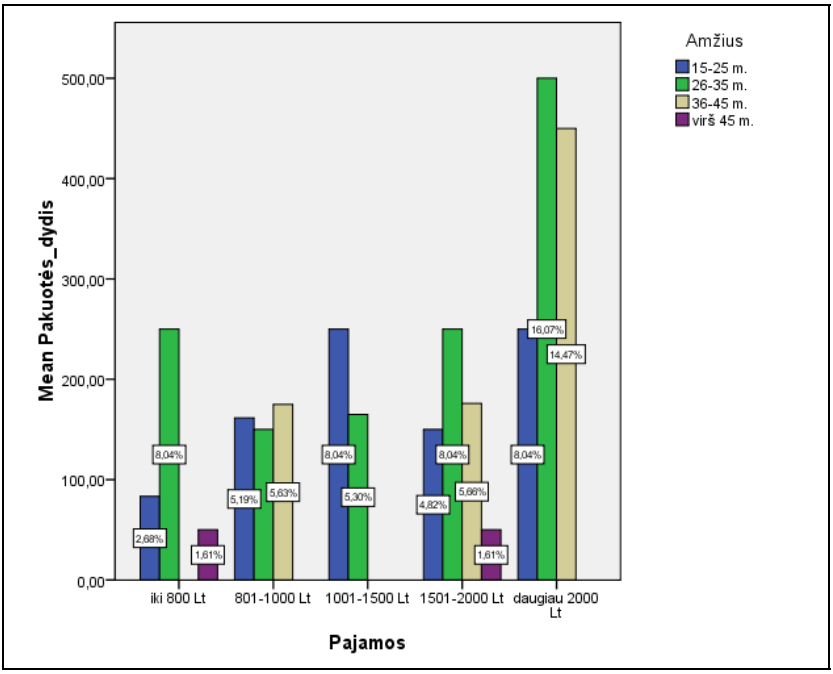
3.41 pav. ***Chart Editor. Properties. Data Value Labels***

Rodoma automatiškai parinkta kintamojo *Mean Pakuotės_dydis* reikšmės žymena. Užvėrę *Properties* ir *Chart Editor* langus, gausite diagramą, kurios kintamieji stulpeliuose turi žymenas – kintamųjų vidutines reikšmes (žr. 3.42 pav.):



3.42 pav. **Sugrupuota stulpelinė diagrama. Žymenos – kintamųjų vidutinės reikšmės**

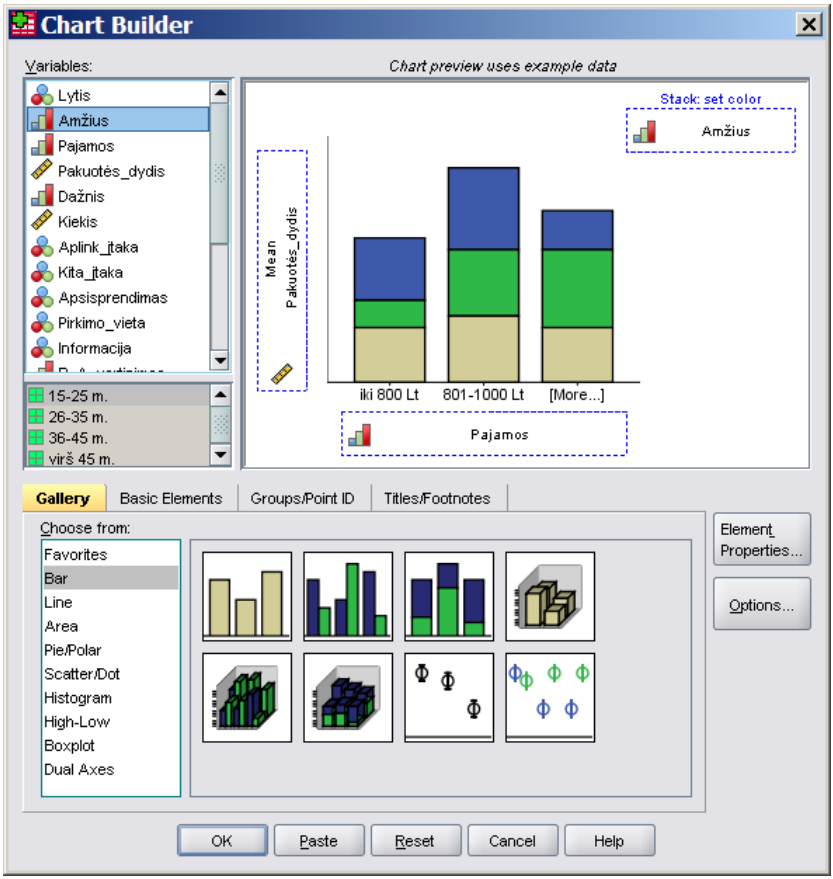
Jei *Properties* lango duomenų žymenų kortelėje (žr. 3.41 pav.) srityje *Displayed* pasirinksite *Percent* (tai galite padaryti iš srities *Not Displayed* žalia rodykle perkėlę *Percent* į sritį *Displayed*, o *Mean Pakuotės_dydis* raudonu mygtuku ištrynę), gausite diagramą (žr. 3.43 pav.):



3.43 pav. **Sugrupuota stulpelinė diagrama. Žymenos – procentai**

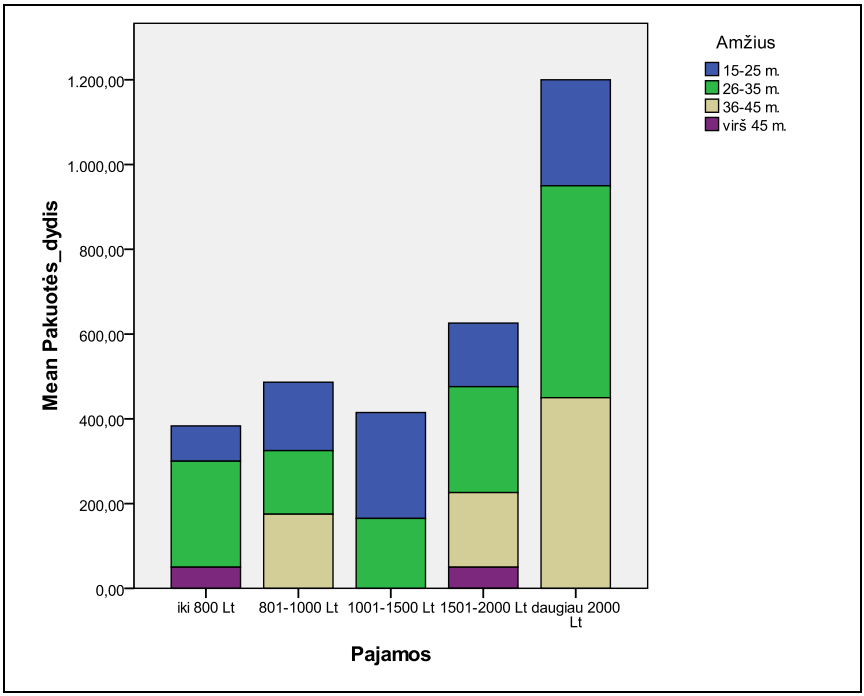
Pasirinkę vietoj *Clustered* variantą *Stacked*, gausite sudėtinę diagramą, kurioje kelios sekos bus pavaizduotos viename stulpelyje (žr. 3.44 pav.).

- Atverkite bylą su analizuojamais duomenimis.
- Nurodykite komandas *Graphs* → *Chart Builder...*
- Pasirinkite srityje *Gallery* stulpelinę diagramą (*Stacked Bar*) ir nuvilkite ją į diagramos peržiūros langą.
- Lauke *Statistic* nustatytas y ašyje esantis kintamasis *Count* (mūsų analizuojamo kintamojo pasirodymo dažnis). Pakeiskite jį į *Mean* (vidurkis) ir spragtelėkite mygtuką *Apply*.
- Pažymėkite srityje *Variables* kintamąjį *Pakuotės_dydis* ir nuvilkite jį į y ašies laukelį diagramos peržiūros lange; kintamąjį *Pajamos* nuvilkite į x ašies laukelį; kintamąjį *Amžius* – į laukelį *Stack: set color.*



3.44 pav. **Dialogo langas *Chart Builder***

- Spragtelėkite mygtuką *OK* (žr. 3.45 pav.).

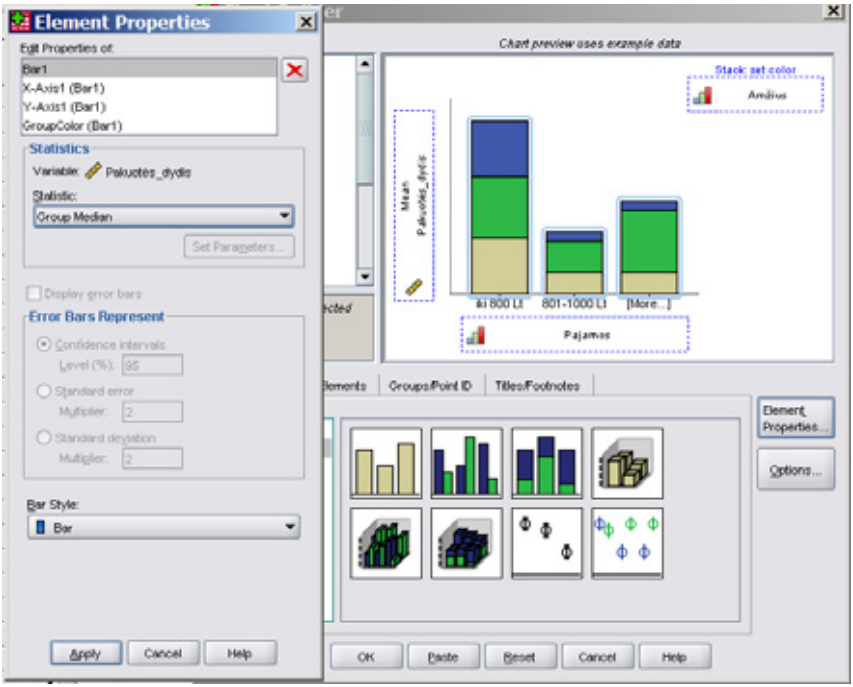


3.45 pav. ***Stacked* tipo stulpelinė diagrama**

Žymenos pridedamos analogiškai kaip ir *Clustered* tipo stulpelinei diagramai (žr. 3.41 pav.).

Stacked tipo stulpelinės diagramos y ašyje galima turėti įvairius kintamuosius:

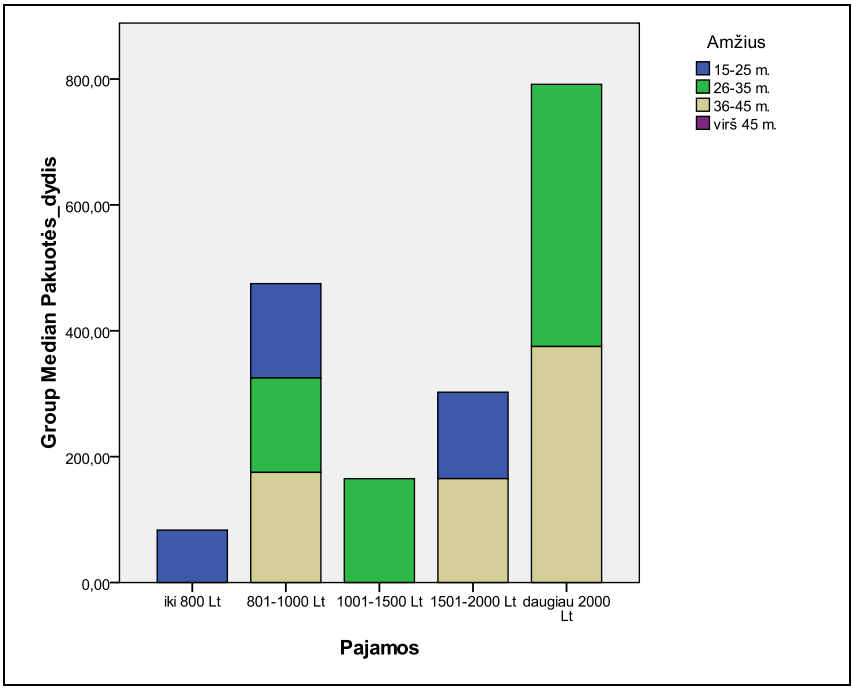
- Lauke *Statistic* nustatytas y ašyje esantis kintamasis *Count* (mūsų analizuojamo kintamojo pasirodymo dažnis). Pakeiskite jį į *Group Median* (mediana atitinkamai grupei) ir spragtelėkite mygtuką *Apply* (žr. 3.46 pav.).



3.46 pav. ***Element Properties*** **langas**

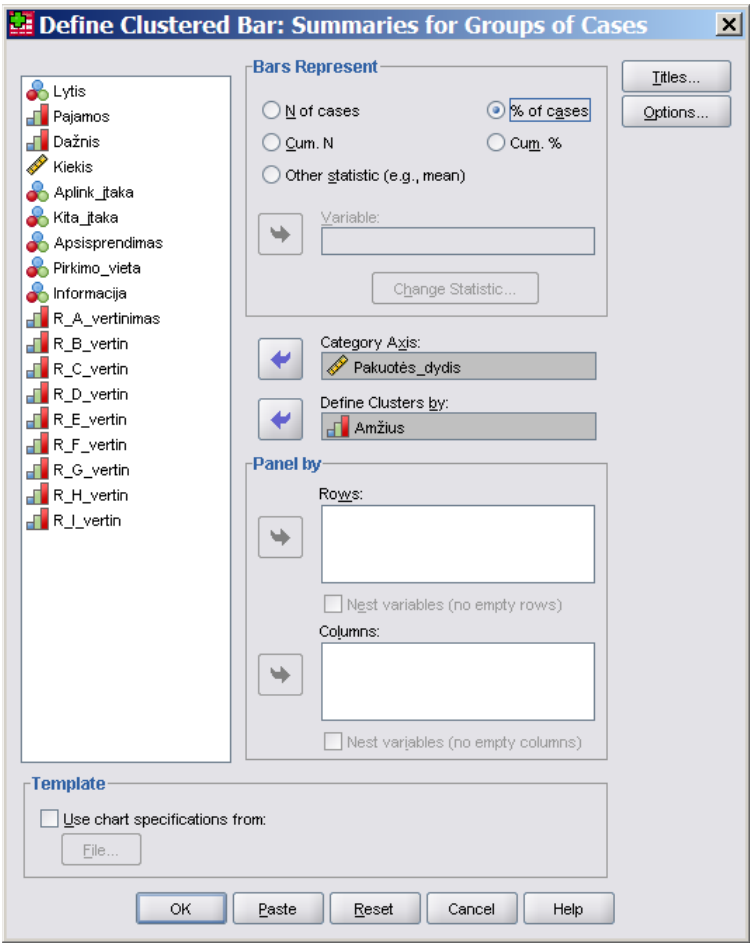
- Spustelėkite mygtuką *OK*.

Kaip jau minėta, sugrupuotas stulpelines diagramas galima nubrėžti ir naudojantis meniu *Graphs* komanda *Legacy Dialogs*. Sugrupuotoje stulpelinėje diagramoje pavaizduosime perkamos pakuotės dydžio priklausomybę nuo amžiaus grupės (žr. 3.47 pav.).



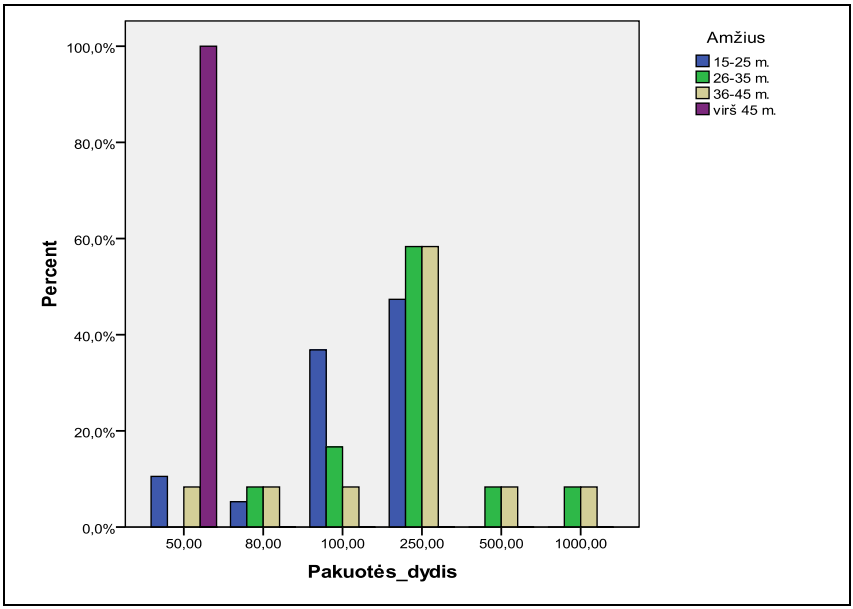
3.47 pav. ***Stacked* tipo stulpelinė diagrama**

- Atverkite bylą su analizuojamais duomenimis.
- Nurodykite komandas *Graphs* → *Legacy Dialogs* → *Bar...*
- Atsivėrusiame dialogo lange *Bar Charts* pasirinkite *Clustered* stulpelinės diagramos tipą ir spragtelėkite mygtuką *Define* (nustatyti).
- Į laukelį *Category Axis* įkelkite kintamąjį *Pakuotės_dydis*; į laukelį *Define Clusters by* – kintamąjį *Amžius*.
- Srityje *Bars Represent* pažymėkite *% of cases* – stebėjimų skaičius procentais (paprastai būna nustatyta *N of cases* – stebėjimų skaičius) (žr. 3.48 pav.):



3.48 pav. ***Define Clustered Bar: Summaries for Groups of Cases***

- Spragtelėkite mygtuką *OK* (žr. 3.49 pav.).

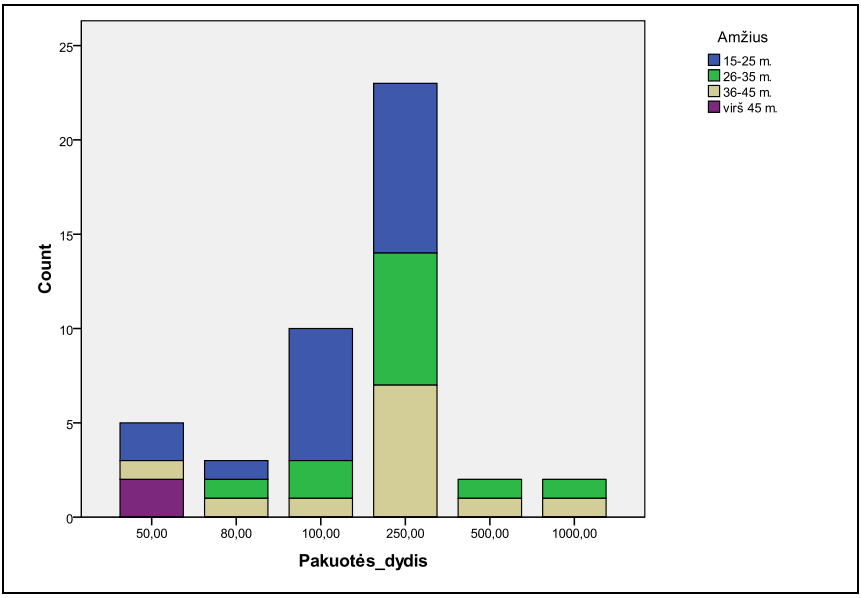


3.49 pav. **Perkamos pakuotės dydžio priklausomybė nuo amžiaus grupės procentais**

Prireikus stulpeliams pridedamos žymenos (*Labels*).

Norėdami kelias sekas pavaizduoti viename stulpelyje, taikome *Stacked* tipo diagramas.

- Atverkite bylą su analizuojamais duomenimis.
- Nurodykite komandas *Graphs* → *Legacy Dialogs* → *Bar...*
- Atsivėrusiame dialogo lange *Bar Charts* pasirinkite *Stacked* stulpelinės diagramos tipą ir paspauskite mygtuką *Define*.
- Į laukelį *Category Axis* įkelkite kintamąjį *Pakuotės_dydis*; į laukelį *Define Clusters by* – kintamąjį *Amžius*.
- Srityje *Bars Represent* palikite *N of cases* – stebėjimų skaičius.
- Spragtelėkite mygtuką *OK* (žr. 3.50 pav.).

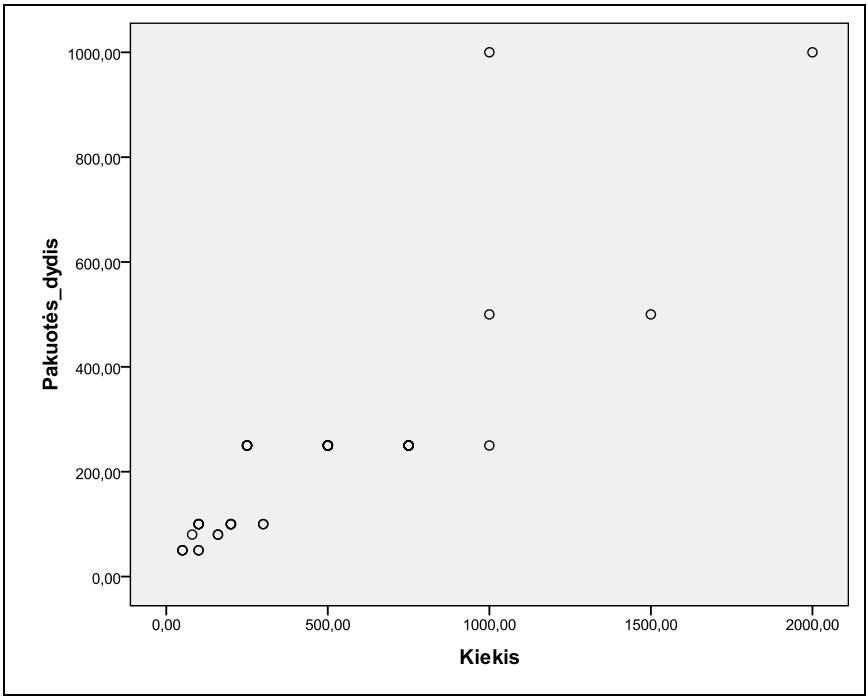


3.50 pav. **Perkamos pakuotės dydžio priklausomybė nuo amžiaus grupės**

Praktikoje dažnai taikomos sklaidos (taškinė) *Scatter/Dot* bei linijinė *Line* diagramos.

Nubrėžkime perkamos kavos pakuotės sklaidos diagramą (žr. 3.51 pav.):

- Atverkite bylą su analizuojamais duomenimis.
- Nurodykite komandas *Graphs* → *Legacy Dialogs* → *Scatter/Dot...*
- Atsivėrusiame dialogo lange *Scatter/Dot* pasirinkite *Simple Scatter* diagramos tipą ir spustelėkite mygtuką *Define*.
- Į laukelį *Y Axis* įkelkite kintamąjį *Pakuotės_dydis*; į laukelį *X Axis* – kintamąjį *Kiekis*.
- Spragtelėkite mygtuką *OK* (žr. 3.51 pav.).



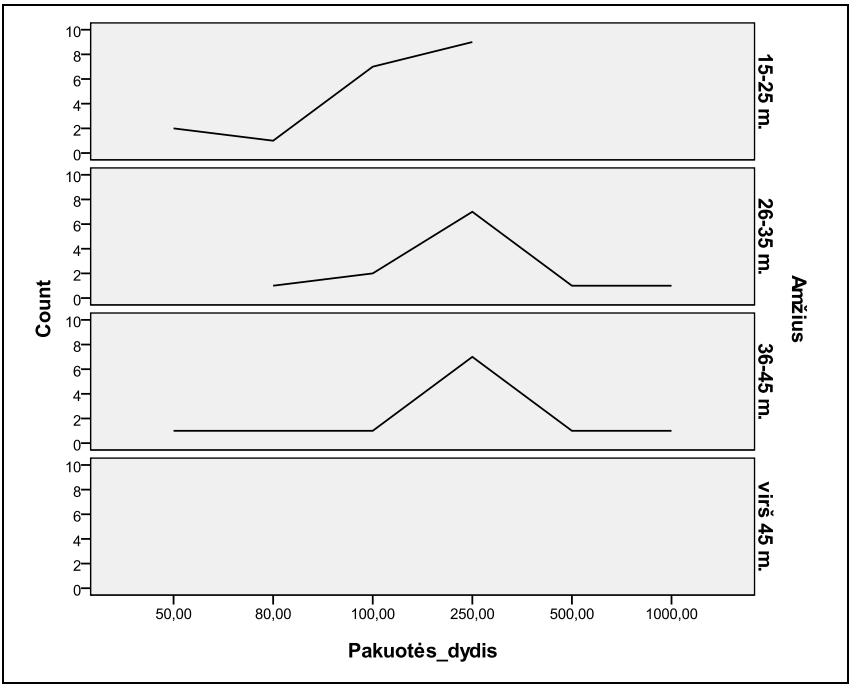
3.51 pav. **Perkamos pakuotės sklaidos diagrama**

Nubrėšime linijinę diagramą, kurioje pavaizduosime perkamos pakuotės priklausomybę nuo amžiaus grupės:

- Atverkite bylą su analizuojamais duomenimis.
- Nurodykite komandas *Graphs* → *Legacy Dialogs* → *Line...*
- Atsivėrusiame dialogo lange *Line Charts* pasirinkite *Multiple* diagramos tipą. Palikite nustatytą duomenų vaizdavimo diagramoje tipą *Summaries for groups of cases* (susumuoja atskiroms stebėjimų grupėms) ir spragtelėkite mygtuką *Define*.

Nubrėšime diagramą, kurioje kiekviena amžiaus grupė pavaizduota atskirame rėmelyje (*panel*) (žr. 3.52 pav.):

- Į laukelį *Category Axis* įkelkite kintamąjį *Pakuotės_dydis*; į srities *Panel by* laukelį *Rows* – kintamąjį *Amžius*.
- Paspauskite mygtuką *OK* (žr. 3.52 pav.).

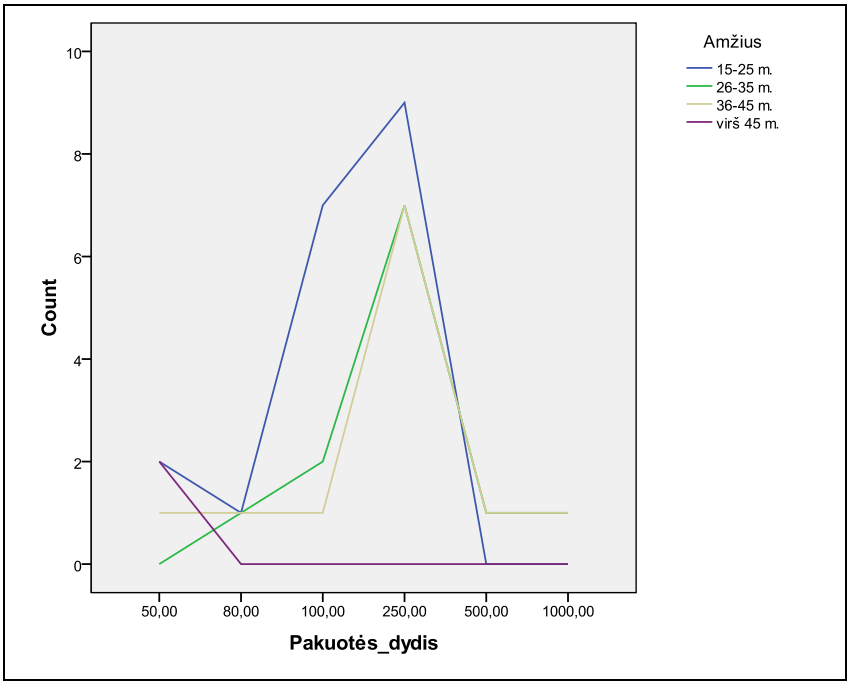


3.52 pav. **Perkamos pakuotės dydžio priklausomybė nuo amžiaus grupės**

Galima nubrėžti diagramą, kurioje amžiaus grupės vaizduojamos linijomis (žr. 3.53 pav.):

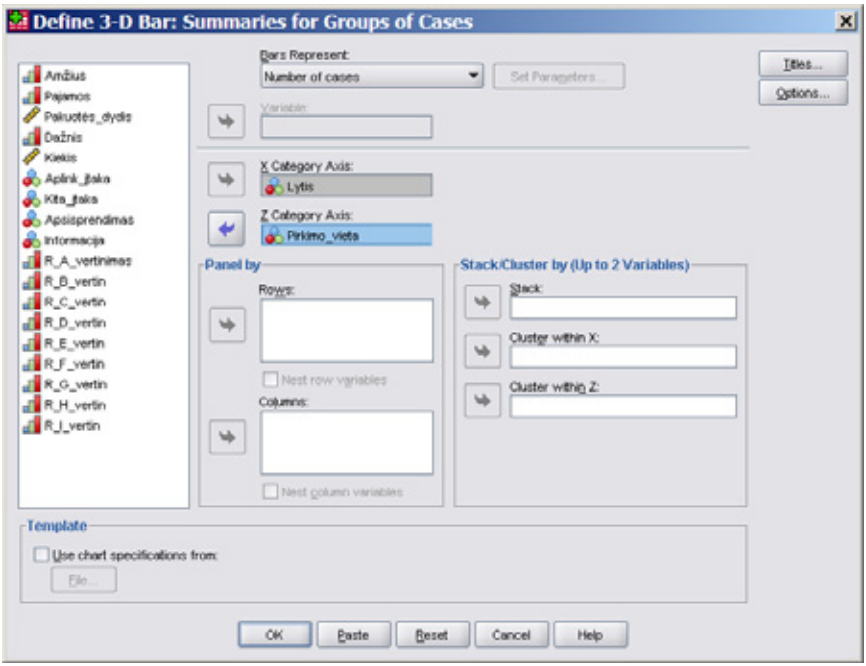
- Į laukelį *Category Axis* įkelkite kintamąjį *Pakuotės_dydis*; į laukelį *Define Lines by* – kintamąjį *Amžius*.
- Paspauskite mygtuką *OK*.

Porinėms dažnių lentelėms vaizduoti SPSS siūlo ir trimates (erdvines) diagramas. Šias diagramas lengva sukioti ir pasirinkti tinkamiausią pavidalą.



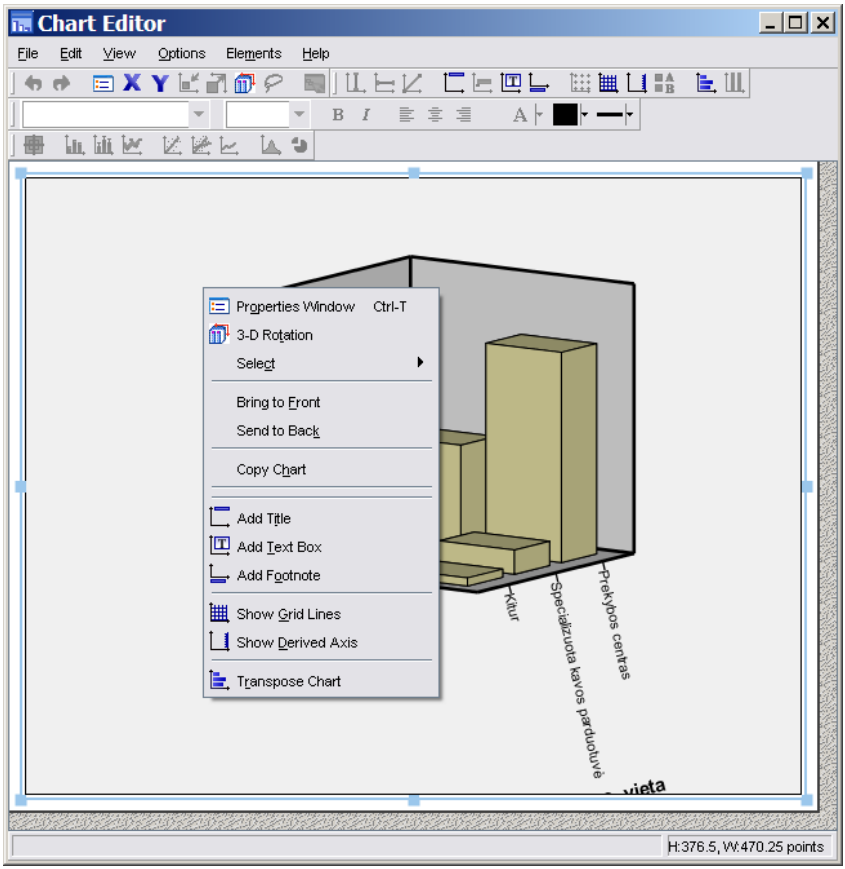
3.53 pav. **Perkamos pakuotės dydžio priklausomybė nuo amžiaus grupės**

- Atverkite bylą su analizuojamais duomenimis.
- Nurodykite komandas *Graphs* → *Legacy Dialogs* → *3-D Bar...*
- Dialogo lange *3-D Bar Charts* spragtelėkite mygtuką *Define.*
- Atsivėrusiame dialogo lange *Define 3-D Bar: Summaries for Groups of Cases* pasirinkite kintamąjį *Lytis* ir paspausdę mygtuką su rodykle įkelkite jį į laukelį *X Category Axis*. Tokiu pat būdu kintamąjį *Pirkimo_vieta* įkelkite į laukelį *Z Category Axis* (žr. 3.54 pav.).



3.54 pav. **Dialogo langas *Define 3-D Bar: Summaries for Groups of Cases***

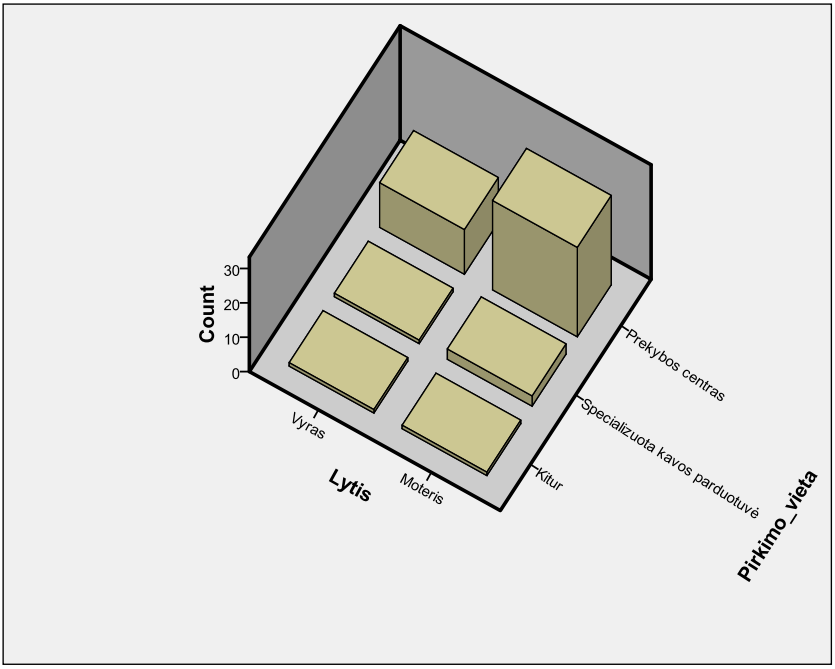
- Spragtelėkite mygtuką *OK*.
- Norėdami paredaguoti gautą diagramą, du kartus paspauskite ją pele. Diagrama atsidurs diagramų redaktoriaus lange. Spragtelėkite dešiniuoju pelės klavišu. Atsivėrusiame sąraše pasirinkite *3-D Rotation* komandą (žr. 3.55 pav.).



3.55 pav. **Dialogo langas *Chart Editor***

- Pele galite pasukti diagramą ir pasirinkti tinkamiausią pavidalą (žr. 3.56 pav.).

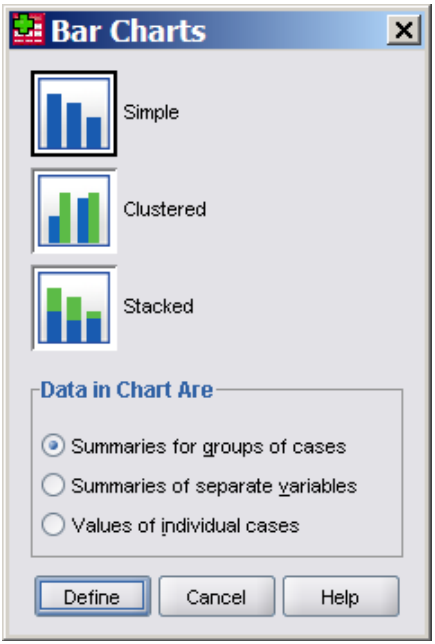
Statistiniuose tyrimuose dažnai kiekvienas respondento atsakymas vaizduojamas atskiru kintamuoju. Pavyzdžiui, respondento klausiama, kokios šalies virtuvę jis mėgsta, ir prašoma tai įvertinti balais. Atsakymai apie kiekvienos šalies virtuvę priskiriami atskiriems kintamiesiems. Juos galima nagrinėti atskirai vienas nuo kito arba pateikti visų suvestinę. Susipažinsime, kaip tokio tipo uždaviniai vaizduojami grafiškai.



3.56 pav. **Trimatė diagrama: kavos pirkimo vietos priklausomybė nuo lyties**

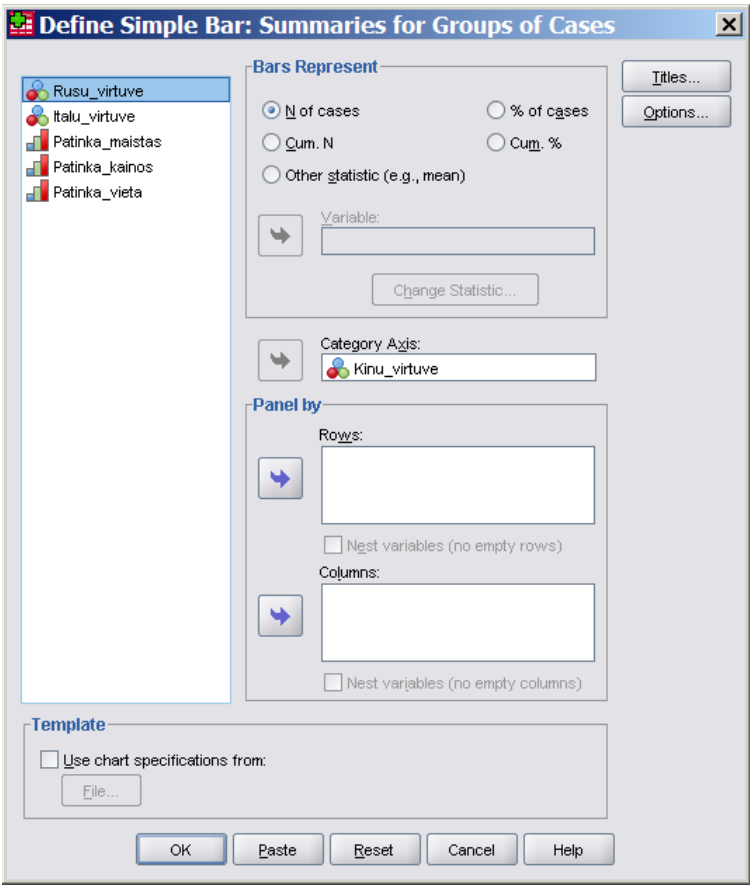
Studentai tyrė, kaip restoranų lankytojai vertina kinų, rusų ir italų virtuvę. Atsakymai buvo priskirti atskiriems kintamiesiems. Respondentų atsakymus apie kiekvieną virtuvę galima pavaizduoti grafiškai atskiromis diagramomis arba viena sumuojančia.

- Atverkite bylą su analizuojamais duomenimis.
- Nurodykite komandas *Graphs* → *Legacy Dialogs* → *Bar...*
- Dialogo lange *Bar Charts* pasirinkite diagramos tipą *Simple* ir srityje *Data in Chart Are* pažymėkite *Summaries for groups of cases* (žr. 3.57 pav.).



3.57 pav. **Dialogo langas *Bar Charts***

- Spragtelėkite mygtuką *Define*.
- Atsivėrusiame dialogo lange *Define Simple Bar: Summaries for Groups of Cases* mygtuku su rodykle kintamąjį *Kinu_virtuve* įkelkite į lauką *Category Axis* (žr. 3.58 pav.).
- Spragtelėkite mygtuką *OK*. Gausite stulpelinę diagramą, kurioje pavaizduoti respondentų atsakymai apie kinų virtuvę („taip“ – mėgsta kinų virtuvę, „ne“ – nemėgsta kinų virtuvės) (žr. 3.59 pav.):

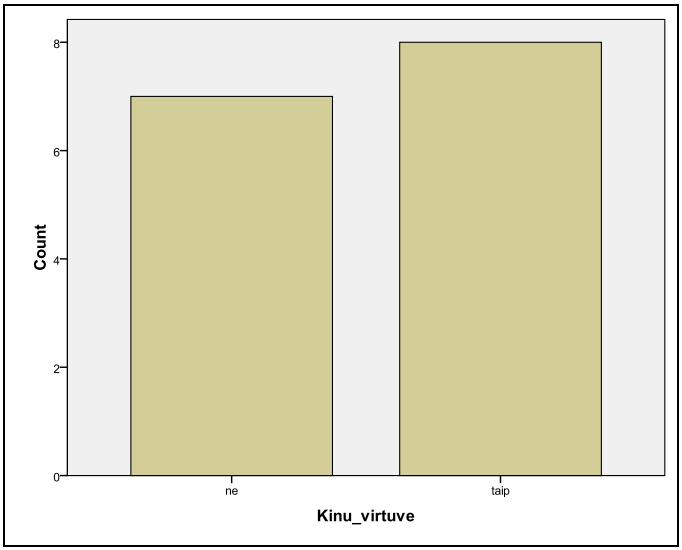


3.58 pav. **Dialogo langas *Define Simple Bar: Summaries for Groups of Cases***

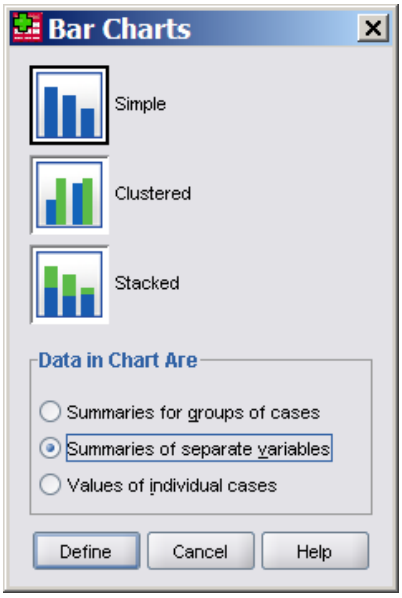
Analogiškai galima nubrėžti atskiras diagramas ir apie rusų bei italų virtuves.

Dažnai tyrėjams ataskaitoje informatyviau atrodo vienoje diagramoje pateikti visi kintamieji.

- Atverkite bylą su analizuojamais duomenimis.
- Nurodykite komandas *Graphs* → *Legacy Dialogs* → *Bar...*

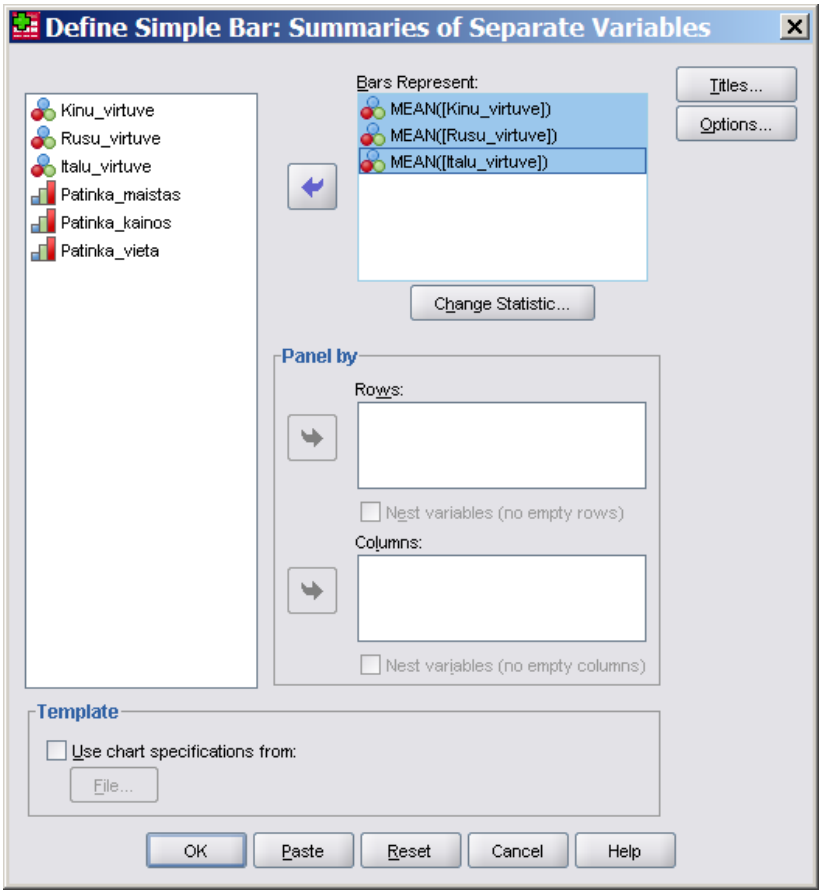


3.59 pav. **Respondentų atsakymai apie kinų virtuvę**



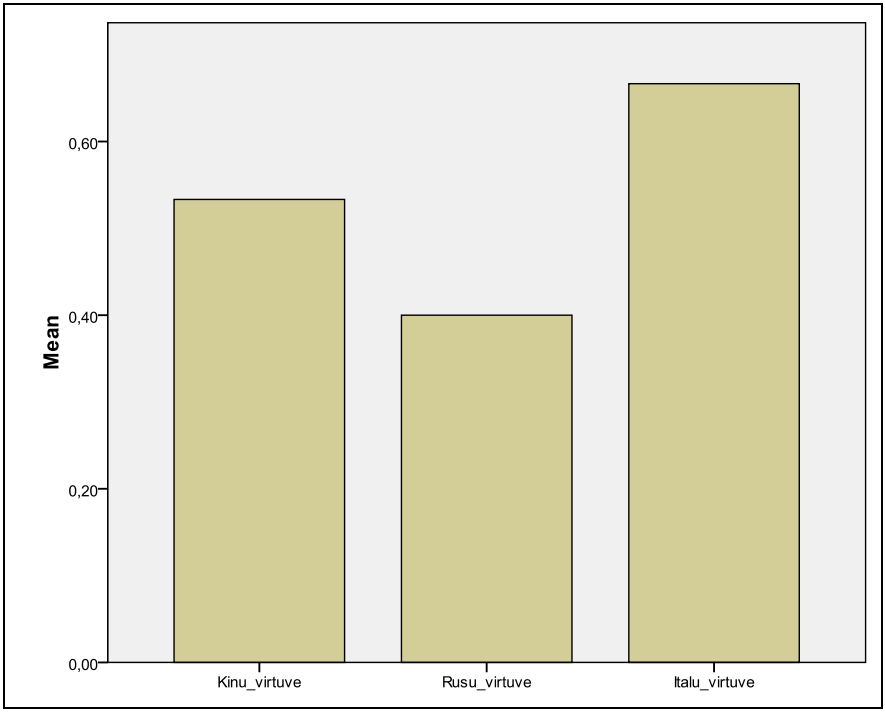
3.60 pav. **Dialogo langas *Bar Charts***

- Dialogo lange *Bar Charts* pasirinkite diagramos tipą *Simple* ir srityje *Data in Chart Are* pažymėkite *Summaries of separate variables* (žr. 3.60 pav.).
- Paspauskite mygtuką *Define*.
- Atsivėrusiame dialogo lange *Define Simple Bar: Summaries of Separate Variables* mygtuku su rodykle kintamuosius *Kinu_virtuve*, *Rusu_virtuve* ir *Italu_virtuve* įkelkite į lauką *Bars Represent* (žr. 3.61 pav.).



3.61 pav. **Dialogo langas *Define Simple Bar: Summaries of Separate Variables***

- Spragtelėkite mygtuką *OK*. Gausite diagramą, kurios stulpeliuose matysite respondentų atsakymų apie kiekvieną virtuvę vidutines reikšmes (žr. 3.62 pav.):

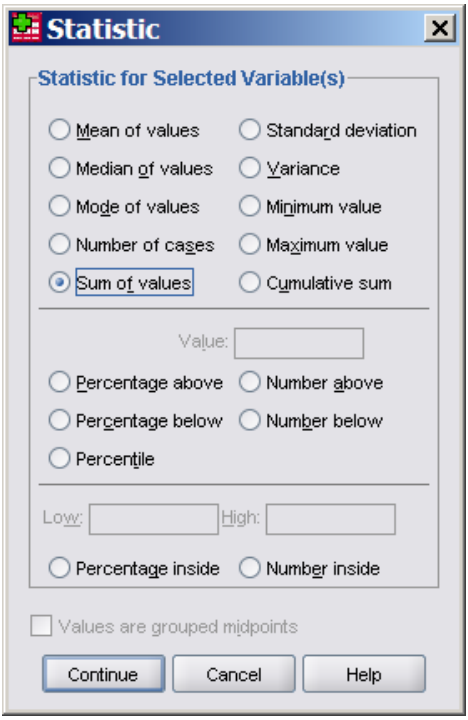


3.62 pav. **Restoranų lankytojų įvairių tautų virtuvių vertinimo vidurkiai**

Jei atsakant kintamieji vertinami balais (pavyzdžiui, nuo 0 iki 10), tai diagramos stulpeliuose galima matyti ne tik vidutines, bet ir sumines vertinimų reikšmes.

- Atverkite bylą su analizuojamais duomenimis.
- Nurodykite komandas *Graphs* → *Legacy Dialogs* → *Bar...*
- Dialogo lange *Bar Charts* pasirinkite diagramos tipą *Simple* ir srityje *Data in Chart Are* pažymėkite *Summaries of separate variables.*
- Spragtelėkite mygtuką *Define*.

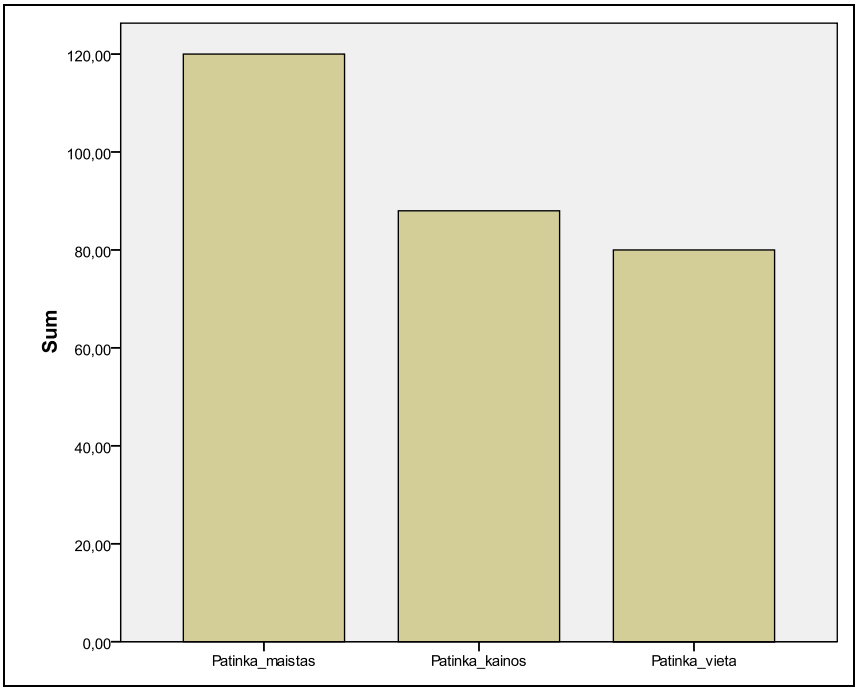
- Atsivėrusiame dialogo lange *Define Simple Bar: Summaries of Separate Variables* spustelėję mygtuku su rodykle kintamuosius *Patinka_maistas*, *Patinka_kainos* ir *Patinka_vieta* įkelkite į lauką *Bars Represent*.
- Paspauskite mygtuką *Change Statistic...*
- Atsivėrusiame dialogo lange *Statistic* pažymėkite *Sum of values* ir spragtelėkite mygtuką *Continue* (žr. 3.63 pav.).



3.63 pav. **Dialogo langas *Statistic***

Grįžkite į dialogo langą *Define Simple Bar: Summaries of Separate Variables.* Lauke *Bars Represent* turite pasirinktų kintamųjų sumas *SUM*([*Patinka_maistas*]), *SUM*([*Patinka_kainos*]) ir *SUM*([*Patinka_vieta*]).

- Spragtelėkite mygtuką *OK*.



3.64 pav. **Suminis restoranų vertinimas**

Diagramoje apskaičiuoti respondentų vertinimai pagal atskirus kintamuosius (žr. 3.64 pav.).

3.4.5. Padėties ir sklaidos charakteristikos

Kintamiesiems, išmatuotiems intervalų ar santykių skalėmis ir turintiems normalųjį skirstinį, dažniausiai skaičiuojamas vidurkis ir standartinis nuokrypis. Kartais skaičiuojamas vidurkio standartinis nuokrypis (standartinė klaida). Statistinės programos pateikia daug galimų duomenų padėties ir sklaidos charakteristikos skaičiavimo pasirinkimo galimybių. Tačiau pateikiant skaičiavimo rezultatus patartina apsiriboti viena, geriausiai duomenims tinkančia, centro charakteristika (vidurkiu

arba mediana) ir viena sklaidos charakteristika (dažniausiai pateikiamas standartinis nuokrypis arba standartinė klaida).

Statistinės duomenų analizės programos centro ir sklaidos charakteristikas grupuoja į kelias grupes.

Pirmąją grupę sudaro šios duomenų padėties (centro) charakteristikos:

Vidurkis arba **aritmetinis vidurkis** (*Mean*) – tai populiariausias duomenų vidurio (centro) matas. Jis skaičiuojamas taip: sudedant visas reikšmes ir padalijant iš jų skaičius. Vidurkis, apskaičiuotas iš imties, yra empirinis ir žymimas $\bar{x}$. Kitaip nei empirinis vidurkis, populiacijos vidurkis žymimas raide μ. Vidurkis yra kiekybinis rodiklis, kuris apibendrina ir išreiškia vienarūšių reiškinių tam tikro varijuojančio požymio tipišką lygį konkrečiomis vietos ir laiko sąlygomis.

Tarkime, tam tikrą objektų visumą tiriame pagal mus dominantį požymį X, o šio požymio galimos reikšmės (variantai) visumos objektams yra $x_1, x_2, x_3, \ldots, x_n$. Aritmetinis vidurkis skaičiuojamas pagal formulę

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x_i}{n}. \qquad (3.2)$$

Tuo atveju, jei variantų reikšmės kartojasi, skaičiuojamas svertinis aritmetinis vidurkis

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_n f_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum x_i f_i}{\sum f_i}. \qquad (3.3)$$

Čia f_i – yra reikšmės x_i pasikartojimų skaičius – dažniai.

Pavyzdys. Tirtas studentų kelionės į universitetą laikas.

Laikas, min.	*0–5*	*5–10*	*10–15*	*15–20*	*20–25*	*25–30*
Studentų skaičius	12	4	3	15	10	6

Apskaičiuosime, kiek vidutiniškai užtrunka studento kelionė į universitetą.

Tuo atveju, kai požymio reikšmės yra duotos intervalais, variantų reikšmėmis imamos intervalų vidurio reikšmės

$$x_i = \frac{x_{i1} + x_{i2}}{2}, \tag{3.4}$$

čia: x_{i1} – apatinė intervalo reikšmė,
x_{i2} – viršutinė intervalo reikšmė.

Pertvarkysime duotų intervalų pasiskirstymo eilutę į diskrečiąją, variantų reikšmėmis imdami intervalo vidurio taškus.

x_i	2,5	7,5	12,5	17,5	22,5	27,5
f_i	12	4	3	15	10	6

Skaičiuojame svertinį vidurkį:

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{2{,}5*12 + 7{,}5*4 + 12{,}5*3 + 17{,}5*15 + 22{,}5*10 + 27{,}6*6}{12+4+3+15+10+6} = 15.$$

Mediana (*Median*) – tai matavimų skalės taškas, kuris stebimas reikšmes dalija į dvi lygias dalis. Pavyzdžiui, jei išmatavę gavome reikšmes

7 1 8 8 9 1 6 8 5,

tai pirmiausia jas reikia išdėstyti pagal didėjimą, t. y. sudaryti ranžiruotą variacinę eilutę:

1 1 5 6 7 8 8 8 9

ir surasti vidurinę reikšmę (šiuo atveju 7). Jei stebėjimų skaičius yra lyginis, tai surandamas dviejų vidurinių reikšmių aritmetinis vidurkis. Kurių narių vidurkį imti randame pagal formulę (N + 1)/2.

Pavyzdžiui, jei

N = 10. Pagal formulę 11/2 = 5,5. Mediana bus 5 ir 6 ranžiruotos eilutės narių aritmetinis vidurkis.

N = 9. Pagal formulę 10/2 = 5. Mediana bus 5 ranžiruotos eilutės narys.

Moda (*Mode*) – tai dažniausiai pasikartojanti reikšmė. Jeigu kelios reikšmės kartojasi vienodą skaičių kartų, tai pasirenkama mažiausia reikšmė.

Antrąją grupę sudaro duomenų sklaidos charakteristikos

Požymio atskirų reikšmių skirtumai tam tikros visumos viduje statistikoje vadinami požymio sklaida. Visumos elementai apibūdinami kiekybinėmis ir kokybinėmis požymių reikšmėmis. Sklaida statistikoje suprantama kaip nagrinėjamą reiškinį apibūdinančio požymio dydžio kiekybiniai pokyčiai dėl skirtingų veiksnių nenuspėjamos įtakos. Kaip jau minėta, būdingiausius visumos bruožus apibūdina vidurkis. Sklaidos rodikliai įvertina požymio reikšmių išsisklaidymo apie vidurkį dydį ir amplitudę.

Sklaidos plotis – tai maksimaliosios ir minimaliosios požymio reikšmės skirtumas.

$$R = x_{\max} - x_{\min}, \tag{3.5}$$

čia: x_{max} – didžiausia požymio reikšmė,

x_{min} – mažiausia požymio reikšmė.

Standartinis nuokrypis (*Standard Deviation*) – tai matuojamųjų dydžių sklaidos apie vidurkį matas. Jis apskaičiuojamas ištraukiant kvadratinę šaknį iš dispersijos. Imties (empirinis) standartinis nuokrypis žymimas raide s, o populiacijos standartinis nuokrypis – raide σ. Jei duomenų skirstinys yra normalusis, tai į intervalą, kuris gaunamas atidėjus į abi puses nuo vidurkio po standartinį nuokrypį s, $(\bar{x} - s, \bar{x} + s)$ patenka apytiksliai 68 procentai visų duomenų. Į intervalą $(\bar{x} - 2s, \bar{x} + 2s)$ patenka apytikriai 95,4 procentai visų duomenų, o į intervalą $(\bar{x} - 3s, \bar{x} + 3s)$ – beveik visi duomenys (tiksliau, 99,7 procentai).

$$s = \sqrt{s^2}. \tag{3.6}$$

Dispersija (*Variance*) apskaičiuojama sudedant kiekvienos reikšmės atstumo (skirtumo) nuo vidurkio kvadratus ir gautą sumą padalijant iš reikšmių skaičiaus minus 1. Dispersija, kaip ir standartinis nuokrypis, parodo duomenų sklaidą apie vidurkį. Kadangi dispersija apskaičiuojama keliant reikšmes kvadratu, tai ir dispersijos matavimo vienetai yra pakelti kvadratu. Standartinis nuokrypis (kvadratinė šaknis iš dispersijos) matuojamas tokiais pačiais matavimo vienetais kaip ir patys duomenys, todėl jį lengviau suprasti ir interpretuoti.

Dispersija

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}. \tag{3.7}$$

Sugrupuotiems duomenims

$$s^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i}. \tag{3.8}$$

Vidurkio standartinis nuokrypis arba **standartinė klaida** (*Standard Error of Mean*) yra apskaičiuojama imties standartinį nuokrypį s padalijant iš kvadratinės šaknies, ištrauktos iš visų duomenų skaičiaus (imties tūrio n). Vidurkio standartinis nuokrypis, kaip ir imties standartinis nuokrypis, padeda nustatyti atskirų reikšmių išsidėstymo intervalą. Su 99,7 procentų tikimybe galima teigti, kad populiacijos vidurkis yra intervale, kuris gaunamas atidėjus į abi puses nuo vidurkio po tris vidurkio standartinius nuokrypius $(\bar{x} - 3s, \bar{x} + 3s)$; su 95,4 procentų tikimybe galima teigti, kad populiacijos vidurkis yra intervale, gautame atidėjus į abi puses nuo vidurkio po du vidurkio standartinius nuokrypius $(\bar{x} - 2s, \bar{x} + 2s)$; o jei nuo vidurkio atidėsime po vieną vidurkio standartinį nuokrypį $(\bar{x} - s, \bar{x} + s)$, tai tikimybė, kad į šį intervalą pateks populiacijos vidurkis, yra apytikriai 68 procentai.

Vidurkio patikimumo intervalas (*Confidence Limits of Mean*). Vidurkio standartinis nuokrypis arba standartinė klaida leidžia apskaičiuoti vidurkio patikimumo (pasikliautinąjį) intervalą. Su 95,4 procentų tikimybe galima teigti, kad intervale, kurį sudaro du vidurkio standartiniai nuokrypiai, atidėti į abi puses nuo imties vidurkio, bus populiacijos vidurkis. Jei atidėsime ne po 2, o po 1,96 vidurkio standartinius nuokrypius gausime tikslią 95 procentų tikimybę. Šis intervalas vadinamas 95 procentų vidurkio patikimumo (pasikliautinuoju) intervalu.

Variacijos (kitimo) koeficientas. Šis rodiklis nustato sklaidos laipsnį:

$$CV = \frac{s}{\bar{x}} * 100\ \%. \qquad (3.9)$$

Vertinama taip: iki 10 % – sklaida maža; 10–20 % – sklaida vidutinė; 20–30 % – sklaida didelė; 30 % ir daugiau – sklaida labai didelė.

Į atskirą (trečiąją) grupę išskirti procentiliai

Skaičiai, suskirstantys variacinę eilutę į 100 apytikriai lygių dalių, vadinami *procentiliais*, o skaičiai, dalijantys variacinę eilutę į keturias maždaug lygias dalis, vadinami *kvartiliais*. Dažniausiai kvartiliai žymimi raidėmis Q1, Q2 ir Q3. Pirmasis kvartilis Q1 – tai taškas, išmatuotų reikšmių skalėje, nuo kurio kairėje pusėje lieka 25 procentai reikšmių. Antrasis kvartilis Q2 – taškas, kurio kairėje pusėje yra 50 procentų reikšmių. Antrasis kvartilis sutampa su mediana. Trečiojo kvartilio Q3 kairėje pusėje yra 75 procentai reikšmių.

Kaip vienas iš galimų duomenų centro charakteristikų taikomas pirmojo ir trečiojo kvartilių vidurkis $(Q_1 + Q_3)/2$.

Skaičiai, dalijantys duomenų eilutę į 10 lygių dalių, vadinami *deciliais*.

Ketvirtąją grupę sudaro du asimetriškumo matai

Asimetrijos koeficientas (*Skewness*) matuoja, ar duomenų išsidėstymas skiriasi nuo simetrinio išsidėstymo, t. y. nuo tokio duomenų išsidėstymo, kai reikšmės išsidėsto vienodais atstumais į abi puses nuo vidurkio. Jei duomenys turi normalųjį skirstinį, tai asimetrijos koeficientas lygus nuliui. Asimetrijos koeficientą galima taikyti kaip vieną iš būdų

tikrinant duomenų normališkumą. Jei asimetrijos koeficientas reikšmingai skiriasi nuo nulio, tai duomenų skirstinys nėra normalusis. Kai asimetrijos koeficientas didesnis už nulį, asimetrija teigiama, kai mažesnis – neigiama.

Asimetrijos koeficientas skaičiuojamas pagal formulę

$$As = \frac{1}{\frac{n-1}{s^3}} \sum (x_i - \bar{x})^3. \qquad (3.10)$$

Eksceso koeficientas (*Kurtosis*) parodo, ar histograma yra aukšta su trumpomis „uodegomis“, ar atvirkščiai, plokščia su ilgomis „uodegomis“. Eksceso koeficientas parinktas taip, kad tuo atveju, kai duomenų skirstinys yra normalusis, eksceso koeficientas lygus nuliui.

$$Ex = \frac{1}{\frac{n-1}{s^4}} \sum (x_i - \bar{x})^4 - 3. \qquad (3.11)$$

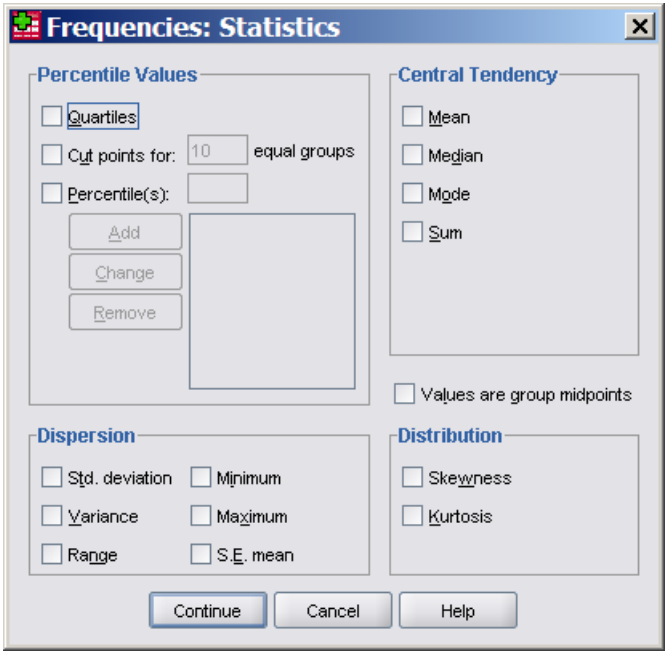
SPSS panagrinėsime kintamųjų padėties ir sklaidos charakteristikas. Kaip pavyzdį pateikime apklausą, kurioje sukaupti respondentų atsakymai apie jų gaunamas pajamas – intervalinis kintamasis *Menesio grynosios pajamos*.

- Atverkite bylą su analizuojamais duomenimis.
- Nurodykite komandas *Analyze* → *Descriptive Statistics* → *Frequencies...*
- Įkelkite kintamąjį *Menesio grynosios pajamos* į kintamųjų sąrašą *Variable*(*s*).
- Paspauskite dialogo langelio *Frequencies* mygtuką *Statistics*... Pasirodys dialogo langelis *Frequencies: Statistics* (žr. 3.65 pav.)

Grupėje *Percentile Values* (procentilių reikšmės) galite pasirinkti šiuos variantus:

Quartiles (kvartiliai). Bus pateikti pirmasis, antrasis ir trečiasis kvartiliai. Pirmasis kvartilis (Q1) – tai taškas surikiuotų pagal didėjimą duomenų skalėje, žemiau (kairėje) jo yra ketvirtis (25 procentai) išmatuotų reikšmių. Antrasis kvartilis (Q2) atskiria du ketvirčius (50 procen-

tų), o trečiasis (Q3) – tris ketvirčius (75 procentus) išmatuotų reikšmių. Antrasis kvartilis vadinamas mediana.



3.65 pav. **Dialogo langelis *Frequencies: Statistics***

Cut points for (dalijimo taškai). Bus pateiktos procentilių, dalijančių surikiuotos pagal didėjimą imties duomenis į lygias, t. y. turinčias vienodą duomenų skaičių grupes, reikšmės. Galima nurodyti norimą grupių skaičių (pirminis nustatytasis grupių skaičius – 10).

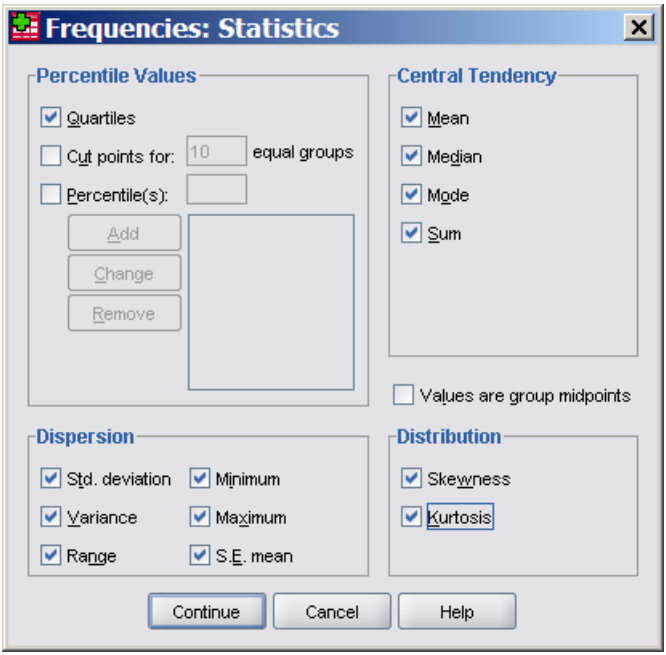
Percentile(*s*). Čia galima nurodyti procentilių reikšmes savo nuožiūra. Laukelyje įrašykite pirmąją procentilio reikšmę, ribojamą nuo 0 iki 100, ir spragtelėkite mygtuką *Add*. Pakartokite šiuos veiksmus visoms kitoms procentilių reikšmėms.

Grupėje *Central Tendency* galima pasirinkti šias duomenų padėties (centro) charakteristikas: *Mean* (vidurkį), *Median* (medianą), *Mode* (modą) arba *Sum* (sumą) – visų reikšmių sumą.

Grupėje *Dispersion* galima pasirinkti sklaidos charakteristikas, apibūdinančias duomenų išsidėstymą: *Std. deviation* (standartinis nuokrypis), *Variance* (dispersija), *Range* (plotis), *Minimum* (mažiausia reikšmė), *Maximum* (didžiausia reikšmė) arba *S.E. mean* (standartinė paklaida).

Grupėje *Distribution* (pasiskirstymas) galima pasirinkti šiuos pasiskirstymo asimetriškumo matus: *Skewness* (asimetrijos koeficientas) arba *Kurtosis* (ekscesas). Normaliojo skirstinio asimetrijos ir ekscesо koeficientai yra lygūs nuliui.

- Pažymėkite varnele žymimuosius pasirinktų rodiklių langelius ir spragtelėkite mygtuką *Continue* (tęsti) (žr. 3.66 pav).



3.66 pav. **Dialogo langelis *Frequencies: Statistics*. Skaičiuojamų charakteristikų pasirinkimas**

- Grįžtame į dialogo langą *Frequencies*. Paspauskite mygtuką *OK*. Rezultatus matome rodinio (*Viewer*) lange.

Statistics		
Mėnesio grynosios pajamos		
N	Valid	119
	Missing	0
Mean		2543,7143
Std. Error of Mean		124,64500
Median		2759,0000
Mode		1888,00[a]
Std. Deviation		1359,71638
Variance		1848828,630
Skewness		-,147
Std. Error of Skewness		,222
Kurtosis		-1,120
Std. Error of Kurtosis		,440
Range		4746,00
Minimum		194,00
Maximum		4940,00
Sum		302702,00
Percentiles	25	1347,0000
	50	2759,0000
	75	3623,0000
a. Multiple modes exist. The smallest value is shown		

3.67 pav. **Dažnių statistikos rezultatai**

Iš gautų rezultatų (žr. 3.67 pav.) galima nustatyti visas skirstinio charakteristikas. Matome, kad asimetrijos koeficientas artimas nuliui, todėl galime teigti, kad skirstinio forma neiškreipta. Vidurkis ir mediana artimi – duomenys pasiskirsto apie vidurinę reikšmę. Tai rodo ir mažas vidutinis nuokrypis.

Atliktos procedūros sintaksė:

```
GET
FILE='C:\...\Apklausa_men_pajamos_neto.sav'.
FREQUENCIES VARIABLES=Men_pajamos
/NTILES=4
/STATISTICS=STDDEV VARIANCE RANGE MINIMUM
MAXIMUM SEMEAN MEAN MEDIAN MODE SUM SKEWNESS
SESKEW KURTOSIS SEKURT
/ORDER=ANALYSIS.
```

KONTROLINIAI KLAUSIMAI

3.1. Kokie yra pagrindiniai aprašomosios statistikos tikslai (uždaviniai)?
3.2. Kokie aprašomosios statistikos metodai taikomi kintamiesiems, išmatuotiems nominaliąja matavimo skale?
3.3. Kokie aprašomosios statistikos metodai taikomi kintamiesiems, išmatuotiems intervalų matavimo skale?
3.4. Koks skirstinys vadinamas normaliuoju?
3.5. SPSS programa tikrindami tyrimo rezultatų stebimąjį reišmingumą gavome Sig. = 0,01. Esame pasirinkę reikšmingumo lygmenį 0,05. Kokią išvadą galima daryti apie tyrimo rezultatų skirstinio normalumą?
3.6. Kada vidurkis lygus medianai?
3.7. Tarkime, kad sklaidos plotis yra 105.4, o standartinis nukrypimas – 260.6. Kokią išvadą galite padaryti?
3.8. Gavote dispersiją lygią –45.5. Kokią išvadą galite padaryti?

UŽDUOTYS

3.1 užduotis

Pridedame po 100 prie kiekvieno sekos nario. Kas atsitiks standartiniam nuokrypiui, sklaidos pločiui, vidurkiui ir medianai?

3.2 užduotis

Lentelėje pateiktos planetų charakteristikos. Raskite medianą kiekvienam dydžiui. Jei kas būtų sakę, kad gavo diametro medianą 142 851,50, kokia klaida būtų padaryta?

Planeta	*Atstumas nuo Saulės (mln. km)*	*Diametras (km)*	*Metai (dienomis)*
Merkurijus	57,89	4 872,24	88
Venera	75,58	12 092,16	225
Žemė	149,54	12 745,01	365
Marsas	228,34	6 780,94	687
Jupiteris	778,27	142 851,50	4 332
Saturnas	1 426,30	120 432,77	10 760
Uranas	2 838,12	51 073,30	30 684
Neptūnas	4 487,93	49 484,59	30 188
Plutonas	5 875,63	2 296,22	90 467

4. STATISTINĖS HIPOTEZĖS

4.1. Imties skirstiniai. Įverčiai

Pagrindinė statistikos problema – remiantis imtimi gauti išvadas apie visą populiaciją. Bendrasis atsitiktinės imties modelis aprašomas taip. Tarkime, kad n kartų matuojame atsitiktinį dydį X. Tuomet atsitiktinę imtį sudaro atsitiktinis vektorius (C), kurio visi atsitiktiniai dydžiai X_1, X_2, ..., X_n yra nepriklausomi, vienodai pasiskirstę ir turintys tą patį kaip ir matuojamasis dydis X skirstinį. Atsitiktinė imtis X_1, X_2, ..., X_n ir bet kuri jos funkcija yra atsitiktiniai dydžiai. Statistinėms išvadoms taikoma kokia nors imties duomenų funkcija. Atsitiktinės imties funkcija $f(X_1, X_2, ..., X_n)$ vadinama *statistika*. Statistikos realizacija (konkretus skaičius) skiriasi nuo pačios statistikos (atsitiktinio dydžio). Statistiką žymėsime didžiosiomis raidėmis, o jos realizaciją – mažosiomis. Pavyzdžiui, $\overline{X}$ yra statistika, o $\overline{x}$ – realizacija.

Dažniausiai mus domina ne visa informacija apie stebimojo atsitiktinio dydžio skirstinį, o tik kai kurios skaitinės skirstinio charakteristikos. Kitaip sakant, remdamiesi imties duomenimis, norime įvertinti populiacijos parametrą. Matematiškai tokią situaciją aprašo parametrinis modelis.

Parametrinis modelis $X \sim P_\theta$: stebimasis atsitiktinis dydis X, kurio skirstinys P_θ priklauso nuo nežinomo parametro θ iš aibės Θ.

Statistika, kuri taikoma nežinomam parametrui θ įvertinti, vadinama θ taškiniu įverčiu (toliau – įverčiu) ir žymima $\hat{\theta}$. Nežinomas parametras θ yra skaičius, parametro įvertis $\hat{\theta}$ – atsitiktinis dydis, įverčio realizacija – skaičius, randamas konkrečiai imties realizacijai.

Parametro įverčio realizacija – tai apytikslė parametro reikšmė. Dažnai prireikia populiacijos parametrą palyginti su konkrečiu skaičiumi arba kitos populiacijos analogišku parametru. Taikant matematinės

statistikos metodus galima remiantis imtimi priimti sprendimą apie nežinomą populiacijos parametro reikšmę.

4.2. Statistinių hipotezių tikrinimas

Hipotezė – tai teiginys, aprašantis ir aiškinantis reiškinį, kurio teisingumas iš anksto nežinomas.

Statistinė hipotezė – tai prielaida apie empirinį atsitiktinio dydžio pasiskirstymo dėsnį ir (arba) apie jo empirines skaitines charakteristikas; tai teiginys apie masinius, pasikartojančius reiškinius. Statistinė hipotezė gali būti patikrinta taikant statistinius kriterijus. Statistiniai kriterijai pagrindžiami matematinės statistikos metodais. Formuluojant hipotezes tyrinėtojo nuomonė sistematizuojama ir pateikiama konkrečiu pavidalu.

Hipotezių tikrinimo procedūrą sudaro šie pagrindiniai žingsniai:

1. Imties skirstinio tipo nustatymas.
2. Hipotezių formulavimas.
3. Statistinio kriterijaus skaičiavimas.
4. Statistinio reikšmingumo lygmens parinkimas.
5. Sprendimo priėmimas.

1. Imties skirstinio tipo nustatymas. Atliekant statistinį tyrimą gaunami kintamojo X stebėjimo rezultatai x_1, x_2, x_3, ... , x_i,..., x_n (n – imties dydis). Sąryšis, siejantis atsitiktinio dydžio reikšmes ir jų dažnius, vadinamas atsitiktinio dydžio skirstiniu. Prieš parenkant statistinį kriterijų svarbu žinoti, ar išmatuotų duomenų skirstinys yra normalusis. Normalusis arba Gauso skirstinys grafiškai turi varpo formą. Gauso skirstinio tikimybės tankis užrašomas formule:

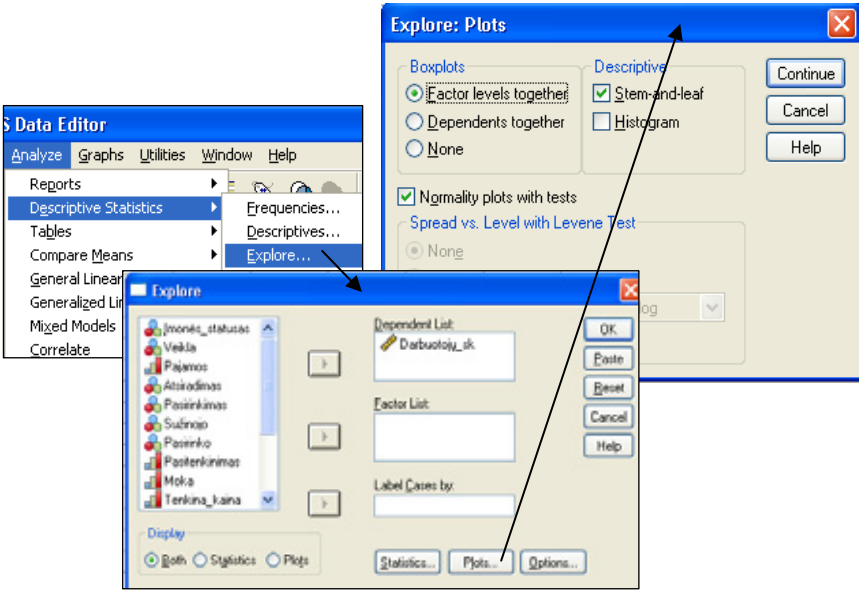
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad (4.1)$$

čia: σ – dispersija,
μ – vidurkis,
$e = 2{,}71$ – Neperio skaičius.

Skirstinio normalumas ypač svarbus esant mažoms imtims. Nuo imties skirstinio tipo priklauso statistinio kriterijaus parinkimas.

4.3. Kintamojo skirstinio normalumo tikrinimas SPSS paketu

Kiekybinio kintamojo skirstinio normalumą galima patikrinti taikant aprašomosios statistikos procedūrą *Explore*.



4.1 pav. **Dialogo langai *Explore* ir *Explore: Plots***

Patikrinsime, ar analizuojamų įmonių darbuotojų skaičius turi normalųjį skirstinį. Dialogo langas *Explore* atveriamas komanda *Analyze* → *Descriptive Statistics* → *Explore* (žr. 4.1 pav.). Kintamasis *Darbuotojų_sk* perkeliamas į langelį *Dependent List.* Mygtuko *Plots* paspaudi-

mu atveriamas dialogo langas *Explore: Plots.* Šiame lange reikia pažymėti langelį *Normality plots with tests.* Paspaudę mygtuką *Continue* grįžtame į pradinį langą. Procedūra baigiama mygtuko *OK* paspaudimu. Lange *Output-SPSS Viewer* rodomi analizės rezultatai (žr. 4.2 pav.).

Tests of Normality

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Darbuotojų_sk	,117	30	,200(*)	,974	30	,644

* This is a lower bound of the true significance.

4.2 pav. **Imties skirstinio normalumo tikrinimo rezultatas**

Lentelėje *Tests of Normality* pateikti Kolmogorovo–Smirnovo (*Kolmogorov-Smirnov*(*a*)) ir Šapiro–Vilko (*Shapiro-Wilk*) kriterijų skaičiavimo rezultatai. Kriterijų reikšmingumo lygmuo *p-level* (*Sig.*) > α = *0,05*, kintamojo skirstinio skirtumas nuo normaliojo statistiškai nereikšmingas.

2. Hipotezių formulavimas. Suformuluotos hipotezės išreiškia spėjimus apie tiriamus duomenis. Statistinių hipotezių klasifikacija pateikta 4.1 lentelėje.

4.1 lentelė. **Statistinių hipotezių klasifikacija**

Statistinės hipotezės					
Parametrinės hipotezės			*Neparametrinės hipotezės*		
Nulinė	*Alternatyvioji*		*Nulinė*	*Alternatyvioji*	
$\theta \in \Theta_0$	$\theta \in \Theta_1$		*X* ir *Y* skirstiniai nesiskiria	*X* ir *Y* skirstiniai skiriasi	
	Dvipusė	Vienpusė		Dvipusė	Vienpusė
$\theta_1 = \theta_0$	$\theta_1 \neq \theta_0$	$\theta_1 > \theta_0$	$P(X > t) = P(Y > t)$ kiekvienam t	$P(X > t) > P(Y > t)$ arba $P(X > t) < P(Y > t)$ kiekvienam t	$P(X > t) > P(Y > t)$ kiekvienam t
		$\theta_1 < \theta_0$			$P(X > t) < P(Y > t)$ kiekvienam t

Statistinės hipotezės skirstomos į nulines ir alternatyviąsias:

Nulinė hipotezė H_0 – tai spėjimas apie skirtumų nebuvimą. Ji teigia, kad lyginamųjų populiacijų parametrų, skirstinių arba populiacijų statistikų skirtumas yra atsitiktinis ir nereikšmingas.

Alternatyvioji hipotezė H_1 – tai spėjimas apie esančius skirtumus, kurių negalima paaiškinti atsitiktiniais svyravimais. Ji dažniausiai yra tai, ką mes norime įrodyti.

Nulinė ir alternatyvioji hipotezės gali būti *parametrinės* ir *neparametrinės*. Tuo atveju, kai duomenų skirstinys artimas normaliajam ir duomenys matuojami kiekybine skale, galima tikrinti teiginius apie tiriamos populiacijos parametrus (dažniausiai vidurkį ir dispersiją). Bet koks teiginys apie populiacijos parametro reikšmę vadinamas parametrine hipoteze. Statistinę parametrinę hipotezę sudaro du alternatyvūs teiginiai apie galimas parametro reikšmes. Problema formuluojama kaip spėjimas apie galimas parametro reikšmes, priklausančias aibei Θ_0, pateikiant alternatyvą, kad θ priklauso aibei Θ_1:

$$\begin{cases} H_0 : \theta \in \Theta_0, \\ H_1 : \theta \in \Theta_1, \end{cases} \tag{4.2}$$

čia: H_0 – parametrinė (nulinė) hipotezė,
H_1 – alternatyva (alternatyvioji hipotezė).

Alternatyvos skirstomos į dvipuses $\theta_1 \neq \theta_0$ ir vienpuses $\theta_1 > \theta_0$ (arba $\theta_1 < \theta_0$).

Tais atvejais, kai apie duomenų skirstinį nieko nežinome arba kai tiriamos mažos imtys, formuluojami teiginiai, nesusiję su parametrų įvertinimu. Tokios hipotezės vadinamos neparametrinėmis. Jei duomenys matuojami pavadinimų ir rangų skalėmis, dažniausiai formuluojamos neparametrinės hipotezės. Neparametrinių hipotezių atveju dažniausiai lyginami skirstiniai. Neparametrinės hipotezės taip pat būna vienpusės ir dvipusės. Tarkime, norime palyginti kintamųjų X ir Y skirstinius. Neparametrinės hipotezės formuluojamos taip:

$$\begin{cases} H_0\text{: } X \text{ ir } Y \text{ skirstiniai nesiskiria,} \\ H_1\text{: } X \text{ ir } Y \text{ skirstiniai skiriasi.} \end{cases} \tag{4.3}$$

Matematiškai dėl dvipusių hipotezių tai galima užrašyti kaip:

$$\begin{cases} H_0\text{: } P(X > t) = P(Y > t) \text{ kiekvienam } t, \\ H_1\text{: } P(X >) > P(Y > t) \text{ arba } P(X > t) < P(Y > t) \text{ kiekvienam } t. \end{cases} \quad (4.4)$$

Vienpusių alternatyvų statistinės hipotezės matematiškai užrašomos taip:

$$\begin{cases} H_0\text{: } P(X > t) = P(Y > t) \text{ kiekvienam } t, \\ H_1\text{: } P(X > t) > P(Y > t) \text{ kiekvienam } t, \end{cases} \quad (4.5)$$

arba

$$\begin{cases} H_0\text{: } P(X > t) = P(Y > t) \text{ kiekvienam } t, \\ H_1\text{: } P(X > t) < P(Y > t) \text{ kiekvienam } t. \end{cases} \quad (4.6)$$

3. Statistinio kriterijaus skaičiavimas. Taisyklė, kuria remiantis iš imties rezultatų darome išvadą apie hipotezės teisingumą ar klaidingumą, vadinama statistiniu kriterijumi. Statistiniai kriterijai skirstomi į parametrinius ir neparametrinius. Dauguma parametrinių hipotezių tikrinimo kriterijų gali būti taikomi tik tada, kai kintamojo skirstinys yra normalusis (kintamieji pasiskirstę pagal Gauso dėsnį). Praktiškai ši sąlyga ne visada tenkinama. Be to, mažoms imtims šios sąlygos patikrinti iš esmės neįmanoma. Yra grupė kriterijų, kuriuos taikant nereikalaujama, kad stebimojo kintamojo skirstinys būtų normalusis. Šie kriterijai vadinami nepriklausomais nuo skirstinio arba neparametriniais kriterijais. Jie nėra skirti hipotezėms apie populiacijų reikšmes tikrinti. Neparametriniai kriterijai gali būti taikomi mažoms imtims. Parametriniai ir neparametriniai kriterijai papildo vieni kitus.

Parametriniai kriterijai yra *galingesni*, suteikia didesnę tikimybę nustatyti lyginamų populiacijų parametrų skirtumus, tačiau jie ne tokie universalūs. Neparametriniai kriterijai yra universalesni, mažiau jautrūs išskirtims, tačiau jie silpnesni.

4. Statistinio reikšmingumo lygmens parinkimas. Tikrinant hipotezę dažniausiai taikomas toks principas: hipotezę atmetame, jei atsitinka tai, kas, esant teisingai hipotezei, atsitikti praktiškai negalėjo, arba tokio įvykio tikimybė yra labai maža. Tam pateikiama reikšmingumo lygmens sąvoka. Matematikoje reikšmingumo lygmuo paprastai žymimas raide α, $\alpha > 0$. Nusprendžiame įvykį laikyti iš esmės negalimu, jei

jo tikimybė mažesnė už α. Kuo mažesnė α reikšmė, tuo didesnė tikimybė, kad bus priimtas teisingas sprendimas. Priimdami arba atmesdami hipotezę H_0, galime padaryti dviejų rūšių klaidas. Pirmosios rūšies klaida: H_0 atmetame, kai ji teisinga. Antrosios rūšies klaida: H_0 priimame, kai ji klaidinga. Kalbant formalizuotai, galima pasakyti, kad reikšmingumo lygmuo – tai tikimybė atmesti H_0 hipotezę, kai ji yra teisinga (pirmosios rūšies klaida). Taisyklė, pagal kurią iš imties rezultatų daromе išvadą apie hipotezės teisingumą ar klaidingumą, vadinama statistiniu kriterijumi. Galimos statistinio kriterijaus taikymo baigtys pateiktos 4.2 lentelėje.

4.2 lentelė. **Pirmosios ir antrosios rūšies klaidos**

	H_0 teisinga	H_0 neteisinga
atmetame H_0	I rūšies klaida	neteisingas sprendimas
neatmetame H_0	teisingas sprendimas	II rūšies klaida

Reikšmingumo lygmuo, kuriuo remiantis hipotezė atmetama, yra sutartinis. Daugelyje mokslo sričių taikomas 5 procentų reikšmingumo lygmuo. Tačiau reikia atminti, kad šis reikšmingumo lygmuo susijęs su gana didele klaidos tikimybe (5 procentų). Pagal bendrai priimtą terminologiją, išvados, padarytos remiantis klaidos tikimybe $\alpha \leq 0,05$, vadinamos reikšmingomis, remiantis klaidos tikimybe $\alpha \leq 0,01$ labai reikšmingomis, o esant klaidos tikimybei $\alpha \leq 0,001$ maksimaliai reikšmingomis. Kartais literatūroje tokios išvados žymimos atitinkamai viena, dviem arba trimis žvaigždutėmis (žr. 4.3 lentelę).

Kompiuterinėse duomenų analizės programose skaičiuojamas *mažiausias* reikšmingumo lygmuo, su kuriuo teisinga nulinė hipotezė gali būti atmesta turimiems duomenims. Šis reikšmingumo lygmuo vadinamas *stebimuoju reikšmingumo lygmeniu p-level*.

4.3 lentelė. **Klaidos tikimybės interpretacija**

Klaidos tikimybė	*Reikšmingumas*	*Žymėjimas*
$\alpha > 0,05$	Nereikšminga	ns (non-significant)
$\alpha \leq 0,05$	Reikšminga	*
$\alpha \leq 0,01$	Labai reikšminga	**
$\alpha \leq 0,001$	Maksimaliai reikšminga	***

5. Sprendimo priėmimas. Visos galimos statistinio kriterijaus reikšmės dalijamos į dvi sritis: pirmoji sritis, į kurią patekus apskaičiuoto kriterijaus reikšmei nulinė hipotezė atmetama, ir antroji, į kurią patekus apskaičiuoto kriterijaus reikšmei nulinės hipotezės atmesti negalima, t. y. ji priimama. Skaičiai C_1, C_2, …, kurie atskiria hipotezės atmetimo ir neatmetimo sritis, vadinami *kritinėmis reikšmėmis*. Jei apskaičiuota (empirinė) statistinio kriterijaus reikšmė patenka į kritinę sritį, nulinė hipotezė atmetama. Jei apskaičiuota (empirinė) statistinio kriterijaus reikšmė nepatenka į kritinę sritį, nulinė hipotezė neatmetama (priimama).

Skaičiuojant kompiuteriu rezultatuose pateikiama apskaičiuota statistinio kriterijaus reikšmė ir stebimasis reikšmingumo lygmuo *p-level*. Tada sprendimo priėmimo taisyklė formuluojama taip:

- Jeigu apskaičiuotas stebimasis reikšmingumo lygmuo *p-level* yra mažesnis už pasirinktą reikšmingumo lygmenį α (*p-level* $< \alpha$), nulinė hipotezė atmetama.
- Jeigu apskaičiuotas stebimasis reikšmingumo lygmuo *p-level* yra didesnis arba lygus α (*p-level* $\geq \alpha$), nulinė hipotezė neatmetama (priimama).

KONTROLINIAI KLAUSIMAI

4.1. Nuo ko priklauso parametrinių arba neparametrinių hipotezių pasirinkimas?
4.2. Kaip priimamas sprendimas priimti arba atmesti hipotezę H_0?
4.3. Ką reiškia pirmosios ir antrosios rūšies klaidos?
4.4. Ką reiškia reikšmingumo lygmuo?
4.5. Pateikite statistinio kriterijaus sampratą.
4.6. Kada taikomi parametriniai ir neparametriniai kriterijai?

UŽDUOTYS

4.1 užduotis

Tikrinant respondentų amžiaus skirstinio normalumą gautas rezultatas (žr. 4.3 pav.). Paaiškinkite gautą rezultatą. Ar galima šiam kintamajam taikyti parametrinius kriterijus?

Tests of Normality

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Amžius	,198	60	,000	,874	60	,000

a Lilliefors Significance Correction

4.3 pav. **Imties skirstinio normalumo tikrinimo rezultatas**

5. RYŠIŲ ANALIZĖ

Ryšių analizė leidžia ne tik analizuoti kintamuosius atskirai, bet ir nustatyti jų tarpusavio ryšį. Galima išskirti parametrinius ir neparametrinius ryšius.

5.1. Kintamųjų, išmatuotų nominaliąja arba rangų skale su nedideliu kategorijų skaičiumi, statistiniai ryšiai

Praktiškai atliekant bet kokį tyrimą, duomenų analizė pradedama nuo dažnių lentelių, ypač kai analizuojami nominalieji ir ranginiai kintamieji. Paprasčiausia porinės dažnių lentelės forma yra 2 × 2 lentelė, gaunama tuo atveju, kai tiriami du kintamieji, kurių kiekvienas gali įgyti tik dvi reikšmes (dichotominiai kintamieji). Sudarius dviejų ar daugiau kintamųjų dažnių lenteles, taikomi suderinamumo kriterijai (Pirsono χ^2, Fišerio ir panašiai), kuriais tikrinama, ar tarp stebimų ir tikėtinų dažnių yra reikšmingas skirtumas. Nors Pirsono χ^2 suderinamumo kriterijų galima taikyti bet kokio tipo kintamiesiems, geriausią rezultatą ši analizė duoda analizuojant nominaliuosius bei ranginius kintamuosius. Pirsono χ^2 kriterijų galima taikyti ir 2 × 2, ir didesnėms lentelėms.

5.1.1. Pirsono χ^2 suderinamumo kriterijus

Pirsono χ^2 (*Pearson chi-square*) kriterijus remiasi tuo, kad apskaičiuojami *tikėtini* dažniai, t. y. dažniai, kurių reikėtų *tikėtis* tuo atveju, kai tarp kintamųjų nėra jokio ryšio. Pavyzdžiui, apklausta 30 vyrų ir 30 moterų, ar perkant kavą jie pasiduoda aplinkinių įtakai. Jei nėra jokio ryšio tarp lyties ir aplinkinių poveikio, reikėtų *tikėtis*, kad tiek pasiduodančių ir nepasiduodančių aplinkinių įtakai vyrų ir moterų skaičius bus maždaug vienodas.

Pavyzdžiui, reikia nustatyti, ar aplinkinių įtaka perkant kavą priklauso nuo lyties (žr. 5.1 lentelę).

5.1 lentelė. **Porinė dažnių lentelė**

		Lytis		*Iš viso*
		Vyras	*Moteris*	
Aplinkinių įtaka	ne	10	12	22
	taip	20	18	38
Iš viso		30	30	60

Pirsono χ^2 kriterijaus taikymas susijęs su kai kuriais apribojimais, taikomais stebėjimų skaičiui n. Šis kriterijus taikomas, kai stebėjimų skaičius n didesnis nei 30 ir bent 75 procentai dažnių lentelės narvelių tikėtini dažniai nemažesni kaip 5. Jei šios sąlygos netenkinamos, rezultatai gali būti nepatikimi.

Suformuluokime hipotezes:
Hipotezė H_0: aplinkinių įtaka perkant kavą nepriklauso nuo lyties,
Hipotezė H_1: aplinkinių įtaka perkant kavą priklauso nuo lyties.

Statistinio reikšmingumo lygmens parinkimas
Reikšmingumo lygmuo, kuriuo remiantis hipotezė atmetama, yra sutartinis. Daugelyje mokslo sričių taikomas 5 procentų reikšmingumo lygmuo. Pasirenkame $\alpha = 0,05$.

Pirsono χ^2 kriterijaus skaičiavimas
Analizuojant statistinius ryšius pagal χ^2 kriterijų, pirmiausia apskaičiuojami tikėtini dažniai, t. y. kokios turėtų būti lentelės narvelių reikšmės, jeigu tarp kintamųjų nebūtų statistinio ryšio. Du kintamieji yra tarpusavyje nepriklausomi, jei stebimi dažniai sutampa su tikėtinais dažniais. Tikėtini dažniai apskaičiuojami pagal formulę:

$$q_{ij} = \frac{f_{i*} f_{*j}}{n}; \qquad (5.1)$$

čia: q_{ij} – tikėtinas dažnumas narvelyje, esančiame *i*-tos eilutės ir *j*-tojo stulpelio sankirtoje,

f_{i*} – *i*-tosios eilutės dažnių suma, $f_{i*} = f_{i1} + f_{i2}$,

f_{*j} – *j*-tojo stulpelio dažnių suma, $f_{*j} = f_{1j} + f_{2j}$,

n – nagrinėjamų elementų skaičius, šiuo atveju $n = 60$.

Pirsono χ^2 kriterijus apskaičiuojamas pagal formulę:

$$\chi^2 = \sum_{i,j=1}^{n} \frac{\left(f_{ij} - q_{ij}\right)^2}{q_{ij}}, \qquad (5.2)$$

čia f_{ij} – gardelės, esančios *i*-tosios eilutės ir *j*-tojo stulpelio sankirtoje, tikrasis (stebėtas) dažnis.

Iš 5.1 lentelės duomenų gauta reikšmė $\chi^2 = 0{,}287$.
Paskui apskaičiuojami lentelės laisvės laipsniai:

$$df = (s - 1)(t - 1), \qquad (5.3)$$

čia: *df* – lentelės laisvės laipsnių skaičius,
s – eilučių skaičius,
t – stulpelių skaičius.

Nagrinėjamą lentelę sudaro dvi eilutės ir du stulpeliai. Taigi $df = (2 - 1)(2 - 1) = 1$.

Statistinių išvadų formulavimas, skaičiavimui netaikant statistinės programos

Reikia nustatyti, ar gauta χ^2 kriterijaus reikšmė yra pakankama, kad būtų galima teigti, jog tiriamieji požymiai yra priklausomi. Hipotezė apie kintamųjų nepriklausomumą yra atmetama, kai apskaičiuotos statistikos χ^2 reikšmė yra didesnė už kritinę reikšmę. Pasirinkome reikšmingumo lygmenį $\alpha = 0{,}05$. Kritinę χ^2 reikšmę galima rasti χ^2 skirstinio lentelėje (žr. Priedai. 1 lentelė). Šioje lentelėje reikšmingumo lygmens

$\alpha = 0{,}05$ χ^2 kritinė reikšmė yra $\chi^2_{0,05}(1) = 3{,}84$ ir ji didesnė už gautą reikšmę 0,287, todėl hipotezės H_0 atmesti negalima.

Dažnių lentelių ryšio stiprumo matai

Φ koeficientas (*Phi coefficient*)

Apskaičiavus χ^2 kriterijų nustatoma, ar tarp požymių yra statistinis ryšys. Norint tiksliau nustatyti ryšio stiprumą, taikomi nominaliųjų kintamųjų koreliacijos matai. Vienas iš tokių matų – *Φ* koeficientas. Šiuo matu įvertinamas koreliacinio ryšio stiprumas 2 × 2 lentelėse. *Φ* koeficiento reikšmės kinta nuo 0 (tarp požymių nėra jokio ryšio) iki 1 (tarp požymių yra labai stiprus ryšys):

$$\Phi^2 = \frac{\chi^2}{n} \text{ arba } \Phi = \sqrt{\frac{\chi^2}{n}}, \qquad (5.4)$$

čia n – imties dydis.

Šis koeficientas gali būti taikomas ir didesnėms nei 2 × 2 lentelėms, tačiau jose kartais *Φ* koeficiento reikšmė būna didesnė už 1. Norint išvengti šio trūkumo, skaičiavimams taikomi papildomi koeficientai.

Kramerio V koeficientas (*Cramer's V coefficient*) apskaičiuojamas pagal formulę:

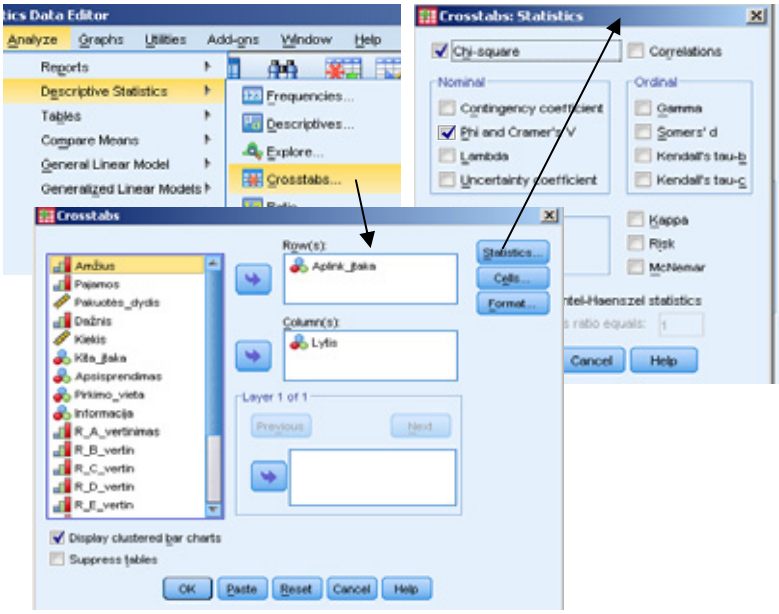
$$V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, c-1)}}, \qquad (5.5)$$

čia: r – dažnių lentelės eilučių skaičius,
c – stulpelių skaičius,
$min(r - 1, c - 1)$ – parenkama mažiausia reikšmė.

Šiuo kriterijumi įvertinama, kad eilučių ir stulpelių skaičius gali būti skirtingas. *Kramerio V koeficientas* gali būti taikomas didesnėms nei 2 × 2 lentelėms.

5.1.2. Pirsono χ^2 suderinamumo kriterijaus skaičiavimas SPSS terpėje

SPSS paketu patikrinsime, ar nuo lyties priklauso aplinkinių įtaka perkant kavą. Analizės dialogo langas atveriamas komanda *Analyze* → *Descriptive Statistics* → *Crosstabs...* (žr. 5.1 pav.). Kintamasis *Aplink_įtaka* perkeliamas į langelį *Row*(*s*), o kintamasis *Lytis* – į langelį *Column*(*s*). Kad būtų iškart nubraižytas porinių dažnių grafikas, reikia pažymėti langelį *Display clustered bar charts.* Mygtuko *Statistics* paspaudimu atveriamas dialogo langas *Crosstabs: Statistics.* Šiame lange reikia pažymėti langelį *Chi-square*, galima pažymėti ir langelį *Phi and Cramer's V*, jei reikia nustatyti ne tik ryšį, bet ir jo stiprumą (ryšio stiprumui nustatyti dėl nominaliųjų kintamųjų taikomi *Phi and Cramer's V* koeficientai – dialogo lango *Crosstabs: Statistics* sritis *Nominal*). Paspaudę mygtuką *Continue* grįžtame į pradinį langą. Procedūra baigiama mygtuko *OK* paspaudimu. Lange *Output-SPSS Viewer* rodomi analizės rezultatai (žr. 5.2 pav.).



5.1 pav. **Dialogo langai *Crosstabs* ir *Crosstabs: Statistics***

Pirmiausia pateikiama analizuojamų kintamųjų dažnių lentelė: *Aplink_įtaka*Lytis Crosstabulation*. Suformuluosime hipotezes:

H_0: aplinkinių įtaka nepriklauso nuo lyties,

H_1: aplinkinių įtaka priklauso nuo lyties.

Aplink_įtaka * Lytis Crosstabulation

Count

		Lytis		
		Vyras	Moteris	Total
Aplink_įtaka	ne	10	12	22
	taip	20	18	38
Total		30	30	60

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,287[a]	1	,592		
Continuity Correction[b]	,072	1	,789		
Likelihood Ratio	,287	1	,592		
Fisher's Exact Test				,789	,395
Linear-by-Linear Association	,282	1	,595		
N of Valid Cases	60				

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 11,00.

b. Computed only for a 2x2 table

5.2 pav. **Pirsono χ^2 kriterijaus skaičiavimo rezultatas**

Lentelėje *Chi-Square Tests* pateikti analizės rezultatai (žr. 5.2 pav.). Užrašas po lentele informuoja, ar kintamieji tinka analizei. χ^2 kriterijaus taikymas turi apribojimus. Skaičiuojant šį kriterijų taikomas aproksimacijos metodas. Pagrindiniai reikalavimai: aproksimacija laikoma gana tiksli, jei bendras imties dydis $m \geq 30$ elementų bei 75 procentai porinių dažnių lentelės narvelių tikėtini dažniai ≥ 5, nagrinėjami kintamieji tinkami Pirsono χ^2 kriterijaus analizei. Nurodytas skaičius narvelių su dažniais mažiau 5 (*0 cells have exspected count less than 5*) yra mažesnis, negu leidžiamas tokių narvelių skaičius (*The minimum expected count is 11,00*).

Pearson Chi-Square eilutės stulpelyje *Value* rodoma χ^2 kriterijaus reikšmė, stulpelyje *df* – lentelės laisvės laipsnių skaičius, stulpelyje *Asymp. Sig.* (*2-sided*) – stebimojo reikšmingumo lygmens *p-level* reikšmė. Skaičiuoti taikant statistinę programą, dėl hipotezės H_0 atmetimo sprendžiama iš stebimojo reikšmingumo lygmens *p-level*. *P-level = 0,592 > α = 0,05*, hipotezės H_0, kad aplinkinių įtaka nepriklauso nuo lyties, atmesti negalima. Eilutė *Linear-by-Linear Association* leidžia daryti išvadą dėl *nenominaliųjų* kintamųjų tiesinio ryšio – nagrinėjamu atveju šis parametras reikšmės neturi.

Ryšio stiprumui nustatyti dėl nominaliųjų kintamųjų taikomi *Phi and Cramer's V* testai (žr. 5.3 pav.). Lentelėje *Symmetric Measures* pateiktos šių testų reikšmės ir jų stebimas reikšmingumo lygmuo *p-level* (stulpelis *Approx. Sig.*). *P-level = 0,592 > α = 0,05*, koeficientai nereikšmingi. 5.4 pav. pateikta porinių dažnių stulpelinė diagrama.

Symmetric Measures

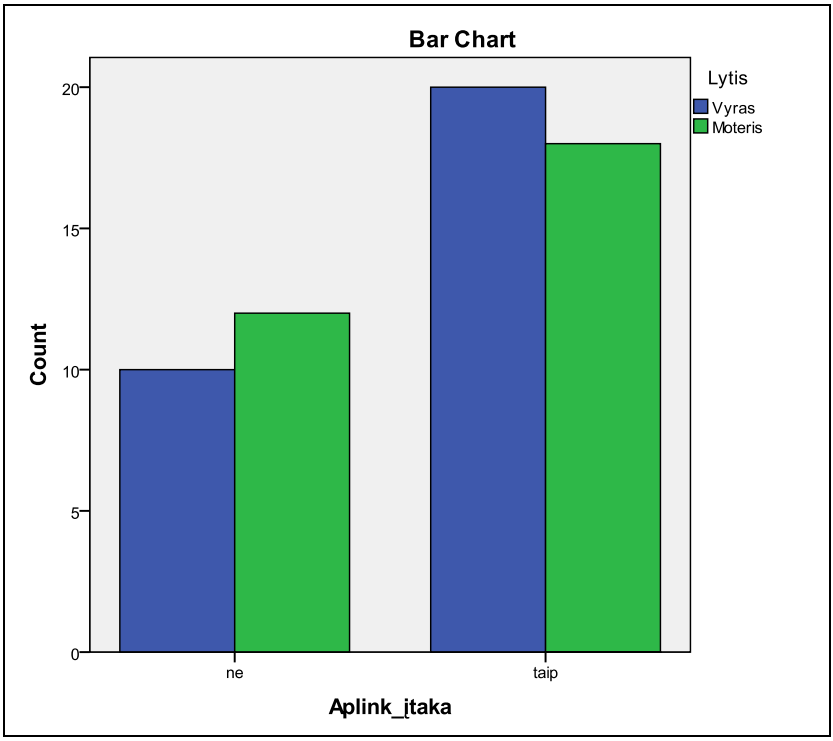
		Value	Approx. Sig.
Nominal by Nominal	Phi	-,069	,592
	Cramer's V	,069	,592
N of Valid Cases		60	

a Not assuming the null hypothesis.
b Using the asymptotic standard error assuming the null hypothesis.

5.3 pav. ***Phi* ir *Cramer's V* koeficientų skaičiavimo rezultatas**

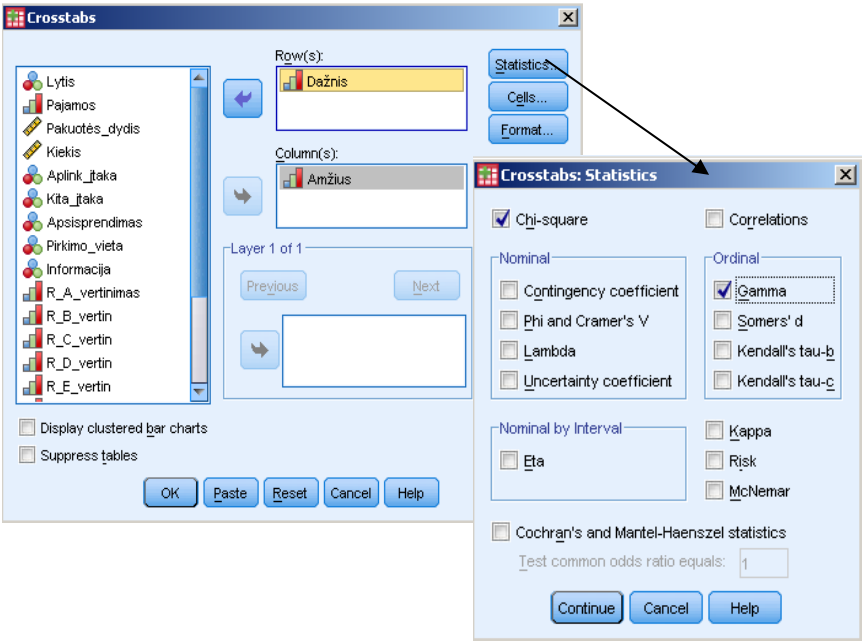
Atliktos procedūros sintaksė:

```
CROSSTABS
  /TABLES=Aplink_įtaka BY Lytis
  /FORMAT= AVALUE TABLES
  /STATISTIC=CHISQ PHI
  /CELLS= COUNT
  /COUNT ROUND CELL .
```



5.4 pav. **Porinių dažnių stulpelinė diagrama**

χ^2 kriterijumi patikrinsime, kaip pirkimo dažnis priklauso nuo amžiaus (ranginiai kintamieji). Analizės dialogo langas atveriamas komanda *Analyze* → *Descriptive Statistics* → *Crosstabs*. Kintamasis *Dažnis* perkeliamas į langelį *Row*(*s*), o kintamasis *Amžius* – į langelį *Column*(*s*). Kad būtų iškart nubraižytas porinių dažnių grafikas, reikia pažymėti langelį *Display clustered bar charts.* Mygtuko *Statistics* paspaudimu atveriamas dialogo langas *Crosstabs: Statistics.* Šiame lange reikia pažymėti langelį *Chi-square*, galima pažymėti ir langelį *Gamma* (dialogo lango *Crosstabs: Statistics* sritis *Ordinal*), jei reikia nustatyti ne tik ryšį, bet ir jo stiprumą (žr. 5.5 pav.). Paspaudę mygtuką *Continue* grįžtame į pradinį langą. Procedūra baigiama mygtuko *OK* paspaudimu. Lange *Output-SPSS Viewer* rodomi analizės rezultatai (žr. 5.6 pav.).



5.5 pav. **Dialogo langai *Crosstabs* ir *Crosstabs: Statistics***

Suformuluosime hipotezes:

H_0: pirkimo dažnis nepriklauso nuo amžiaus,

H_1: pirkimo dažnis priklauso nuo amžiaus.

Lentelėje *Chi-Square Tests* pateikti analizės rezultatai (žr. 5.6 pav.). Užrašas po lentele informuoja, kad kintamieji tinka analizei. *P-level* = *0,182* > α = *0,05*, hipotezės H_0, kad pirkimo dažnis nepriklauso nuo amžiaus, atmesti negalima. Eilutė *Linear-by-Linear Association* leidžia daryti išvadą dėl nenominaliųjų kintamųjų tiesinio ryšio, *p-level* = *0,707* > α = *0,05*, kintamųjų tiesinis ryšys nereikšmingas.

Apskaičiavus χ^2 kriterijų nustatoma, ar tarp požymių yra statistinis ryšys. Ryšio stiprumui nustatyti dėl *ranginių* kintamųjų taikomas *Gamma* testas (žr. 5.7 pav.). Lentelėje *Symmetric Measures* pateikta šio testo reikšmė (*Gamma* = *–0,068*) ir jos stebimas reikšmingumo lygmuo *p-level* (stulpelis *Approx. Sig.*). *P-level* = *0,710* > α = *0,05*, *Gamma* koe-

ficientas nereikšmingas. 5.8 pav. pateikta porinių dažnių stulpelinė diagrama.

Dažnis * Amžius Crosstabulation

Count

		Amžius			Total
		15–25 m.	26–35 m.	Daugiau nei 36 m.	
Dažnis	1,00	8	6	6	20
	2,00	6	7	8	21
	3,00	11	1	7	19
Total		25	14	21	60

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6,237(a)	4	,182
Likelihood Ratio	7,311	4	,120
Linear-by-Linear Association	,141	1	,707
N of Valid Cases	60		

a 1 cells (11,1 %) have expected count less than 5. The minimum expected count is 4,43.

5.6 pav. **Pirsono χ^2 kriterijaus skaičiavimo rezultatas**

Symmetric Measures

		Value	Asymp. Std. Error(a)	Approx. T(b)	Approx. Sig.
Ordinal by Ordinal	Gamma	-,068	,182	-,372	,710
N of Valid Cases		60			

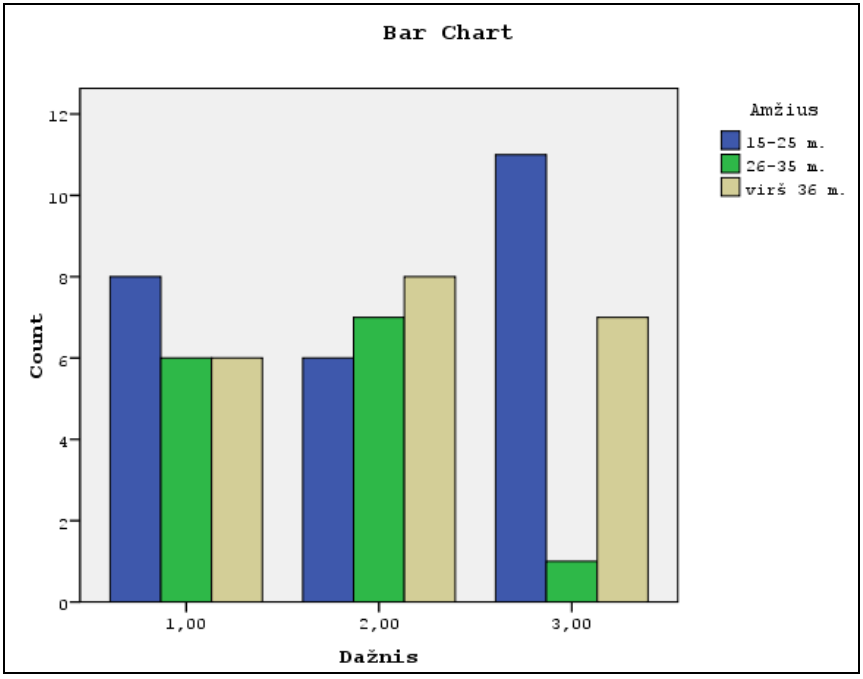
a Not assuming the null hypothesis.

b Using the asymptotic standard error assuming the null hypothesis.

5.7 pav. ***Gamma* koeficiento skaičiavimo rezultatas**

Atliktos procedūros sintaksė:

```
CROSSTABS
  /TABLES=Amžius BY Dažnis
  /FORMAT= AVALUE TABLES
  /STATISTIC=CHISQ GAMMA
  /CELLS= COUNT
  /COUNT ROUND CELL .
```



5.8 pav. **Porinių dažnių stulpelinė diagrama**

5.2. Koreliacinė analizė

Duomenys, apibūdinantys tiriamąjį objektą, dažnai būna susiję tarpusavyje arba priklauso vieni nuo kitų. Tokius ryšius galima nustatyti tik su tam tikra tikimybe. Statistinio ryšio (koreliacijos) stiprumą galima įvertinti koreliacijos koeficientais. Kalbant apie koreliaciją, vartojami du terminai: *koreliacinis ryšys* ir *koreliacinė priklausomybė*. Priklausomybė reiškia vieno kintamojo įtaką kitam, o ryšys – bendrą ar darnų dviejų kintamųjų kitimą. Koreliacijos koeficientas neparodo, kuris kintamasis yra priežastis, o kuris pasekmė. Jis tik nustato, ar kintamųjų kitimas yra susietas. Jei galima nustatyti, kad vieno kintamojo kitimas veikia kitus kintamuosius, šis kintamasis vadinamas nepriklausomuoju. Kintamieji, kurie, mūsų nuomone, kinta veikiant nepriklausomiesiems kintamiesiems, vadinami priklausomais. Dažniausiai tiriant sunku nustatyti, kuris kintamasis yra priklausomasis, o kuris nepriklausomasis. Tai gali numatyti tyrėjas, analizuodamas logiškai nagrinėjamo proceso veiksnius.

Koreliacinė analizė skirta *ranginių* arba *kiekybinių* kintamųjų ryšiui, jo stiprumui ir krypčiai nustatyti. Kiekybinių kintamųjų tiesiniam ryšiui įvertinti skaičiuojamas Pirsono (*Pearson*) koreliacijos koeficientas, o ranginių kintamųjų (tik skaitmeninių reikšmių) ryšiui įvertinti – Spirmeno (*Spearman*) koreliacijos koeficientas. Turinčius dvi kategorijas (binarinius) nominaliuosius kintamuosius koreliacinėje analizėje galima laikyti ranginiais kintamaisiais.

Didelėms imtims (kai *n* daugiau už 30) yra nusistovėjusios koreliacijos stiprumo nustatymo taisyklės (žr. 5.2 lentelę). Tiksliau įvertinti koreliacinį ryšį galima patikrinus hipotezę dėl koreliacijos koeficiento lygybės nuliui.

5.2 lentelė. **Koreliacijos stiprumo interpretacija**

Koreliacijos koeficiento reikšmė	*Interpretacija*
Nuo 0,9 iki 1,0 arba nuo –0,9 iki –1,0	Labai stipri koreliacija
Nuo 0,7 iki 0,9 arba nuo –0,7 iki –0,9	Stipri koreliacija
Nuo 0,5 iki 0,7 arba nuo –0,5 iki –0,7	Vidutinė koreliacija
Nuo 0,3 iki 0,5 arba nuo –0,3 iki –0,5	Silpna koreliacija
Nuo 0,3 iki –0,3	Nereikšminga koreliacija

5.2.1. Spirmeno (S*pearman*) koreliacijos koeficientas

Ryšio stiprumui tarp dviejų ranginių kintamųjų arba kiekybinių kintamųjų, kurių skirstinys nėra normalusis, taikomas Spirmeno (*Spearman*) koreliacijos koeficientas, kuris ryšio stiprumui įvertinti taiko ne pačias kintamųjų reikšmes, o jų rangus.

Skaičiuojant šį koreliacijos koeficientą lyginami dviejų sekų rangai:

$$r_S = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n^3 - n}, \qquad (5.6)$$

čia: r_s – Spirmeno koreliacijos koeficientas,
n – imties dydis,
D_i – skirtumas tarp vienos poros rangų.

Šis paprastas Spirmeno koeficiento skaičiavimo būdas taikomas tada, kai visi rangai skirtingi, t. y. kai tarp stebimų kintamųjų nebuvo pasikartojančių reikšmių. Jeigu yra sutampančių rangų, Spirmeno koreliacijos koeficientas skaičiuojamas pagal apibendrintą formulę:

$$r_S = \frac{\sum_{i=1}^{n}\left(R_{xi} - \frac{n+1}{2}\right)\left(R_{yi} - \frac{n+1}{2}\right)}{\sqrt{\sum_{i=1}^{n}\left(R_{xi} - \frac{n+1}{2}\right)^2}\sqrt{\sum_{i=1}^{n}\left(R_{yi} - \frac{n+1}{2}\right)^2}}, \qquad (5.7)$$

čia: $R_{x1}, R_{x2}, ..., R_{xn}$ – pirmosios sekos rangai,
$R_{y1}, R_{y2}, ..., R_{yn}$ – antrosios sekos rangai.

Jeigu imties dydis yra $n \geq 10$, tai Spirmeno koreliacijos koeficiento skirstinys yra apytiksliai Stjudento t skirstinys su $(n - 2)$ laisvės laipsniais. Hipotezę dėl šio koreliacijos koeficiento reikšmingumo galima tikrinti t kriterijumi:

$$t = r_S \sqrt{\frac{n-2}{1-r_S^2}}, \qquad (5.8)$$

čia: r_S – apskaičiuotas Spirmeno koreliacijos koeficientas,

n – imties dydis.

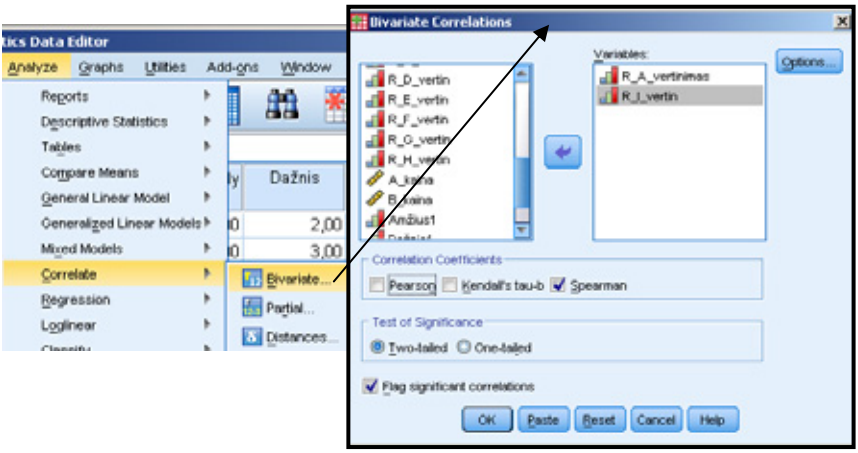
Sprendimo priėmimo taisyklė pasirenkama iš skaičiavimo rezultatų pateikimo būdo (ar skaičiavimams taikyta statistinė programa, ar kitokia skaičiavimo priemonė).

Kai statistinio t kriterijaus reikšmė apskaičiuota ir stebimasis reikšmingumo lygmuo *p-level* nežinomas, sprendimą galima priimti lyginant apskaičiuotą ir kritinę t reikšmes. Parenkamas reikšmingumo lygmuo α, statistinėje lentelėje (žr. Priedai. 2 lentelė) randama Stjudento skirstinio su (n – *2*) laisvės laipsniais *α/2* lygmens kritinė reikšmė $t_{\alpha/2}(n-2)$, čia n – imties dydis. Hipotezė H_0 atmetama, jei apskaičiuota t kriterijaus absoliuti reikšmė didesnė nei $t_{\alpha/2}(n-2)$.

Kai skaičiuoti taikoma statistinė programa, ar atmesti hipotezę H_0 sprendžiama iš stebimojo reikšmingumo lygmens *p-level*. Jei stebimojo reikšmingumo lygmens *p-level* reikšmė mažesnė už pasirinktą reikšmingumo lygmenį α (*p-level* < α), tai hipotezė H_0 atmetama ir daroma išvada, kad koreliacija yra reikšminga. Jei *p-level* reikšmė didesnė ar tokia pati, kaip reikšmingumo lygmuo α (*p-level* ≥ *α*), tai koreliacija nereikšminga, o kintamieji nepriklausomi.

5.2.2. Spirmeno koreliacijos koeficiento skaičiavimas SPSS terpėje

SPSS paketu patikrinsime, ar yra ryšys tarp skirtingos kavos rūšies įvertinimo. Analizės dialogo langas *Bivariate Correlations* atveriamas komanda *Analyze* → *Correlate* → *Bivariate...* (žr. 5.9 pav.). Kad nustatytume ryšį tarp A rūšies ir I rūšies įvertinimo, kintamieji *R_A_vertinimas* ir *R_I_vertin* perkeliami į langelį *Variables*. Paskui pažymimas langelis *Spearman*. Procedūra baigiama mygtuko *OK* paspaudimu. Lange *Output-SPSS Viewer* rodomi analizės rezultatai (žr. 5.10 pav.).



5.9 pav. **Dialogo langas *Bivariate Correlations***

Spirmeno koreliacijos koeficientas (*Spearman's rho*) $r_s = –0,556$, koreliacija vidutinė, neigiama: kuo geriau vertinama A rūšis, tuo blogiau vertinama I rūšis.

Gavus skaičiavimo rezultatus tenka priimti sprendimą dėl koreliacijos koeficiento reikšmingumo. Koreliacijos koeficiento reikšmingumui tikrinti suformuluojame hipotezes:

H_0: koreliacijos koeficientas lygus nuliui,

H_1: koreliacijos koeficientas nelygus nuliui.

Kai skaičiuoti taikoma statistinė programa, ar atmesti hipotezę H_0 sprendžiama iš stebimojo reikšmingumo lygmens *p-level*. Jei stebimojo reikšmingumo lygmens *p-level* reikšmė mažesnė už pasirinktą reikšmingumo lygmenį α (*p-level* < α), tai hipotezė H_0 atmetama ir daroma išvada, kad koreliacija yra reikšminga. Jei *p-level* reikšmė didesnė ar tokia pati, kaip reikšmingumo lygmuo α (*p-level* ≥ α), tai koreliacija nereikšminga, o kintamieji nepriklausomi.

Nagrinėjamu atveju stebimasis reikšmingumo lygmuo (*Sig. (2-tailed)*) *p-level = 0,000*, tai mažiau už pasirinktą reikšmingumo lygmenį $\alpha = 0,05$. Patvirtinta hipotezė H_1.

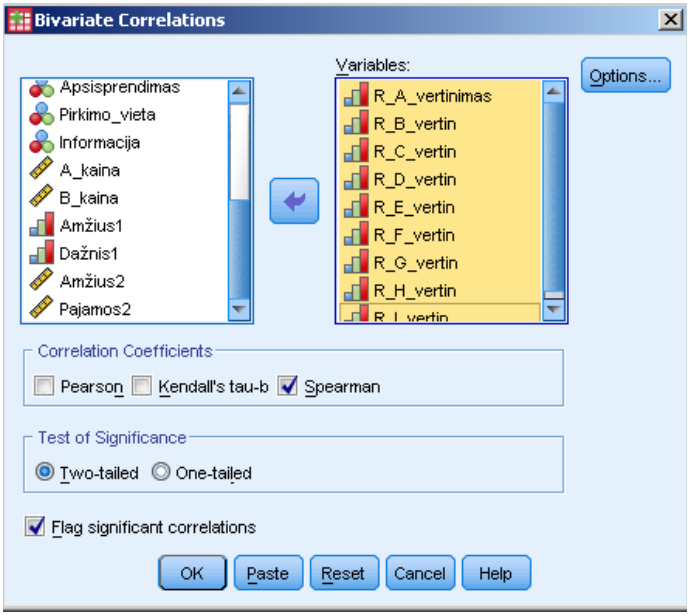
Correlations				
			R_A_vertinimas	R_I_vertin
Spearman's rho	R_A_vertinimas	Correlation Coefficient	1,000	-,556**
		Sig. (2-tailed)	.	,000
		N	60	60
	R_I_vertin	Correlation Coefficient	-,556**	1,000
		Sig. (2-tailed)	,000	.
		N	60	60

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

5.10 pav. **Spirmeno koreliacijos koeficiento skaičiavimo rezultatas**

Atliktos procedūros sintaksė:

```
NONPAR CORR
  /VARIABLES=R_A_vertinimas R_I_vertin
  /PRINT=SPEARMAN TWOTAIL NOSIG
  /MISSING=PAIRWISE .
```



5.11 pav. **Dialogo langas *Bivariate Correlations***

SPSS paketu galima nustatyti kelių kintamųjų koreliacijos koeficientus vienu metu. Į dialogo lango *Bivariate Correlations* langelį *Variables* perkeliami kintamieji *R_A_vertinimas*, *R_B_vertin*, *R_C_vertin*, *R_D_vertin*, *R_E_vertin*, *R_F_vertin*, *R_G_vertin*, *R_H_vertin* ir *R_I_vertin.* Tada pažymimas langelis *Spearman* (žr. 5.11 pav.). Procedūra baigiama mygtuko *OK* paspaudimu. Lange *Output-SPSS Viewer* rodomi analizės rezultatai (žr. 5.12 pav.). Lentelėje *Correlations* pateikti kintamųjų poriniai koreliacijos koeficientai ir jų reikšmingumo įverčiai.

Correlations

			R_A_vertinimas	R_B_vertin	R_C_vertin	R_D_vertin	R_E_vertin	R_F_vertin	R_G_vertin	R_H_vertin	R_I_vertin
Spearman's rho	R_A_vertinimas	Correlation Coefficient	1,000	-,055	,116	-,280*	,094	-,225	-,136	,150	-,556**
		Sig. (2-tailed)	.	,678	,377	,030	,474	,084	,299	,254	,000
		N	60	60	60	60	60	60	60	60	60
	R_B_vertin	Correlation Coefficient	-,055	1,000	,016	-,165	-,167	,029	-,013	-,238	-,107
		Sig. (2-tailed)	,678	.	,906	,209	,202	,827	,922	,067	,414
		N	60	60	60	60	60	60	60	60	60
	R_C_vertin	Correlation Coefficient	,116	,016	1,000	-,390**	,119	-,149	-,224	,029	,023
		Sig. (2-tailed)	,377	,906	.	,002	,365	,256	,086	,824	,861
		N	60	60	60	60	60	60	60	60	60
	R_D_vertin	Correlation Coefficient	-,280*	-,165	-,390**	1,000	-,102	-,079	-,073	-,237	,125
		Sig. (2-tailed)	,030	,209	,002	.	,439	,548	,580	,069	,342
		N	60	60	60	60	60	60	60	60	60
	R_E_vertin	Correlation Coefficient	,094	-,167	,119	-,102	1,000	-,183	-,085	-,012	-,238
		Sig. (2-tailed)	,474	,202	,365	,439	.	,161	,520	,925	,067
		N	60	60	60	60	60	60	60	60	60
	R_F_vertin	Correlation Coefficient	-,225	,029	-,149	-,079	-,183	1,000	-,206	-,081	,012
		Sig. (2-tailed)	,084	,827	,256	,548	,161	.	,115	,538	,930
		N	60	60	60	60	60	60	60	60	60
	R_G_vertin	Correlation Coefficient	-,136	-,013	-,224	-,073	-,085	-,206	1,000	-,311*	-,132
		Sig. (2-tailed)	,299	,922	,086	,580	,520	,115	.	,016	,316
		N	60	60	60	60	60	60	60	60	60
	R_H_vertin	Correlation Coefficient	,150	-,238	,029	-,237	-,012	-,081	-,311*	1,000	,014
		Sig. (2-tailed)	,254	,067	,824	,069	,925	,538	,016	.	,913
		N	60	60	60	60	60	60	60	60	60
	R_I_vertin	Correlation Coefficient	-,556**	-,107	,023	,125	-,238	,012	-,132	,014	1,000
		Sig. (2-tailed)	,000	,414	,861	,342	,067	,930	,316	,913	.
		N	60	60	60	60	60	60	60	60	60

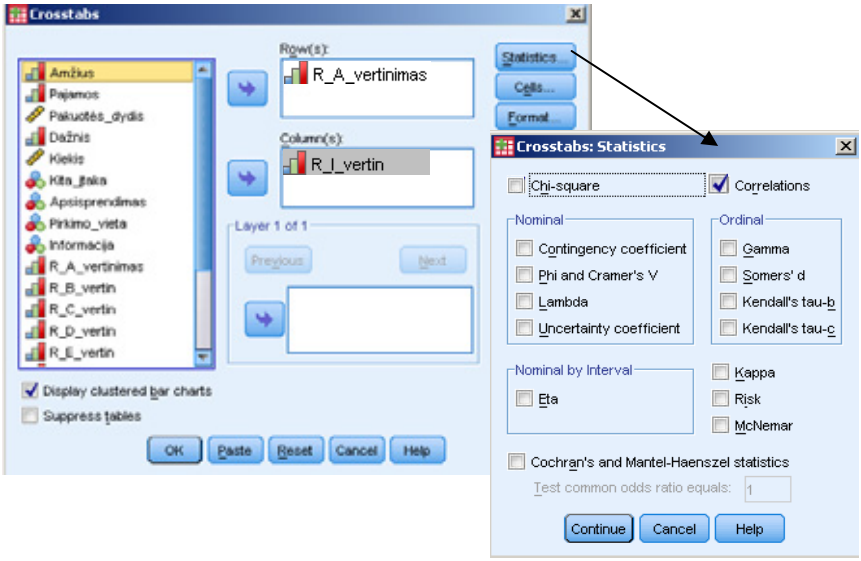
*. Correlation is significant at the 0.05 level (2-tailed).

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

5.12 pav. **Spirmeno koreliacijos koeficiento skaičiavimo rezultatas**

Apskaičiuoti Spirmeno koreliacijos koeficientą galima ir kitaip. Patikrinsime, ar yra ryšys tarp skirtingos kavos rūšies įvertinimo. Analizės dialogo langas atveriamas komanda *Analyze → Descriptive Statistics →*

Crosstabs.... Kintamasis *R_A_vertinimas* perkeliamas į langelį *Row*(*s*), o kintamasis *R_I_vertin* – į langelį *Column*(*s*). Mygtuko *Statistics* paspaudimu atveriamas dialogo langas *Crosstabs: Statistics.* Šiame lange reikia pažymėti langelį *Correlations* (žr. 5.13 pav.). Paspaudę mygtuką *Continue* grįžtame į pradinį langą. Procedūra baigiama mygtuko *OK* paspaudimu. Lange *Output-SPSS Viewer* rodomi analizės rezultatai (žr. 5.14 pav.).



5.13 pav. **Dialogo langas *Crosstabs: Statistics***

Symmetric Measures

		Value	Asymp. Std. Er-ror(a)	Approx. T(b)	Approx. Sig.
Interval by Interval	Pearson's R	-,544	,091	-4,932	,000(c)
Ordinal by Ordinal	Spearman Correlation	-,556	,088	-5,093	,000(c)
N of Valid Cases		60			

a Not assuming the null hypothesis.
b Using the asymptotic standard error assuming the null hypothesis.
c Based on normal approximation.

5.14 pav. **Spirmeno koreliacijos koeficiento skaičiavimo rezultatas**

Šiuo atveju lentelėje *Symmetric Measures* pateikti du koreliacijos koeficientai: Pirsono ir Spirmeno. Tyrėjas turi žinoti, koks koeficientas tinkamas. Nagrinėjamas ranginių kintamųjų ryšys, Spirmeno koreliacijos koeficientas (*Spearman Correlation*) $r_s = –0,556$, koreliacija vidutinė, neigiama: kuo geriau vertinama A rūšis, tuo blogiau vertinama I rūšis. Stebimasis reikšmingumo lygmuo (*Approx. Sig.*) *p-level = 0,000*, tai mažiau už pasirinktą reikšmingumo lygmenį $\alpha = 0,05$. Patvirtinta hipotezė H_1. Galima pasakyti, kad rezultatai nesiskiria, atsižvelgiant į tai, koks skaičiavimo būdas taikomas. Lentelėje *Symmetric Measures* pateikta ir *t* kriterijaus reikšmė (*Approx. T*(*b*)). Norint koreliacijos koeficiento reikšmingumą galima patikrinti ir *T* statistika (žr. 5.2.1 poskyrį).

Atliktos procedūros sintaksė:

```
CROSSTABS
  /TABLES=R_A_vertinimas BY R_I_vertin
  /FORMAT= AVALUE TABLES
  /STATISTIC=CORR
  /CELLS= COUNT
  /COUNT ROUND CELL .
```

5.2.3. Kiti kokybinių kintamųjų ryšio koeficientai

Eta (*Nominal by Interval*) koeficientas yra taikomas tada, kai nepriklausomasis kintamasis yra nominalusis arba ranginis, o priklausomasis kintamasis – kiekybinis. Nominalusis arba ranginis kintamasis turi turėti skaitmeninį kodavimą (*Numeric* tipo).

Patikrinsime, ar yra pajamų ir pakuotės dydžio ryšys. Analizės dialogo langas atveriamas komanda *Analyze → Descriptive Statistics → Crosstabs*.... Kintamasis *Pajamos* perkeliamas į langelį *Row*(*s*), o kintamasis *Pakuotės_dydis* – į langelį *Column*(*s*). Mygtuko *Statistics* paspaudimu atveriamas dialogo langas *Crosstabs: Statistics.* Šiame lange reikia pažymėti langelį *Eta* srityje *Nominal by Interval.* Paspaudę mygtuką *Continue* grįžtame į pradinį langą. Procedūra baigiama mygtuko *OK* paspaudimu. Lange *Output-SPSS Viewer* rodomi analizės rezultatai (žr. 5.15 pav.).

Directional Measures

			Value
Nominal by Interval	Eta	Pajamos Dependent	,569
		Pakuotės_dydis Dependent	,616

5.15 pav. ***Eta* (*Nominal by Interval*) koreliacijos koeficiento skaičiavimo rezultatas**

Nagrinėjamu atveju *Pajamos* – ranginis kintamasis (nepriklausomasis), o *Pakuotės_dydis* – kiekybinis kintamasis (priklausomasis). Pasirenkame koeficiento reikšmę iš eilutės *Pakuotės_dydis Dependent.* Eta koeficiento reikšmė *0,616*. Egzistuoja vidutinė teigiama pakuotės dydžio priklausomybė nuo pajamų.

Jeigu suranguotume pakuotės dydžius (naujas kintamasis *Pakuotė*) ir apskaičiuotume kintamųjų *Pajamos* ir *Pakuotė* Spirmeno koreliacijos koeficientą, gautume $r_s = 0,528$ (žr. 5.16 pav.). Šis koeficientas rodo silpnesnį kintamųjų ryšį nei *Eta* koeficientas.

Symmetric Measures

		Value	Asymp. Std. Error(a)	Approx. T(b)	Approx. Sig.
Interval by Interval	Pearson's R	,462	,097	3,755	,000(c)
Ordinal by Ordinal	Spearman Correlation	,528	,098	4,483	,000(c)
N of Valid Cases		54			

a Not assuming the null hypothesis.
b Using the asymptotic standard error assuming the null hypothesis.
c Based on normal approximation.

5.16 pav. **Spirmeno koreliacijos koeficiento skaičiavimo rezultatas**

Atliktos procedūros sintaksė:

```
CROSSTABS
  /TABLES=Pajamos BY Pakuotės_dydis
  /FORMAT= AVALUE TABLES
  /STATISTIC=ETA
  /CELLS= COUNT
  /COUNT ROUND CELL .
```

Rizikos laipsnis (*Risk*) skaičiuojamas 2 × 2 dažnių lentelėms, sudarytoms laikantis tam tikrų toliau išdėstytų reikalavimų. Skaičiuojant rizikos matą, analizuojamas vadinamasis rizikos kintamasis, kuris turi dvi kategorijas ir nurodo, įvyko tam tikras įvykis ar ne. Rizikos kintamasis nagrinėjamas priežastinio (nepriklausomojo) kintamojo atžvilgiu, kuris irgi turi būti binarinis (turintis dvi kategorijas). Pavyzdžiui, atlikta 60 respondentų apklausa apie aplinkinių įtaką perkant kavą. Apklausos rezultatai pateikti 5.3 lentelėje.

5.3 lentelė. **Porinė dažnių lentelė**

Aplinkinių įtaka	*Taip*	*Ne*
Vyrai	a = 20	b = 10
Moterys	c = 18	d = 12

Aplinkinių įtaka yra rizikos kintamasis, o *lytis* – nepriklausomasis (priežastinis) kintamasis. SPSS rizikos matas apskaičiuojamas pagal šias formules:

$$R_0 = \frac{ad}{bc}, \text{ galimybių santykis (}\textit{Odds Ratio}\text{)}, \qquad (5.9)$$

$$R_1 = \frac{a(c+d)}{c(a+b)}, \text{ santykinės rizikos koeficientas}, \qquad (5.10)$$

$$R_2 = \frac{b(c+d)}{d(a+b)}, \text{ santykinės rizikos koeficientas}, \qquad (5.11)$$

čia a, b, c, d – lentelėje pateikti dažniai.

Apskaičiuosime riziką pasiduoti aplinkinių įtakai. Analizės dialogo langas atveriamas komanda *Analyze → Descriptive Statistics → Crosstabs*.... Nepriklausomasis kintamasis *Lytis* perkeliamas į langelį *Row*(*s*), o rizikos kintamasis *Aplink_įtaka* – į langelį *Column*(*s*). Mygtuko *Statistics* paspaudimu atveriamas dialogo langas *Crosstabs: Statistics*. Šiame lange reikia pažymėti langelį *Risk*. Paspaudę mygtuką *Continue* grįžtame į pradinį langą. Procedūra baigiama mygtuko *OK* paspaudimu. Lange *Output-SPSS Viewer* rodomi analizės rezultatai (žr. 5.17 pav.).

Risk Estimate

	Value	95 % Confidence Interval	
		Upper	Lower
Odds Ratio for Lytis (Moteris / Vyras)	1,333	,465	3,823
For cohort Aplink_įtaka = ne	1,200	,614	2,344
For cohort Aplink_įtaka = taip	,900	,611	1,325
N of Valid Cases	60		

5.17 pav. **Rizikos laipsnio įvertinimo rezultatas**

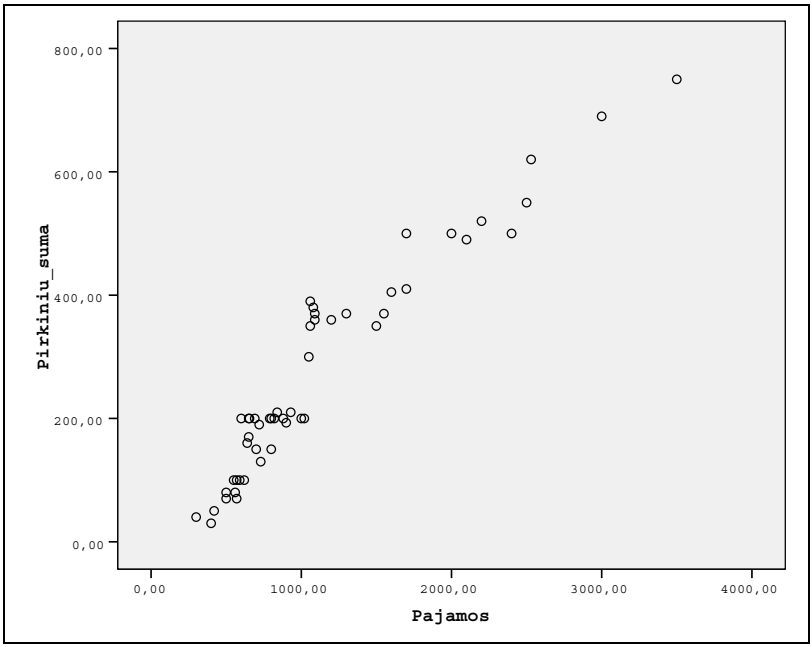
Rizikos įvertinimas (*Risk Estimate*) lentelėje pateiktas dydis – galimybių santykis (*Odds Ratio for Lytis* (*Moteris / Vyras*)) rodo, kad rizika pasiduoti aplinkinių įtakai moterims vidutiniškai yra 1,333 karto didesnė, negu vyrams. Reikšmės *Lower = 0,465* ir *Upper = 3,823* rodo 95 procentų rizikos pasikliautinojo intervalo $[\bar{x} - 2s;\ \bar{x} + 2s]$ ribas, čia $\bar{x}$ – rizikos santykio reikšmės vidurkis, *s* – šios reikšmės standartinis nuokrypis. Galima teigti, kad su 95 procentų patikimumu rizika pasiduoti aplinkinių įtakai moterims sudaro nuo 0,465 iki 3,823 vyrų rizikos. Kiti du koeficientai vadinami santykinės rizikos koeficientais.

Atliktos procedūros sintaksė:

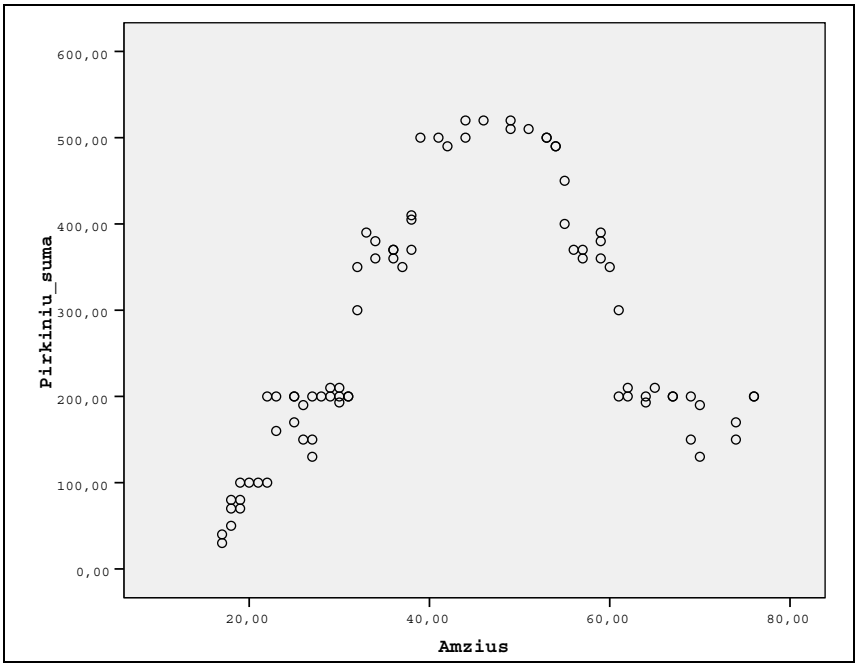
```
CROSSTABS
  /TABLES=Lytis1 BY Aplink_įtaka
  /FORMAT= AVALUE TABLES
  /STATISTIC=CORR RISK
  /CELLS= COUNT
  /COUNT ROUND CELL .
```

5.2.4. Pirsono koreliacijos koeficientas

Koreliaciniai ryšiai skiriasi pagal formą, kryptį ir stiprumą. Pagal *formą* koreliaciniai ryšiai skirstomi į tiesinius ir kreivinius. Ryšio formą galima pamatyti analizuojant taškines (*Scatter*) diagramas. Tiesinis ryšys gali būti, pavyzdžiui, tarp pirkinių sumos ir pajamų (žr. 5.18 pav.). Netiesiniai ryšiai gali būti stebimi, tarkim, tarp pirkinių sumos ir amžiaus (žr. 5.19 pav.).



5.18 pav. **Tiesinis koreliacinis ryšys**

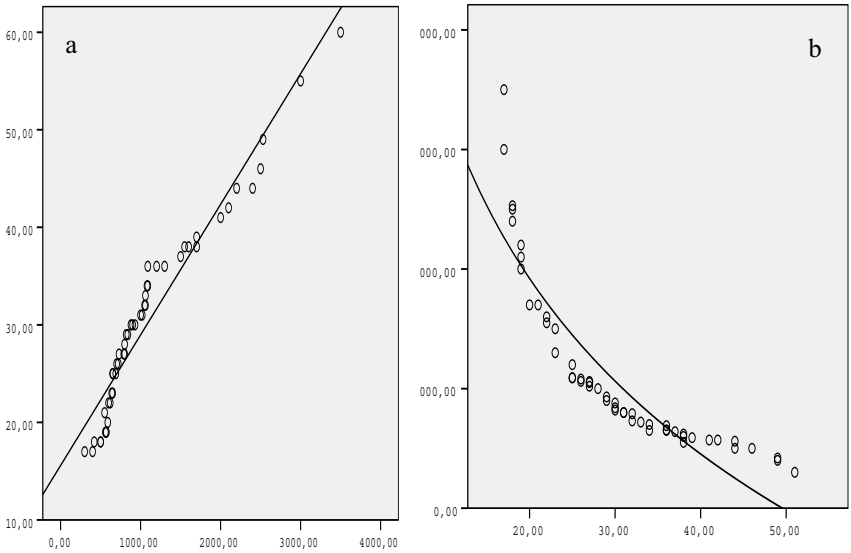


5.19 pav. **Netiesinis koreliacinis ryšys**

Pagal *kryptį* koreliaciniai ryšiai yra teigiami (*tiesioginiai*) ir neigiami (*atvirkštiniai*). Esant teigiamam ryšiui, vieno kintamojo reikšmėms didėjant, didėja ir kito kintamojo reikšmės (5.20 pav., a). Esant neigiamam ryšiui, vieno kintamojo reikšmėms didėjant, kito kintamojo reikšmės mažėja (5.20 pav., b).

Kiekybinių kintamųjų ryšio *stiprumą* galima išmatuoti koreliacijos koeficientu. Didelės šio koeficiento reikšmės, nepaisant ar jos teigiamos, ar neigiamos, atitinka tai, ką vadiname stipria koreliacija, o mažos reikšmės – silpna koreliacija. Jei koreliacija yra nereikšminga, tai nereiškia, kad koreliacijos koeficientas tiksliai lygus nuliui, tačiau jo reikšmė yra arti nulio. Kiekybinių kintamųjų tyrimui taikomi keli skirtingi koreliacijos koeficientai. Dažniausiai taikomas tiesinio ryšio Pirsono (*Pearson*) koreliacijos koeficientas *r*. Jei Pirsono koreliacijos koeficientas rodo kintamųjų statistinio ryšio nebuvimą, tai nereiškia, kad šio

ryšio visiškai nėra, tarp kintamųjų nėra *tiesinio* ryšio, bet gali būti sudėtingesnės formos ryšys.



5.20 pav. **Teigiamas tiesinis ir neigiamas netiesinis koreliacinis ryšys**

Empirinis Pirsono koreliacijos koeficientas yra tiesinio ryšio stiprumo tarp kintamųjų *X* ir *Y* matas. Jis apskaičiuojamas iš turimų stebėjimų pagal formulę:

$$r = \frac{\frac{1}{n-1}\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{S_x S_y}, \qquad (5.12)$$

čia: $\bar{x}$ ir $\bar{y}$ atitinkamai yra stebėjimų *x* ir *y* vidurkiai,

S_x ir S_y – standartiniai nuokrypiai.

Koreliacijos koeficiento reikšmingumas tikrinamas *t* kriterijumi. Suformuluokime statistines hipotezes:

H_0: koreliacijos koeficientas lygus nuliui,

H_1: koreliacijos koeficientas nelygus nuliui.

Hipotezės H_0 teisingumui tikrinti taikysime *t* kriterijų:

$$t = r\sqrt{\frac{n-2}{1-r^2}}\,, \tag{5.13}$$

čia: *r* – apskaičiuota koreliacijos koeficiento reikšmė,

n – imties dydis.

Parenkamas reikšmingumo lygmuo α. Hipotezė H_0 atmetama, jei absoliuti *t* reikšmė viršija Stjudento skirstinio su (*n–2*) laisvės laipsniais *α/2* lygmens kritinę reikšmę $t_{\alpha/2(n-2)}$.

Dalinis koreliacijos koeficientas. Kai kintamieji *X* ir *Y* nagrinėjami kartu su kitais kintamaisiais (*Z*, *W*,…), dažnai būna, kad kintamieji *X* ir *Y* kinta dėl to, kad kinta kiti kintamieji. Dalinis koreliacijos koeficientas parodo tiesinio ryšio tarp *X* ir *Y* stiprumą tuo atveju, kai pašalinama kitų kintamųjų įtaka. Trijų kintamųjų atveju dalinį koreliacijos koeficientą ρ galima apskaičiuoti pagal formulę:

$$\rho_{XY(Z)} = \frac{r_{XY} - r_{xz}r_{YZ}}{\sqrt{\left(1-r_{XZ}^2\right)\left(1-r_{YZ}^2\right)}}\,, \tag{5.14}$$

čia r_{XY} – koreliacijos koeficientas tarp kintamųjų *X* ir *Y*.

Determinacijos koeficientas. Vertinant pasirinkto kintamųjų ryšio modelio tinkamumą analizuojamos liekanos ir skaičiuojamas *determinacijos* koeficientas. Determinacijos koeficientas parodo, kuri vieno požymio bendro kitimo dalis gali būti paaiškinta kito požymio reikšmių kitimu. Šis koeficientas interpretuojamas taip:

$$R^2 = 1 - \frac{S_\varepsilon^2}{S_y^2} = \frac{\text{Variacijos dalis, kurią paaiškina modelis}}{\text{Visa variacija}}, \quad (5.15)$$

čia: S_ε^2 – liekanų (ε) dispersija,

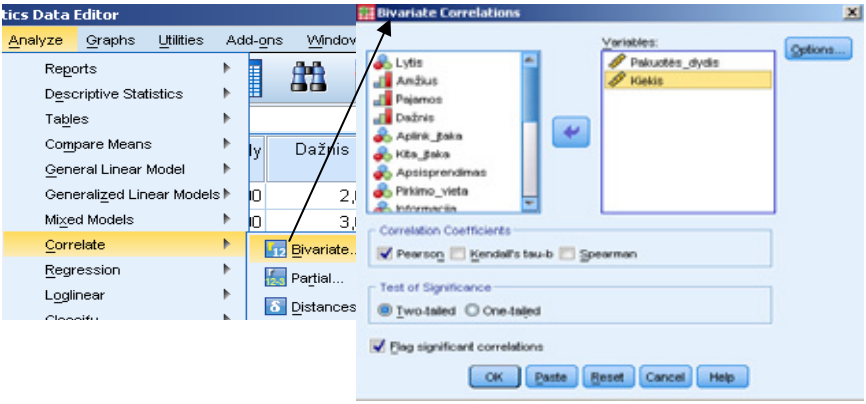
S_y^2 – stebimų kintamojo (Y) reikšmių dispersija.

Determinacijos koeficientas R^2 esant *tiesinei* priklausomybei yra lygus koreliacijos tarp kintamųjų (požymių) X ir Y koeficiento kvadratui.

$$R^2 = r_{xy}^2. \quad (5.16)$$

5.2.5. Pirsono koreliacijos koeficiento skaičiavimas SPSS terpėje

Patikrinsime, ar yra ryšys tarp pakuotės dydžio ir perkamos kavos kiekio. Analizės dialogo langas *Bivariate Correlations* atveriamas komanda *Analyze* → *Correlate* → *Bivariate...* (žr. 5.21 pav.). Kintamieji *Pakuotės_dydis* ir *Kiekis* perkeliami į langelį *Variables.* Tada reikia pažymėti langelį *Pearson*. Procedūra baigiama mygtuko *OK* paspaudimu. Lange *Output-SPSS Viewer* rodomi analizės rezultatai (žr. 5.22 pav.).



5.21 pav. **Dialogo langas *Bivariate Correlations***

Correlations

		Pakuotės_dydis	Kiekis
Pakuotės_dydis	Pearson Correlation	1	,760**
	Sig. (2-tailed)		,000
	N	60	60
Kiekis	Pearson Correlation	,760**	1
	Sig. (2-tailed)	,000	
	N	60	60

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

5.22 pav. **Pirsono koreliacijos koeficiento skaičiavimo rezultatas**

Pirsono koreliacijos koeficientas $r = 0,760$, koreliacija stipri ir teigiama. Koreliacijos koeficiento reikšmingumui tikrinti suformuluojame hipotezes:

H_0: koreliacijos koeficientas lygus nuliui,

H_1: koreliacijos koeficientas nelygus nuliui.

Stebimasis reikšmingumo lygmuo *p-level* = *0,000*, tai mažiau už pasirinktą reikšmingumo lygmenį $\alpha = 0,05$. Patvirtinta hipotezė H_1.

Atliktos procedūros sintaksė:

```
CORRELATIONS
  /VARIABLES=Pakuotės_dydis Kiekis
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE .
```

Apskaičiuoti Pirsono koreliacijos koeficientą galima ir kitaip. Patikrinsime, ar yra ryšys tarp skirtingos kavos rūšies įvertinimo. Analizės dialogo langas atveriamas komanda *Analyze* → *Descriptive Statistics* → *Crosstabs*.... Kintamasis *Pakuotės_dydis* perkeliamas į langelį *Row*(*s*), o kintamasis *Kiekis* – į langelį *Column*(*s*). Mygtuko *Statistics* paspaudimu atveriamas dialogo langas *Crosstabs: Statistics*. Šiame lange reikia pa-

žymėti langelį *Correlations*. Paspaudę mygtuką *Continue* grįžtame į pradinį langą. Procedūra baigiama mygtuko *OK* paspaudimu. Lange *Output-SPSS Viewer* rodomi analizės rezultatai (žr. 5.23 pav.).

Symmetric Measures

		Value	Asymp. Std. Error(a)	Approx. T(b)	Approx. Sig.
Interval by Interval	Pearson's R	,760	,058	8,899	,000(c)
Ordinal by Ordinal	Spearman Correlation	,894	,026	15,186	,000(c)
N of Valid Cases		60			

a Not assuming the null hypothesis.
b Using the asymptotic standard error assuming the null hypothesis.
c Based on normal approximation.

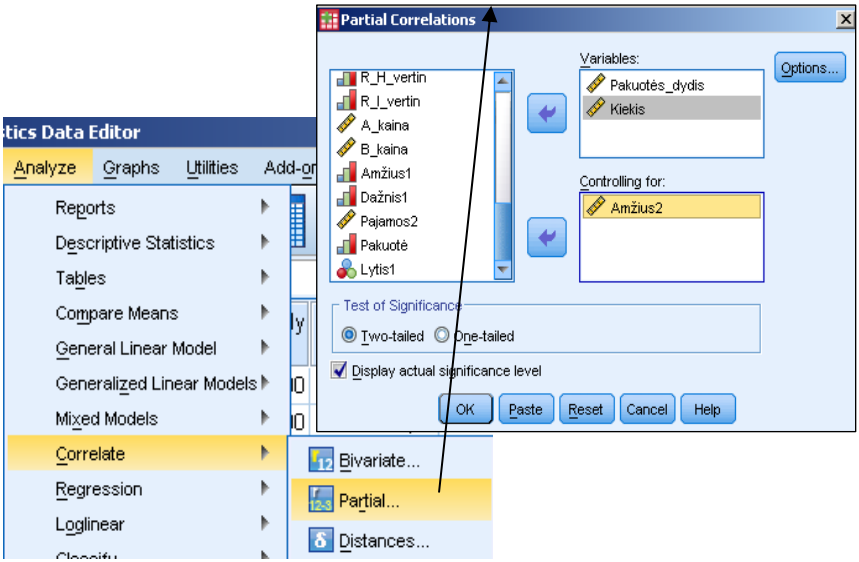
5.23 pav. **Pirsono koreliacijos koeficiento skaičiavimo rezultatas**

Šiuo atveju lentelėje *Symmetric Measures* pateikti du koreliacijos koeficientai: Pirsono ir Spirmeno. Tyrėjas turi žinoti, koks koeficientas tinkamas. Nagrinėjamas kiekybinių kintamųjų ryšys, Pirsono koreliacijos koeficientas (*Pearson's R*) $r = 0,760$, koreliacija stipri ir teigiama. Stebimasis reikšmingumo lygmuo (*Approx. Sig.*) *p-level* $= 0,000$, tai mažiau už pasirinktą reikšmingumo lygmenį $\alpha = 0,05$. Patvirtinta hipotezė H_1. Galima pasakyti, kad rezultatai nesiskiria, atsižvelgiant į tai, koks skaičiavimo būdas taikomas. Lentelėje *Symmetric Measures* pateikta ir *t* kriterijaus reikšmė (*Approx. T*(*b*)). Norint koreliacijos koeficiento reikšmingumą galima patikrinti ir *T* statistika.

Atliktos procedūros sintaksė:

```
/COMPRESSED.
CROSSTABS
  /TABLES=Pakuotės_dydis BY Kiekis
  /FORMAT= AVALUE TABLES
  /STATISTIC=CORR
  /CELLS= COUNT
  /COUNT ROUND CELL .
```

Patikrinsime, koks yra pakuotės dydžio ir perkamos kavos kiekio ryšys, pašalinus amžiaus įtaką (apskaičiuosime dalinį koreliacijos koeficientą). Analizės dialogo langas *Partial Correlations* atveriamas komanda *Analyze → Correlate → Partial...* (žr. 5.24 pav.). Kintamieji *Pakuotės_dydis* ir *Kiekis* perkeliami į langelį *Variables*, kintamasis *Amžius2* – į langelį *Controlling for*. Procedūra baigiama mygtuko *OK* paspaudimu. Lange *Output-SPSS Viewer* rodomi analizės rezultatai (žr. 5.25 pav.).



5.24 pav. **Dialogo langas *Partial Correlations***

Galima pasakyti, kad pašalinus amžiaus įtaką, perkamos pakuotės dydžio ir kavos kiekio ryšys tampa kiek silpnesnis (Pirsono koreliacijos koeficientas $r = 0{,}760$, o dalinis koreliacijos koeficientas $r_{XY} = 0{,}752$).

Atliktos procedūros sintaksė:

```
DATASET NAME DataSet1 WINDOW=FRONT.

PARTIAL CORR
  /VARIABLES= Pakuotės_dydis Kiekis BY Amžius2
  /SIGNIFICANCE=TWOTAIL
  /MISSING=LISTWISE .
```

Correlations

Control Variables			Pakuotės_dydis	Kiekis
Amžius2	Pakuotės_dydis	Correlation	1,000	,752
		Significance (2-tailed)	.	,000
		df	0	57
	Kiekis	Correlation	,752	1,000
		Significance (2-tailed)	,000	.
		df	57	0

5.25 pav. **Dalinio koreliacijos koeficiento skaičiavimo rezultatas**

KONTROLINIAI KLAUSIMAI

5.1. Kokiais atvejais kintamųjų ryšiui nustatyti taikomas Pirsono χ^2 suderinamumo kriterijus?
5.2. Ar galima Pirsono χ^2 suderinamumo kriterijų taikyti kiekybiniams kintamiesiems?
5.3. Paaiškinkite *tikėtinų dažnių* sampratą.
5.4. Kaip nustatomas statistinio ryšio stiprumas, taikant Pirsono χ^2 suderinamumo kriterijų?
5.5. Kokie Pirsono χ^2 suderinamumo kriterijaus taikymo apribojimai?
5.6. Kokiais atvejais taikomas Spirmeno koreliacijos koeficientas?
5.7. Kokiais atvejais taikomas Pirsono koreliacijos koeficientas?
5.8. Kaip priimamas sprendimas, ar atmesti hipotezę H_0: koreliacijos koeficientas lygus nuliui?

5.9. Kaip koreliaciniai ryšiai skirstomi pagal formą?
5.10. Kaip koreliaciniai ryšiai skirstomi pagal kryptį?
5.11. Ką rodo determinacijos koeficientas?

Lytis * Daznis1 Crosstabulation

Count

		Daznis1			
		1,00	2,00	3,00	Total
Lytis	Vyras	13	7	10	30
	Moteris	7	14	9	30
Total		20	21	19	60

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4,186[a]	2	,123
Likelihood Ratio	4,259	2	,119
Linear-by-Linear Association	,631	1	,427
N of Valid Cases	60		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 9,50.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,264	,123
	Cramer's V	,264	,123
N of Valid Cases		60	

a. Not assuming the null hypothesis.

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

5.26 pav. **Pirsono χ^2 suderinamumo kriterijaus skaičiavimo rezultatas**

UŽDUOTYS

5.1 užduotis

Tikrinant pirkimo dažnio priklausomybę nuo lyties, gautas rezultatas (žr. 5.26 pav.). Suformuluokite hipotezes ir paaiškinkite gautą rezultatą.

5.2 užduotis

Tikrinant priimtinos jogurto kainos (litais) priklausomybę nuo amžiaus (metais), gautas rezultatas (žr. 5.27 pav.). Suformuluokite hipotezes ir paaiškinkite gautą rezultatą.

Amzius2 * A_kaina Crosstabulation

Count

		A_kaina					Total
		1,00	2,00	3,00	4,00	5,00	
Amzius2	20,00	80	48	16	40	16	200
	30,00	8	16	24	32	32	112
	40,00	8	16	56	72	16	168
Total		96	80	96	144	64	480

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	151,787[a]	8	,000
Likelihood Ratio	152,301	8	,000
Linear-by-Linear Association	65,558	1	,000
N of Valid Cases	480		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 14,93.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval by Interval	Pearson's R	,370	,039	8,706	,000[c]
Ordinal by Ordinal	Spearman Correlation	,356	,041	8,333	,000[c]
N of Valid Cases		480			

a. Not assuming the null hypothesis.

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

c. Based on normal approximation.

5.27 pav. **Pirsono χ^2 suderinamumo kriterijaus skaičiavimo rezultatas**

5.3 užduotis

5.28 pav. pateikta lyties – pajamų porinių dažnių lentelė. Paaiškinkite, ar nagrinėjamu atveju kintamųjų ryšiui nustatyti galima taikyti Pirsono χ^2 suderinamumo kriterijų? Kodėl?

Lytis * Pajamos Crosstabulation

Count

		Pajamos					
		iki 8000 Lt	801-1000 Lt	1001-1500 Lt	1501-2000 Lt	daugiau 2000 Lt	Total
Lytis	Vyras	0	2	5	13	10	30
	Moteris	5	10	2	7	6	30
Total		5	12	7	20	16	60

5.28 pav. **Porinių dažnių lentelė**

5.4 užduotis

Tyrimo metu gauti kintamieji: respondentų amžius (kiekybinis), respondentų vidutinės pajamos (ranginis), vidutinis pinigų kiekis, išleidžiamas per mėnesį prekei A (kiekybinis), priimtina prekės A kaina (kiekybinis). Apdorojant rezultatus nustatyta, kad amžiaus ir vidutinio pinigų kiekio, išleidžiamo per mėnesį prekei A, duomenų skirstinys yra normalusis, o priimtinos prekės A kainos duomenų skirstinys neatitinka normalumo reikalavimų. Numatykite, kokius koreliacijos koeficientus galima apskaičiuoti duomenų ryšiams nustatyti.

5.5 užduotis

Analizuojant pajamų ir pirkimo dažnio ryšį, buvo apskaičiuotas Spirmeno koreliacijos koeficientas (žr. 5.29 pav.). Suformuluokite statistines hipotezes ir paaiškinkite gautą rezultatą.

Correlations

			Pajamos	Dažnis
Spearman's rho	Pajamos	Correlation Coefficient	1,000	-,050
		Sig. (2-tailed)	.	,704
		N	60	60
	Dažnis	Correlation Coefficient	-,050	1,000
		Sig. (2-tailed)	,704	.
		N	60	60

5.29 pav. **Koreliacinės analizės rezultatas**

5.6 užduotis

Analizuojant amžiaus ir pinigų, išleistų pirkiniams, ryšį, buvo apskaičiuotas Pirsono koreliacijos koeficientas (žr. 5.30 pav.). Suformuluokite statistines hipotezes ir paaiškinkite gautą rezultatą.

Correlations

		Amžius	Suma
Amžius	Pearson Correlation	1	,604
	Sig. (2-tailed)		,003
	N	30	30
Suma	Pearson Correlation	,604	1
	Sig. (2-tailed)	,003	
	N	30	30

5.30 pav. **Koreliacinės analizės rezultatas**

6. REGRESINĖ ANALIZĖ

Apskaičiavus koreliacijos koeficientą nustatomas ryšių egzistavimas, tačiau negalima nustatyti ryšio dėsnio. Kintamųjų ryšiui apibūdinti ir prognozuoti taikomi regresijos modeliai. Regresinės analizės prognozės yra kiekybinės – regresijos funkcija aprašoma tiriamojo (priklausomojo – *pasekmės*) kintamojo priklausomybė nuo kito (nepriklausomojo – *priežasties*) kintamojo reikšmių kitimo. Regresinė analizė leidžia parinkti kintamuosius siejančią funkciją (sudaromas regresijos modelis). Modeliui sudaryti taikomas aproksimacijos arba *kreivės pritaikymo* metodas, kai turimiems duomenims parenkama geriausiai tinkama funkcijos kreivė. Preliminariai nustatyti priklausomybės formą (funkcijos tipą) galima nagrinėjant analizuojamų duomenų grafiką.

6.1. Tiesinis (paprastasis) regresijos modelis

Tiesinės regresijos modelis nustato kiekybinių kintamųjų tiesinę priklausomybę. Šis modelis dažnai taikomas socialiniuose moksluose. Tiesinė kintamojo Y nuo kintamojo X priklausomybė užrašoma lygtimi:

$$y = \alpha + \beta x + \varepsilon, \qquad (6.1)$$

čia: α ir β yra nežinomi koeficientai (konstantos),
ε – atsitiktinė paklaida.

Atsitiktinę paklaidą gali sudaryti matavimo paklaidos arba neįvardytos ir netiriamos priežastys.

Parinkę įvairias koeficientų α ir β reikšmes, gausime tiesių šeimą. Geriausiai tiesę atitinkantys koeficientai apskaičiuojami taikant matematinį *mažiausiųjų kvadratų metodą*. Šiuo metodu gautos koeficientų α ir

β reikšmės – įverčiai (jas žymėsime raidėmis a ir b) randamos minimizuojant reiškinį:

$$SSE = \sum_{i=1}^{n} \left(y_i - \alpha - \beta x_i \right)^2 . \tag{6.2}$$

Sumą SSE minimizuoja šie koeficientų α ir β įverčiai:

$$b = \frac{n\left(\sum_{i=1}^{n} x_i y_i \right) - \left(\sum_{i=1}^{n} x_i \right)\left(\sum_{i=1}^{n} y_i \right)}{n\left(\sum_{i=1}^{n} x_i^2 \right) - \left(\sum_{i=1}^{n} x_i \right)^2}, \tag{6.3}$$

$$a = \frac{\sum_{i=1}^{n} y_i}{n} - b\frac{\sum_{i=1}^{n} x_i}{n}. \tag{6.4}$$

Lygtis $\hat{y}(x) = a + bx,$ kurios koeficientai apibrėžti (6.3) ir (6.4) lygybėmis, yra regresijos tiesės lygtis, o $\varepsilon = y(x) - \hat{y}(x)$ – liekamoji paklaida, arba liekana. Koeficientai a ir b, apskaičiuoti iš imties, skirsis nuo koeficientų α ir β, gaunamų iš populiacijos. Tiesinės regresijos atveju koeficientų α ir β skirstiniai yra normalieji su vidurkiais a ir b.

Regresijos modelio reikšmingumas. Dažniausiai tikrinama hipotezė, kad tarp kintamųjų X ir Y nėra jokio tiesinio ryšio, t. y. kad koeficientas β lygus nuliui.

H_0: $\beta_1 = 0$,
H_1: $\beta_1 \neq 0$.

Šios hipotezės siejamos su regresijos lygties reikšmingumu. Jeigu hipotezė H_0 neatmetama, vadinasi, tarp kintamųjų X ir Y tiesinio ryšio nėra. Norint patikrinti šią hipotezę, galima taikyti t statistiką:

$$t = \frac{\beta}{s_\beta}, \qquad (6.5)$$

čia s_β – β koeficiento standartinis nuokrypis.

Jei nulinė hipotezė, kad nėra jokio tiesinio ryšio, yra teisinga, ši statistika (6.5) turi Stjudento *t* skirstinį su (*n – 2*) laisvės laipsniais. Hipotezę, kad tiesė eina per koordinačių pradžią, t. y. kad koeficientas α lygus nuliui, galima patikrinti naudojantis kita *t* statistika:

$$t = \frac{\alpha}{s_\alpha}, \qquad (6.6)$$

čia s_α – α koeficiento standartinis nuokrypis.

Statistika (6.6) taip pat turi Stjudento *t* skirstinį su (*n – 2*) laisvės laipsniais (n – imties dydis).

Parinkto modelio tikimas turimiems duomenims

Sudarius modelį, būtina ištirti, ar šis modelis gerai tinka turimiems duomenims. Tiesinio modelio tinkamumui tikrinti taikomi: koreliacijos koeficientas, determinacijos koeficientas ir pataisytasis determinacijos koeficientas.

Determinacijos koeficientu R^2 įvertinamas modelio tikimas visai populiacijai dažnai būna netikslus. Pataisytasis determinacijos koeficientas R_a^2 tiksliau įvertina, ar modelis tinka duomenims:

$$R_a^2 = R^2 - \frac{p(1-R^2)}{n-p-1}, \qquad (6.7)$$

čia *p* yra nepriklausomųjų kintamųjų skaičius (nagrinėjamu atveju $p = 1$).

Didelė determinacijos koeficiento reikšmė reiškia, kad duomenys yra mažai nutolę nuo mažiausiųjų kvadratų metodu gautos tiesės. Jei duomenys idealiai atitinka regresijos tiesės lygtį, t. y. visos y_i reikšmės patenka ant regresijos tiesės, determinacijos koeficientas yra 1. Jei de-

terminacijos koeficientas artimas nuliui, tai regresijos tiesės lygtis visiškai netinka prognozei. Praktiškai taikant tiesinę regresinę analizę dažniausiai reikalaujama, kad $R^2 \geq 0{,}25$. Jei $R^2 < 0{,}25$, tiesinės regresijos modelio taikymas labai abejotinas.

Statistinės programos koeficiento β (kartu ir regresijos lygties) reikšmingumui tikrinti taikomi dispersinės analizės metodai (žr. skyrių „Vienfaktorinė dispersinė analizė“). Šiuo metodu taip pat tikrinama hipotezė apie koeficiento β lygybę nuliui. Jeigu nulinė hipotezė H_0: $\beta = 0$ yra teisinga, tai F statistika yra pasiskirsčiusi pagal $F_{1,n-2}$ skirstinį ir nulinė hipotezė atmetama, kai apskaičiuota reikšmė $F_0 > F_{\alpha,1,n-2}$. Dispersinės analizės metodu gaunamas analogiškas rezultatas kaip ir taikant t kriterijų. Abi procedūros pateikia vienodas išvadas.

6.2. Netiesinė (kreivinė) regresija

Daugelyje praktinių rinkos tyrimo uždavinių tiesinis regresijos modelis, jei ir ne visai tiksliai, tai gana gerai aprašo tikrąją statistinio ryšio formą. Tačiau kartais remiantis sklaidos diagrama arba žinomomis teorinėmis prielaidomis tenka pripažinti, kad ryšio forma yra aiškiai netiesinė. Taikant atitinkamas transformacijas kartais netiesinį regresijos modelį galima pavaizduoti tiesiniu. Pavyzdžiui, eksponentinį modelį $y = ae^{bx}$ į tiesinį galima transformuoti skaičiuojant logaritmą $ln(y) = ln(a) + bx$. Modeliai, kurie transformacijos būdu gali būti pertvarkyti į tiesinius, vadinami tiesiniais iš esmės (*Intrinsically Linear Model*).

Regresijos koeficientų skaičiavimas, parinkto modelio tinkamumo turimiems duomenims tikrinimas užima daug laiko, todėl šiam darbui dažniausiai taikomos statistinės programos.

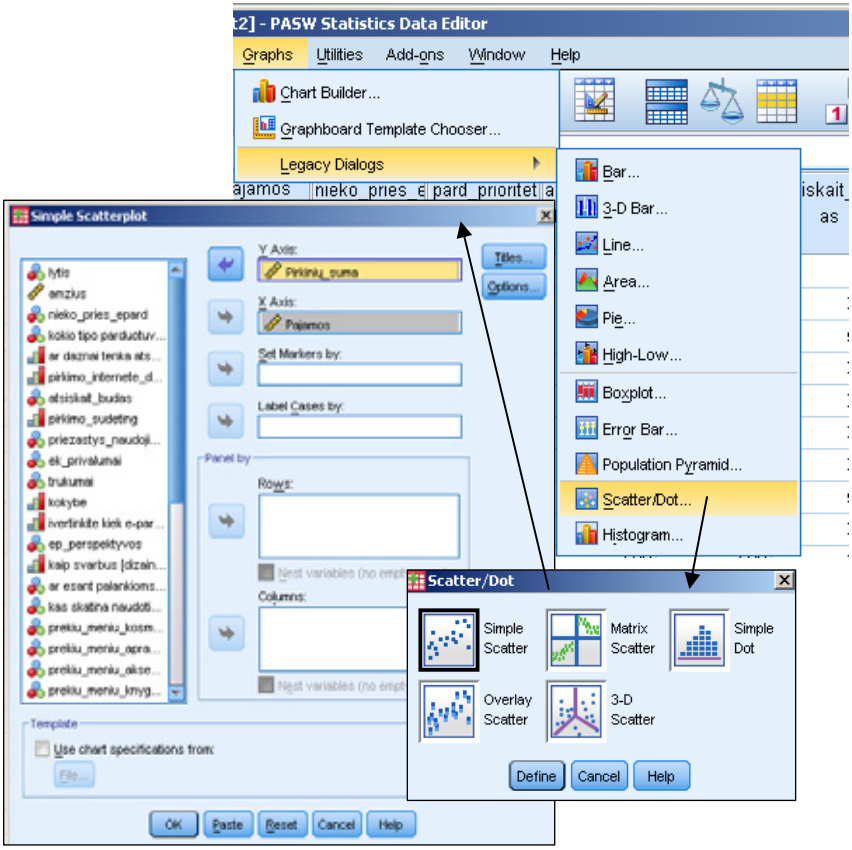
6.2.1. Regresinė analizė SPSS terpėje

Apskaičiuosime pinigų, išleidžiamų pirkiniams, priklausomybę nuo pajamų. Norėdami numatyti regresijos modelio formą nubraižysime taškinę priklausomybės diagramą. Braižydami grafiką įvykdysime koman-

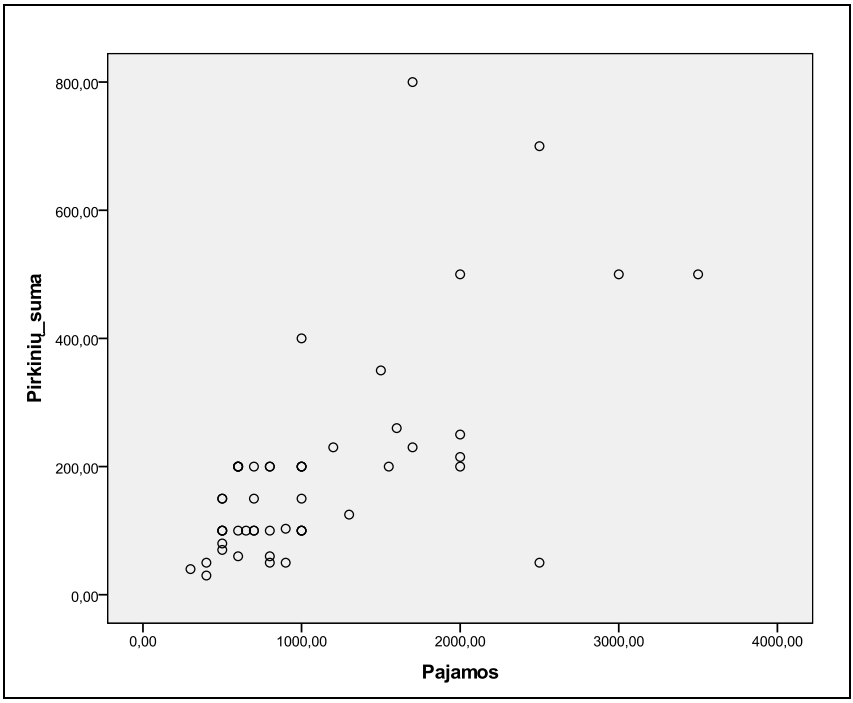
dą *Graphs → Legasy Dialogs → Scatter/Dot...*. Papildomame dialogo lange *Scatter/Dot…* pasirenkame langelį *Simple Scatter*. Atsiveria grafiko braižymo dialogo langas *Simple Scatterplot* (žr. 6.1 pav). Priklausomąjį kintamąjį *Pirkinių_suma* perkeliame į y ašį (Y Axis), o nepriklausomąjį kintamąjį *Pajamos* – į x ašį (X Axis). Procedūra baigiama mygtuko *OK* paspaudimu. Lange *Output-SPSS Viewer* rodomas rezultatas (žr. 6.2 pav.).

Atliktos procedūros sintaksė:

```
GRAPH
  /SCATTERPLOT(BIVAR)=Pajamos WITH Pirkinių_suma
  /MISSING=LISTWISE.
```

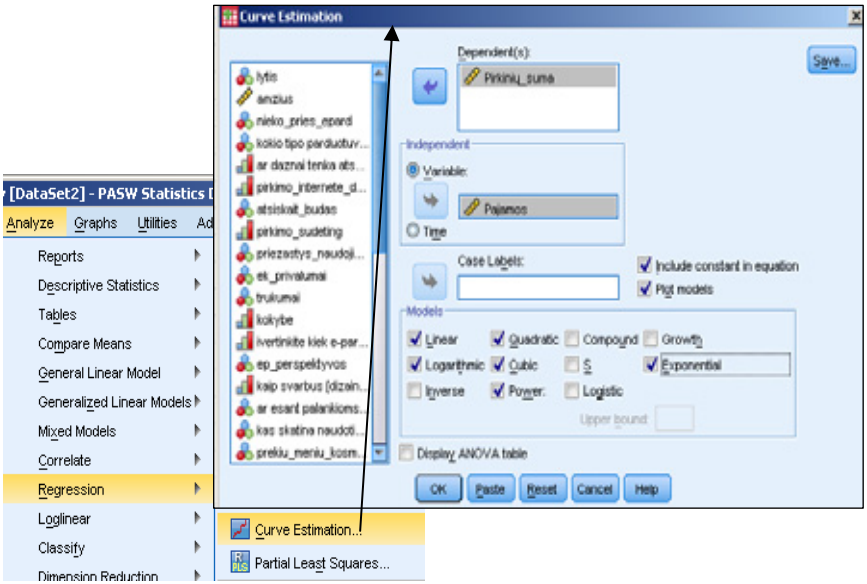


6.1 pav. **Dialogo langas *Simple Scatterplot***



6.2 pav. **Kintamųjų ryšio taškinė diagrama**

Regresijos modelio parametrų vertinimas gaunamas taikant *Regression* procedūrą. Ši procedūra leidžia pritaikyti ne tik tiesinės regresijos modelį, bet ir kitos formos modelį, jei sunku numatyti modelio tinkamumą (pavyzdžiui, kaip nagrinėjamu atveju). Analizės dialogo langas atveriamas komanda *Analyze → Regression → Curve Estimation...* (žr. 6.3 pav.). Priklausomąjį kintamąjį *Pirkinių_suma* perkeliame į langelį *Dependent*(*s*), o nepriklausomąjį kintamąjį *Pajamos* – į langelį *Independent.* Srityje *Models* galima pasirinkti norimą regresijos modelį. Nežinant, koks regresijos modelis tinkamesnis, galima pažymėti keletą modelių. Norint nubraižyti regresijos grafiką, reikia pažymėti langelį *Plot models.* Procedūra baigiama mygtuko *OK* paspaudimu. Lange *Output-SPSS Viewer* rodomi analizės rezultatai (žr. 6.4 pav.).



6.3 pav. **Dialogo langas *Curve Estimation***

Modelio tinkamumas nustatomas pagal determinacijos koeficiento R^2 (*R Square*) reikšmę. Modelis laikomas tinkamu, jei determinacijos koeficiento reikšmė $R^2 > 0{,}25$. Analizės rezultatai ir regresijos lygties koeficientai rodomi lentelėje *Model Summary and Parameter Estimates.* Nagrinėjamu atveju tinkami visi parinkti regresijos modeliai, nes determinacijos koeficientai $R^2 > 0{,}25$. Tokiu atveju, nors kvadratinis (*Quadratic*, $R^2 = 0{,}437$) ir kubinis (*Cubic*, $R^2 = 0{,}438$) modeliai tinkami geriau negu tiesinis (*Linear*, $R^2 = 0{,}433$), patariama pasirinkti tiesinį modelį (jis lengviau analizuojamas ir daugiau tinkamas prognozei). Determinacijos koeficientas rodo, kokią dalį tikro proceso aprašo modelis. Nagrinėjamu atveju – tai tik apie 43 procentus. Apskaičiuotų regresijos lygties koeficientų (konstantų) patikimumą galima patikrinti taikant stebimojo reikšmingumo lygmens *p-level* (*Sig.*) ir pasirinkto reikšmingumo lygmens α palyginimą. Tikrinamos hipotezės H_0: $a = 0$ ir H_0: $b = 0$. Hipotezė atmetama, jei *p-level* $< \alpha$. Nagrinėjamu atveju dėl visų modelių *p-level* $= 0{,}000 < \alpha = 0{,}05$. Koeficientai reikšmingai skiriasi nuo nulio.

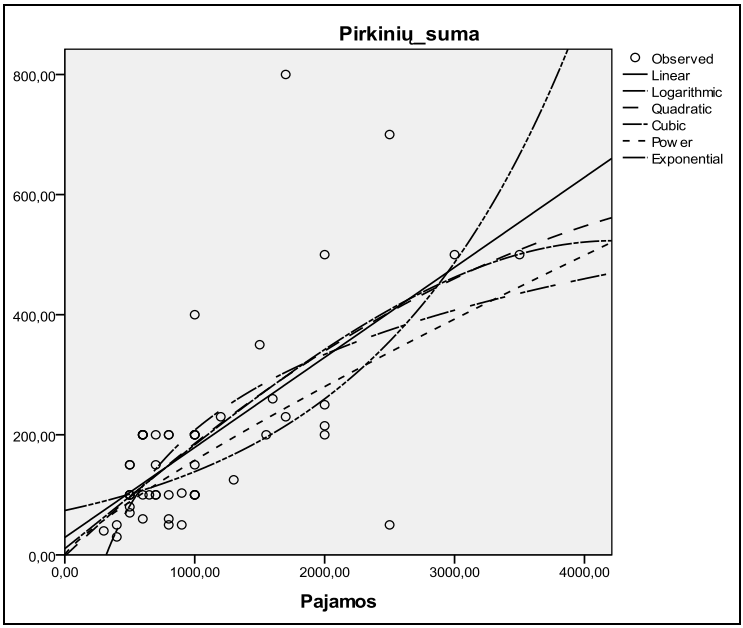
Model Summary and Parameter Estimates

Dependent Variable: Pirkinių_suma

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	,433	37,439	1	49	,000	29,044	,150		
Logarithmic	,415	34,705	1	49	,000	-1048,339	181,840		
Quadratic	,437	18,656	2	48	,000	-1,953	,204	-1,67E-005	
Cubic	,438	12,187	3	47	,000	10,461	,172	4,64E-006	-3,95E-009
Power	,417	35,035	1	49	,000	,501	,832		
Exponential	,365	28,131	1	49	,000	73,994	,001		

The independent variable is Pajamos.

6.4 pav. **Regresinės analizės rezultatas**



6.5 pav. **Regresinės analizės rezultato grafikai**

Galima užrašyti regresijos lygtis:
Tiesinės regresijos modelis: $y = a + b1 \bullet x$.
Logaritminis modelis: $y = a \bullet log\ (b1 \bullet x)$.
Kvadratinis modelis: $y = a + b1 \bullet x + b2 \bullet x^{2.}$
Kubinis modelis: $y = x + b1 \bullet x + b2 \bullet x^2 + b3 \bullet x^{3.}$
Laipsninis modelis: $y = a \bullet x^{b1.}$
Eksponentinis modelis: $y = a \bullet exp\ (b1 \bullet x)$.
Čia: *y – Pirkinių_suma*,
x – Pajamos,
a – Constant (žr. 6.4 pav.),
b1, *b2*, *b3* – koeficientai iš lentelės (žr. 6.4 pav.).

Pavyzdžiui, tiesinės regresijos lygtis: *Pirkinių_suma = 29,044 + 0,15•Pajamos.* Regresijos modelių grafikai pateikti 6.5 pav.

Atliktos procedūros sintaksė:

```
* Curve Estimation.
TSET NEWVAR=NONE.
CURVEFIT
  /VARIABLES=Pirkinių_suma WITH Pajamos
  /CONSTANT
  /MODEL=LINEAR LOGARITHMIC QUADRATIC CUBIC
POWER EXPONENTIAL
  /PLOT FIT.
```

Dar tikslesnį rezultatų įvertinimą galime gauti, jei pritaikytume *ANOVA* procedūrą. Tam dialogo lange *Curve Estimation* reikia pažymėti langelį *Display ANOVA table*. Pasirinksime tiesinį (*Linear*) modelį. Lange *Output-SPSS Viewer* rodomi analizės rezultatai (žr. 6.6 pav.).

Lentelėje *Model Summary* pateiktas ne tik determinacijos koeficientas (*R Square*), bet ir patikslintas determinacijos koeficientas (*Adjusted R Square*), kuris yra pastebimai mažesnis. Lentelėje *Coefficients* pateikti tiesinės regresijos lygties koeficientai, analogiški pateiktiems 6.4 pav. Tačiau koeficientų reikšmingumas įvertintas atskirai. Tikrinamos hipotezės H_0*: a = 0* ir H_0*: b = 0*. Hipotezė atmetama, jei *p-level* (stulpelis *Sig.*) < α. Nagrinėjamu atveju hipotezės H_0*: a = 0* (*p-level = 0,364* > α *= 0,05*) atmesti negalima, todėl atmetama hipotezė H_0*: b = 0* (*p-level = 0,000* < α *= 0,05*). Galima teigti, kad *b* koeficientas yra patikimas, o *a* konstanta nepatikima. Stulpelyje *Beta* pateikiamas standartizuotų duo-

menų regresijos lygties koeficientas. Paprastosios tiesinės regresijos atveju šis koeficientas sutampa su pradinių duomenų koreliacijos koeficientu. Paskutiniuose dviejuose lentelės *Coefficients* stulpeliuose pateikti koeficientų a ir b 95 procentų pasikliautinųjų intervalų rėžiai. Lentelėje *ANOVA* pateikti dispersinės analizės rezultatai. Stulpelyje kvadratų suma (*Sum of Squares*): regresijos kvadratų suma (*Regression*), liekamųjų paklaidų kvadratų suma (*Residual*) ir visa kvadratų suma (*Total*).

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
,658	,433	,422	122,521

The independent variable is Pajamos.

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	562015,233	1	562015,233	37,439	,000
Residual	735553,395	49	15011,294		
Total	1297568,627	50			

The independent variable is Pajamos.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
Pajamos	,150	,025	,658	6,119	,000
(Constant)	29,044	31,709		,916	,364

6.6 pav. **Regresinės analizės rezultatas (tiesinė regresija)**

Atliktos procedūros sintaksė:

```
* Curve Estimation.
TSET NEWVAR=NONE .
CURVEFIT /VARIABLES=Pirkinių_suma WITH Pajamos
  /CONSTANT
  /MODEL=LINEAR
  /PRINT ANOVA
  /PLOT FIT.
```

6.3. Daugialypės tiesinės regresijos modelis

Paprasta tiesinė regresija ir porinis koreliacijos koeficientas tinka tada, kai tiriami du kintamieji – vienas priklausomasis ir vienas nepriklausomasis. Dažnai nepriklausomųjų kintamųjų būna keletas. Tada taikomas daugialypės tiesinės regresijos modelis ir daugialypė koreliacinė analizė. Daugialypės regresijos modelis yra tiesinio (paprastojo) regresijos modelio apibendrinimas atvejui, kai nepriklausomųjų kintamųjų yra daugiau nei vienas:

$$y = \beta + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k + \varepsilon . \qquad (6.8)$$

Kaip ir esant paprastai tiesinei regresijai, daugialypės tiesinės regresijos tyrimo tikslas – apskaičiuoti parametrų β_1, β_2, β_3,... reikšmes.

Apie regresijos lygties reikšmingumą sprendžiama analogiškai dviejų kintamųjų tiesinės regresijos atveju. Koeficientų reikšmingumas tikrinamas *t* kriterijumi, o nepriklausomųjų kintamųjų X įtaka priklausomajam kintamajam *Y* vertinama apskaičiuojant determinacijos koeficientą arba taikant dispersinės analizės metodus. Šiuo atveju dispersine analize tikrinamos hipotezės:

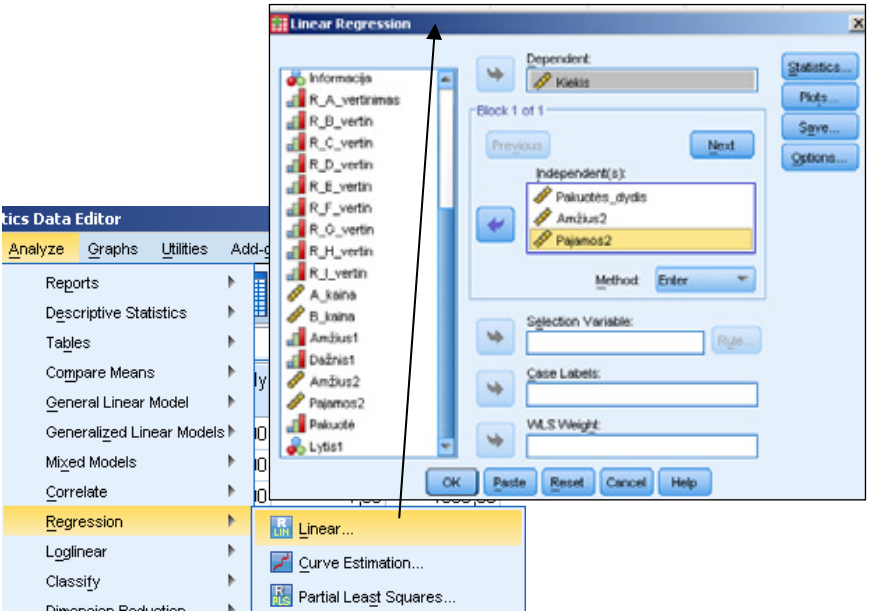
H_0: $\beta_1 = \beta_2 = ... = \beta_k = 0$,
H_1: $\beta_j \neq 0$ bent vienam *j*.

Jeigu taikant *F* kriterijų nulinė hipotezė H_0 neatmetama, tai regresijos modelis netinka. Tuo tarpu atmesta nulinė hipotezė dar nereiškia, kad modelis tikrai tinka. Tai tik vienas iš daugelio rodiklių, kuriuo remiantis sprendžiama apie modelio tinkamumą.

6.3.1. Daugialypė regresija SPSS terpėje

Galima analizuoti vieno priklausomojo kintamojo ryšį su keletu nepriklausomųjų kintamųjų. Pavyzdžiui, nustatysime perkamos kavos kiekio priklausomybę nuo amžiaus, pajamų ir pakuotės dydžio. Dialogo

langas *Linear Regression* atveriamas komanda *Analyze* → *Regression* → *Linear*... (žr. 6.7 pav.). Priklausomąjį kintamąjį *Kiekis* perkeliame į langelį *Dependent*, o nepriklausomuosius kintamuosius *Amžius*, *Pajamos*, *Pakuotės_dydis* – į langelį *Independent*(*s*). Procedūra baigiama mygtuko *OK* paspaudimu. Lange *Output-SPSS Viewer* rodomi analizės rezultatai (žr. 6.8 pav.).



6.7 pav. **Dialogo langas *Linear Regression***

Lentelėje *Model Summary* pateikta determinacijos koeficiento reikšmė (*R Square = 0,585*) ir patikslinta determinacijos koeficiento reikšmė (*Adjusted R Square = 0,563*). $R^2 = 0{,}585 > 0{,}25$, modelis tinkamas. Taip pat pateiktas daugialypės koreliacijos koeficientas $R = 0{,}765$. Lentelėje *ANOVA*(*b*) pateiktas kintamųjų ryšio modelio vertinimas. Gauname vieną iš dviejų išvadų: 1) visi koeficientai $b_j = 0$, t. y. regresijos modelis su duomenimis nesuderintas; 2) bent vienas koeficientas $b \neq 0$, t. y. regresijos modelis bent jau iš dalies prognozėms tinka. Hipotezė H_0: visi koeficientai lygūs nuliui atmetama, kai *p-level*

(stulpelis *Sig.*) < α, čia α – pasirinktas reikšmingumo lygmuo. Nagrinėjamu atveju *p-level* = 0,000 < α = *0,05*, hipotezė H_0 atmetama. Lentelėje *Coefficients*(*a*) pateikiami daugialypės regresijos koeficientai. Regresijos lygtį galima užrašyti taip:

Kiekis = 97,514 + 1,282•Pakuotės_dydis – 1,908•Amžius2 + 0,088•Pajamos2.

Lentelėje *Coefficients*(*a*) pateikti ir standartizuotieji koeficientai (*Standardized Coefficients*, *Beta*). Didesnė absoliuti *Beta* koeficiento reikšmė rodo didesnę Y priklausomybę nuo X_j.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,765(a)	,585	,563	263,96190

a Predictors: (Constant), Pajamos2, Amþius2, Pakuotës_dydis

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5502435,413	3	1834145,138	26,324	,000(a)
	Residual	3901849,587	56	69675,885		
	Total	9404285,000	59			

a Predictors: (Constant), Pajamos2, Amþius2, Pakuotës_dydis
b Dependent Variable: Kiekis

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	97,514	140,251		,695	,490
	Pakuotės_dydis	1,282	,165	,730	7,753	,000
	Amžius2	-1,908	4,134	-,043	-,462	,646
	Pajamos2	,088	,086	,101	1,022	,311

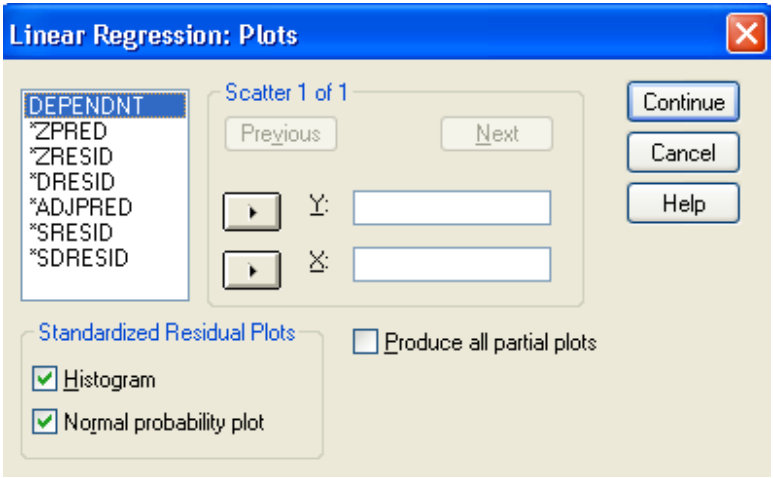
a Dependent Variable: Kiekis

6.8 pav. **Daugialypės regresinės analizės rezultatas**

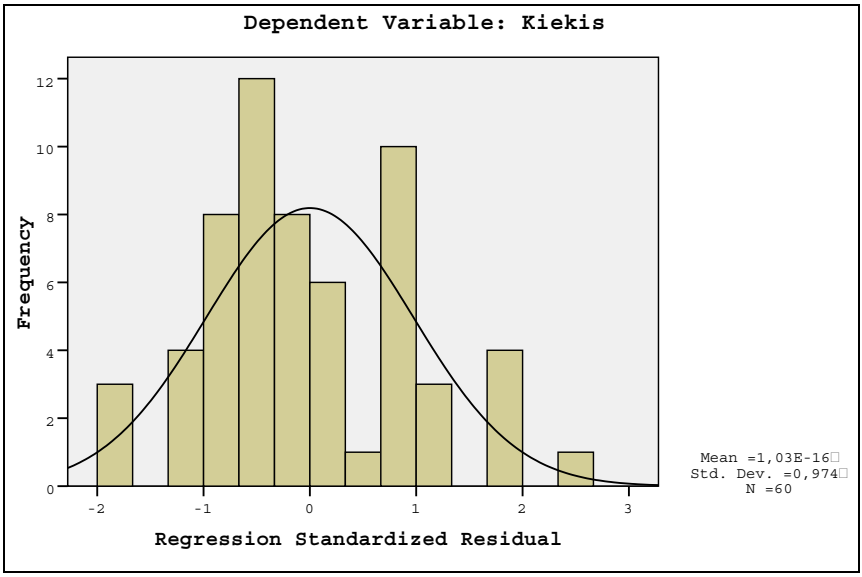
Atliktos procedūros sintaksė:

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Kiekis
  /METHOD=ENTER Pakuotės_dydis Amžius2 Pajamos2 .
```

Svarbu išanalizuoti liekanas, t. y. stebimųjų reikšmių skirtumą nuo statistiškai tikėtinų. Liekanos turi būti atsitiktinės ir turėti normalųjį skirstinį. Tai galima patikrinti standartizuotųjų liekanų histograma. Tuo tikslu mygtuko *Plots* paspaudimu galima iškviesti dialogo langą *Linear Regression: Plots* (žr. 6.9 pav.) ir jame pažymėti langelius *Histogram* (histograma) bei *Normal probability plot* (normaliojo skirstinio kreivė). Gauta histograma pateikta 6.10 pav.



6.9 pav. **Dialogo langas *Linear Regression: Plots***



6.10 pav. **Standartizuotųjų liekanų histograma**

Mygtuko *Statistics* paspaudimu iškviečiamas dialogo langas *Linear Regression: Statistics* (žr. 6.11 pav.). Šiame lange pažymėti nustatytieji langeliai *Regression Coefficients* (regresijos koeficientai): *Estimates* (įverčiai) – pateikiamos regresijos koeficientų reikšmės, *Model fit* (modelio tinkamumas) – pateikiamos determinacijos koeficiento, pataisyto determinacijos koeficiento, standartinės paklaidos reikšmės, taip pat dispersinės analizės (*ANOVA*) rezultatai. Pažymėjus *Durbin-Watson* langelį skaičiuojamas Durbino–Vatsono kriterijus, kuris taikomas prielaidai apie liekamųjų paklaidų nepriklausomumą (taip pat, kad visi Y_i yra nepriklausomi, t. y. nėra autokoreliacijos) patikrinti.

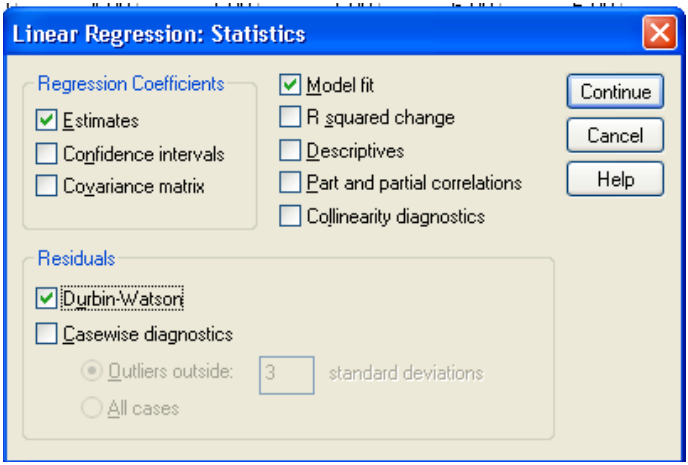
Autokoreliacijos priežastys: nagrinėjamo reiškinio inertiškumas, netiksliai parinkti nagrinėjamą reiškinį veikiantys veiksniai, neteisingai parinkta veiksnių priklausomybės matematinė išraiška. Jei autokoreliacija nustatyta, tai mažiausiuoju kvadratų metodu apskaičiuotas determinacijos koeficientas R^2 yra didesnis už tikrąjį, negalima tikrinti hipotezių *t* kriterijumi.

Durbino–Vatsono statistika *d* kinta nuo 0 iki 4. Kuo *d* arčiau 2, tuo mažiau tikėtina, kad tarp liekamųjų paklaidų (ir tarp kintamojo *Y* reikš-

mių) yra autokoreliacija. *Durbin-Watson* kriterijus pateiktas 6.12 pav. Nagrinėjamu atveju rezultatas beveik idealus.

Atliktos procedūros sintaksė:

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Kiekis
  /METHOD=ENTER Pakuotės_dydis Amžius2 Pajamos2
  /RESIDUALS DURBIN HIST(ZRESID) NORM(ZRESID) .
```



6.11 pav. **Dialogo langas *Linear Regression: Statistics***

Model Summary(b)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,765(a)	,585	,563	263,96190	2,009

a Predictors: (Constant), Pajamos2, Amžius2, Pakuotės_dydis
b Dependent Variable: Kiekis

6.12 pav. **Daugialypės regresinės analizės rezultatas**

6.4. Logistinė regresija

Tais atvejais, kai reikia nustatyti vieno priklausomojo dichotominio ir vieno ar daugiau nepriklausomųjų kintamųjų ryšį, taikomas logistinės regresijos modelis. Logistine regresija naudojamasi tada, kai priklausomasis kintamasis gali įgyti tik dvi reikšmes. Pavyzdžiui, pirkėjas gali pirkti prekę, bet gali jos nepirkti ir panašiai. Šiuo atveju standartiniai regresijos modeliai netinka, nes jie neįvertina, kad priklausomojo kintamojo pobūdis yra dichotominis, todėl sudarytu modeliu galima prognozuoti reikšmes didesnes už 1 ir mažesnes už 0. Tačiau regresijos modelį galima sudaryti taip, kad vietoj dvejetainio kintamojo bus prognozuojamas tolydusis kintamasis, kuris natūraliai kinta nuo 0 iki 1. Dažniausiai taikoma logistinė regresija (*logit regression*).

Logistinės regresijos modelis

Taikant logistinę regresiją galima tirti ir prognozuoti priklausomojo dichotominio kintamojo ir nepriklausomųjų kintamųjų, matuojamų bet kuria skale (nepriklausomieji kintamieji ir liekanos gali neturėti normaliojo skirstinio), ryšius. Prognozuojamos reikšmės niekada nebus didesnės (ir nebus lygios) už 1 ir mažesnės (ir lygios) už 0. Tai pasiekiama taikant logistinės regresijos modelį:

$$y = \frac{\exp(a + b_1 x_1 + b_2 x_2 + ... + b_n x_n)}{1 + \exp(a + b_1 x_1 + b_2 x_2 + ... + b_n x_n)} \qquad (6.9)$$

Akivaizdu, kad ir kokio dydžio x reikšmės būtų, modelio prognozuojamos y reikšmės visada bus nuo 0 iki 1. Kadangi priklausomasis kintamasis y kinta nuo 0 iki 1, jis gali būti nagrinėjamas kaip tolydi tikimybė p, kuri kinta nuo 0 iki 1. Ši tikimybė gali būti transformuojama:

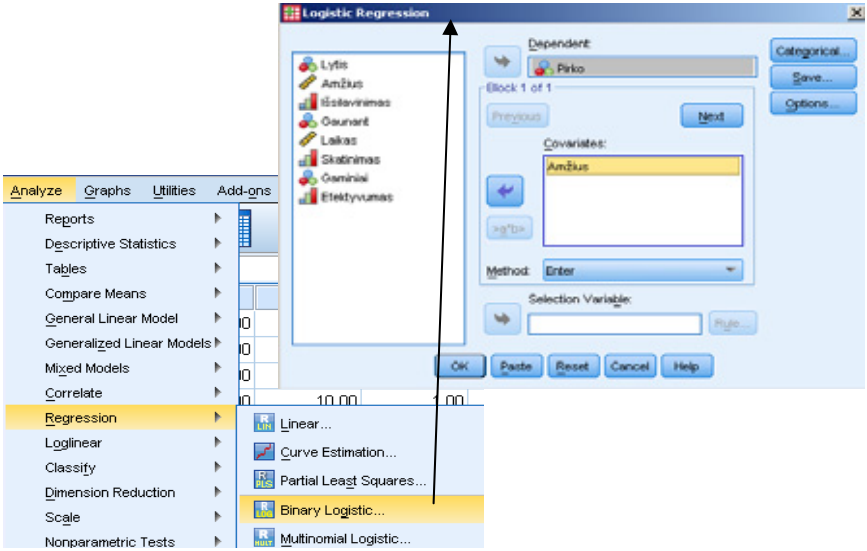
$$p' = \ln\left(\frac{p}{1-p}\right). \qquad (6.10)$$

Ši transformacija vadinama logistine transformacija. Kadangi p' teoriškai gali įgyti bet kurią reikšmę nuo minus iki plius begalybės, tai transformuotas reikšmes jau galima taikyti įprastame tiesinės regresijos modelyje:

$$p' = a + b_1 x_1 + b_2 x_2 + ... + b_n x_n + \text{e}. \qquad (6.11)$$

6.4.1. Logistinė regresija SPSS terpėje

Išnagrinėsime, kaip logistinė regresija gali būti taikoma nustatant binarinio kintamojo priklausomybę nuo kiekybinio kintamojo. Šiuo atveju taikoma binarinė logistinė regresija. Nustatykime, kaip tikimybė pirkti/nepirkti gėles priklauso nuo amžiaus. Dialogo langas *Logistic Regression* atveriamas komanda *Analyze* → *Regression* → *Binary Logistic*... (žr. 6.13 pav.). Priklausomasis binarinis kintamasis *Pirko* perkeliamas į langelį *Dependent*, kiekybinis kintamasis *Amžius* – į langelį *Covariates*. Paspaudę mygtuką *Continue* grįžtame į pradinį langą. Procedūra baigiama mygtuko *OK* paspaudimu. Lange *Output-SPSS Viewer* rodomi analizės rezultatai (žr. 6.14 pav.).



6.13 pav. **Dialogo langas *Logistic Regression***

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	24,110	1	,000
	Block	24,110	1	,000
	Model	24,110	1	,000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	104,546[a]	,228	,305

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than ,001.

Classification Table[a]

			Predicted		
			Pirko		Percentage Correct
	Observed		ne	taip	
Step 1	Pirko	ne	32	12	72,7
		taip	16	33	67,3
	Overall Percentage				69,9

a. The cut value is ,500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Ambius	,081	,019	18,898	1	,000	1,084
	Constant	-3,116	,759	16,852	1	,000	,044

a. Variable(s) entered on step 1: Ambius.

6.14 pav. **Pagrindiniai logistinės regresijos skaičiavimo rezultatai**

Nustatant, ar logistinės regresijos modelis tinka prognozėms, būtina patikrinti, ar dažnai žinomos reikšmės sutampa su reikšmėmis, kurias, remdamiesi logistine regresija, prognozuotume. Jeigu duomenų, iš kurių gauti parametrų įverčiai, prognozės netikslios, tada logistinės regresijos taikyti negalima, net jei kiti rodikliai ir statistinės išvados rodytų, kad ji tinkama. Lentelėje *Classification Table* pateikiamos priklausomojo kin-

tamojo stebimos (*Observed*) ir remiantis apskaičiuotu regresijos modeliu tikėtinos (*Predicted*) reikšmės. Iš 48 kintamojo *Pirko* reikšmių 0 (*ne*) testu buvo patvirtintos 32, o kitos 16 priskirtos prie 1 (*taip*). Iš 45 kintamojo *Pirko* reikšmių 1 (*taip*) buvo patvirtintos 33, o kitos 12 priskirtos prie 0 (*ne*). Bendras teisingo atpažinimo procentas yra 69,9 procentai. Nėra griežtų taisyklių, nusakančių, koks teisingo klasifikavimo procentas laikytinas patenkinamu. Paprastai reikalaujama, kad kiekvienai kategorijai jis būtų ne mažesnis kaip 50 procentų, o logistinė regresija apskritai taikoma tik tuo atveju, kai $y_i = 0$ sudaro ne mažiau kaip 20 procentų ir ne daugiau kaip 80 procentų visų stebėjimų.

Lentelėje *Omnibus Tests of Model Coefficients* pateiktas Pirsono χ^2 modelio suderinamumo kriterijus, kuris yra regresijos *ANOVA* analogas. Formuluojamos hipotezės:

H_0: $b = 0$,
H_1: $b \neq 0$.

Hipotezė H_0 atmetama, jeigu *p-level* < α. Hipotezė H_0 neatmetama, jeigu *p-level* ≥ α, čia α – reikšmingumo lygmuo. Šio pavyzdžio *p-level* (*Sig.*) = *0,000* < *0,05*, todėl nulinę hipotezę atmetame.

Logistinė regresija turi net keletą tiesinės regresijos determinacijos koeficientų analogų. Visiems jiems galioja ta pati taisyklė: kuo didesnis koeficientas, tuo geriau logistinė regresija suderinta su duomenimis, nors patys koeficientai yra nelyginami dėl skirtingo jų apibrėžimo. *Cox & Snell R Square* pseudodeterminacijos koeficientas ir normuotas šio koeficiento variantas – *Nagelkerke R Square* pateikiami lentelėje *Model Summary*. Koeficientai rodo tą dispersijos dalį, kurią paaiškina logistinės regresijos modelis. Šioje lentelėje taip pat pateikiama dviguba tikėtinumo funkcijos logaritmo reikšmė *–2 Log likelihood.*

Lentelėje *Variables in the Equation* pateikiamos apskaičiuotų koeficientų *a* ir *b* reikšmės (*B* stulpelyje) ir jų nelygybės nuliui reikšmingumas. Voldo (*Wald*) kriterijus yra Stjudento kriterijaus tiesinėje regresijoje analogas, t. y. jis atsako į klausimą, ar konkretus koeficientas lygus nuliui. Formuluojamos hipotezės:

H_0: $b = 0$,
H_1: $b \neq 0$.

Hipotezė H_0 atmetama, jeigu *p-level* < α. Hipotezė H_0 neatmetama, jeigu *p-level* ≥ α, čia α – reikšmingumo lygmuo. Visi šio pavyzdžio regresijos lygties koeficientai reikšmingai nelygūs nuliui, nes *p-level* (*Sig.*) < *0,05*. Šios lentelės *Exp*(*B*) stulpelyje pateikiama galimybių santykio reikšmė. Galimybių santykis rodo, kaip kinta Y galimybė įgyti reikšmę 1. Remiantis šiuo pavyzdžiu, galima teigti, kad kiekvieni papildomi metai amžiaus padidina tikimybę pirkti gėles 1,084 karto.

Rezultatų interpretavimas

Gauti koeficientai interpretuojami taip pat, kaip ir tiesinės regresijos. Galima užrašyti regresijos lygtį:

$$\hat{p}' = -3{,}116 + 0{,}081 \cdot amžius\,. \qquad (6.12)$$

Šios lygties parametrai susiję su $\ln\left(p/(1-p)\right)$, o ne su pirkimo tikimybėmis. Norint apskaičiuoti pirkimo tikimybę konkrečiu atveju, taikoma (6.12) formulė. Pavyzdžiui, kai amžius 30 metų, tikimybė pirkti apskaičiuojama taip:

$$\hat{p}' = -3{,}116 + 0{,}081 \cdot 30 = -0{,}686,$$

$$\hat{p}\,(30) = exp\,(-0{,}686) / (1+exp\,(-0{,}686)) = (0{,}504/1{,}504) = 0{,}335.$$

Atliktos procedūros sintaksė:

```
LOGISTIC REGRESSION VARIABLES Pirko
  /METHOD = ENTER Amžius
  /CRITERIA = PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .
```

KONTROLINIAI KLAUSIMAI

6.1. Kam reikalingas regresijos modelis?
6.2. Kaip nustatomas regresijos modelio tinkamumas?
6.3. Kokie būna regresijos modeliai?
6.4. Koks skirtumas tarp paprastosios ir logistinės regresijos?
6.5. Ar logistinės regresijos modelis tinka prognozei?

6.6. Kokiais atvejais taikoma logistinė regresija?
6.7. Koks tiesinės ir logistinės regresijos ryšys?

UŽDUOTYS

6.1 užduotis

Analizuojant vidutiniškai išleidžiamų kavinėje pinigų priklausomybę nuo amžiaus, gautas rezultatas, pateiktas 6.15 pav. Paaiškinkite gautą rezultatą.

Model Summary and Parameter Estimates

Dependent Variable: Vidut_saskaita

	Model Summary					Parameter Estimates			
Equation	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	,235	15,013	1	49	,000	8,607	,478		
Logarithmic	,258	17,023	1	49	,000	-26,455	14,780		
Quadratic	,266	8,676	2	48	,001	-5,486	1,452	-,015	
Cubic	,268	5,741	3	47	,002	-17,866	2,711	-,054	,000
Growth	,182	10,912	1	49	,002	2,480	,019		
Exponential	,182	10,912	1	49	,002	11,944	,019		

The independent variable is Amzius.

6.15 pav. **Regresinės analizės rezultatas**

6.2 užduotis

Išnagrinėkite 6.15 pav. pateiktą rezultatą. Kokie regresijos modeliai yra tinkami? Užrašykite regresijos lygtis. Suformuluokite hipotezes ir patikrinkite lygčių koeficientų patikimumą. Koks modelis yra tinkamiausias, kodėl?

6.3 užduotis

Analizuojant kelionės lėktuvu pasirinkimo tikimybės nuo planuojamų kelionės išlaidų priklausomybę buvo atlikta logistinės regresijos procedūra. Skaičiavimo rezultatas pateiktas 6.16 pav. Išnagrinėkite gautą rezultatą ir įvertinkite logistinės regresijos modelio tinkamumą.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	11,512	1	,001
	Block	11,512	1	,001
	Model	11,512	1	,001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	38,505[a]	,261	,357

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

Classification Table[a]

			Predicted		
			Kelione_lektuvu		Percentage
	Observed		,00	1,00	Correct
Step 1	Kelione_lektuvu	,00	21	3	87,5
		1,00	6	8	57,1
	Overall Percentage				76,3

a. The cut value is ,500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Planuojamos_islaidos	,002	,001	6,464	1	,011	1,002
	Constant	-3,486	1,232	8,010	1	,005	,031

a. Variable(s) entered on step 1: Planuojamos_islaidos.

6.16 pav. **Pagrindiniai logistinės regresijos skaičiavimo rezultatai**

6.4 užduotis

Nagrinėdami logistinės regresijos rezultatą (žr. 6.16 pav.) užrašykite logistinės regresijos lygtį. Apskaičiuokite tikimybę kelionės lėktuvu pasirinkimą dėl skirtingų planuojamų kelionės išlaidų. Kaip keičiasi tikimybė pasirinkti kelionę lėktuvu, jeigu planuojamos išlaidos padidėja 1 litu?

7. SKIRTUMŲ ANALIZĖ

Rinkos tyrimų praktikoje, atliekant rinkos segmentavimą, dažnai reikia palyginti skirtingas respondentų grupes. Skirtumų analizės tikslas – nustatyti respondentų grupes, tarp kurių yra reikšmingas skirtumas. Statistinės procedūros, kurios leidžia nustatyti tokius skirtumus (*t* kriterijus ir dispersinė analizė), lygina kintamųjų vidurkius. *T* kriterijus taikomas nustatant dviejų respondentų grupių skirtumus. Skirtumų analizei tinka kiekybiniai kintamieji. Ranginiai kintamieji gali būti apdorojami kaip kiekybiniai, jei rangų skaičius lygus 15 ir daugiau.

7.1. Hipotezės apie vidurkių lygybę

Vidurkių lygybė dažniausiai tikrinama *t* kriterijumi arba dispersine analize. Dispersinė analizė – tai universalus metodas, *t* kriterijaus taikymas turi imčių skaičiaus apribojimus. Į dispersinę analizę galima žiūrėti kaip į *t* kriterijaus apibendrinimą: lyginant dvi imtis, vienfaktorinės dispersinės analizės ir *t* kriterijaus taikymo rezultatai sutampa. Teoriškai *t* kriterijus gali būti taikomas net ir tada, kai imtis yra labai maža (pavyzdžiui, mažesnė nei 10), tačiau būtina, kad kiekvienos imties kintamųjų skirstinys būtų normalusis ir imčių dispersijų skirtumas neryškus. Normalumą galima patikrinti vizualiai, χ^2 arba Kolmogorovo–Smirnovo kriterijais.

SPSS taiko tris pagrindinius *t* kriterijaus tipus:

- *t* kriterijus vienai imčiai;
- *t* kriterijus dviem nepriklausomoms imtims;
- *t* kriterijus dviem priklausomoms imtims.

Taikomų *t* kriterijui skaičiuoti kintamųjų tipai pateikti 1 lentelėje.

1 lentelė. **Kintamųjų tipai *t* kriterijui skaičiuoti**

T kriterijus vienai imčiai			
Priklausomieji kintamieji		Nepriklausomieji kintamieji	
Kiekis	Tipas	Kiekis	Tipas
Nėra	Nėra	Neribojamas	Kiekybinis
T kriterijus dviem nepriklausomoms imtims			
Priklausomieji kintamieji		Nepriklausomieji kintamieji	
Kiekis	Tipas	Kiekis	Tipas
Vienas	Dichotominis Kiekybinis	Neribojamas	Kiekybinis
T kriterijus dviem priklausomoms imtims			
Priklausomieji kintamieji		Nepriklausomieji kintamieji	
Kiekis	Tipas	Kiekis	Tipas
Nėra	Nėra	Du	Kiekybinis

Priklausomasis kintamasis taikomas tik *t* kriterijaus dviem nepriklausomoms imtims skaičiuoti. Kitais atvejais priklausomųjų kintamųjų nėra, todėl kad analizuojama iš esmės viena imtis. Nepriklausomasis (analizuojamas) kintamasis turi būti tik kiekybinis.

7.2. *T* kriterijus vienai imčiai

***T* kriterijus vienai imčiai** (*One sample t test*) taikomas tada, kai yra viena duomenų imtis ir reikia patikrinti, ar kintamojo vidurkis lygus tam tikram fiksuotam skaičiui.

Dažnai reikia nustatyti, ar nežinomas populiacijos vidurkis skiriasi nuo tam tikro skaičiaus. Paprastai dispersijos reikšmė yra nežinoma. Nežinoma populiacijos dispersija keičiama jos įverčiu s^2. Iš imties apskaičiuota populiacijos dispersijos reikšmė nėra tiksli ir galimas tam tikras skirtumas nuo tikrosios populiacijos dispersijos. Tikrinama *hipotezė apie vidurkio lygybę skaičiui*, *kai dispersija nežinoma.* Pavyzdžiui, norima nustatyti, ar vartotojams priimtina A jogurto kaina statistiškai skiriasi nuo pardavėjų nustatytos 3,5 Lt kainos. Iš imties duomenų galima

apskaičiuoti vidurkį $\bar{x}$ ir standartinį nuokrypį *s*. Formuluojamos statistinės hipotezės:

H_0: $\mu = a$ ($a = 3,5$, priimtina kaina nesiskiria nuo nustatytos),
H_1: $\mu \neq a$ ($a = 3,5$, priimtina kaina skiriasi nuo nustatytos).

Pasirenkamas reikšmingumo lygmuo α. Kadangi standartinis nuokrypis tiksliai nėra žinomas (jo reikšmė apytikriai apskaičiuota iš imties) tikrinant hipotezę taikomas *t* kriterijus:

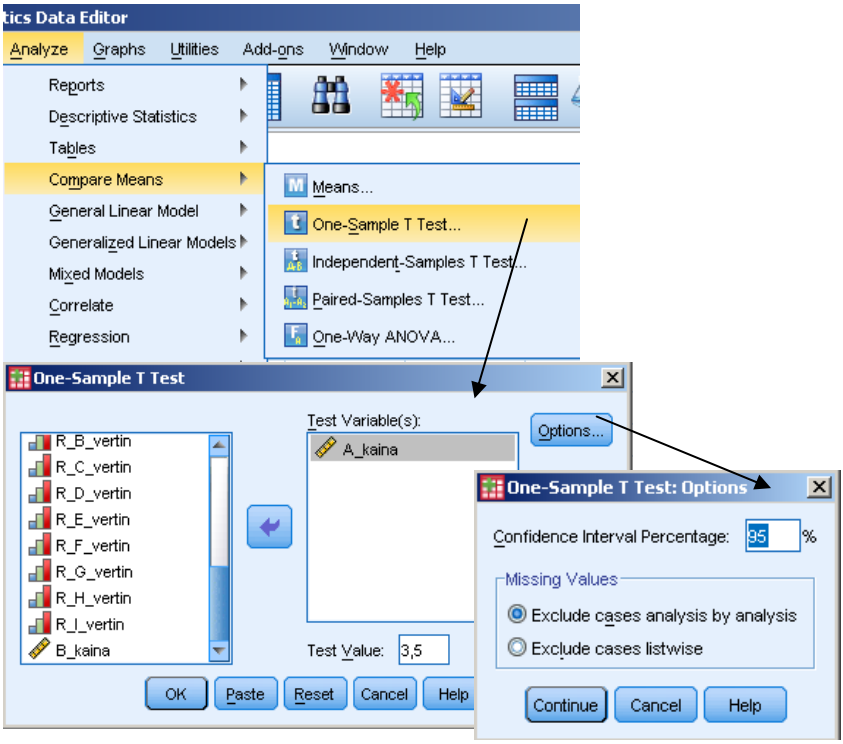
$$t = \frac{\bar{x} - a}{s / \sqrt{n}}, \qquad (7.1)$$

čia: $\bar{x}$ – imties vidurkis,
s – imties standartinis nuokrypis,
n – imties dydis,
a – konstanta, su kuria lyginamas vidurkis.

Šio *t* kriterijaus reikšmės pasiskirstę pagal Stjudento skirstinį su $(n - 1)$ laisvės laipsniu, kai populiacijos vidurkis $\mu = a$. Stjudento skirstinys simetriškas nulio atžvilgiu, todėl esant dvipusei alternatyvai H_1: $\mu \neq a$ kritinės reikšmės yra $t_{\alpha/2}(n-1)$. Priimant sprendimą reikia žinoti $t_{0,025}(n-1)$ reikšmę. Ją galima sužinoti iš statistinių lentelių (žr. Priedai. 2 lentelė). Jei apskaičiuota $t \leq t_{0,025}(n-1)$, H_0 atmesti negalima. Jei apskaičiuota $t > t_{0,025}(n-1)$, H_0 atmetama.

7.2.1. *T* kriterijaus vienai imčiai skaičiavimas SPSS terpėje

Norint nustatyti, ar vartotojams priimtina A jogurto kaina statistiškai skiriasi nuo pardavėjų nustatytos 3,5 Lt kainos, meniu *Analyze* pasirenkame komandą *Compare Means* → *One-Sample T Test...* Atsiveria dialogo langas *One-Sample T Test* (žr. 7.1 pav.).



7.1 pav. **Dialogo langas *One-Sample T Test***

Kintamasis *A_kaina* perkeliamas į langelį *Test Variable*(*s*). Langelyje *Test Value* nurodoma standartinė reikšmė (3,5), su kuria bus lyginamas analizuojamo kintamojo vidurkis. Mygtukas *Options*… leidžia nustatyti patikimumo lygmenį. Procedūra baigiama mygtuko *OK* paspaudimu. Lange *Output-SPSS Viewer* rodomi analizės rezultatai (žr. 7.2 pav.). Pirmoje lentelėje *One-Sample Statistics* pateikiamas apskaičiuotas kintamojo vidurkis (*Mean*) ir standartinis nuokrypis (*Std. Deviation*). Lentelė *One-Sample Test* suteikia galimybę įvertinti statistiškai reikšmingus skirtumus. Išvada daroma vertinant stebimojo reikšmingumo lygmens *p-level* (*Sig.* (*2 tailed*)) reikšmę. Jei $p\text{-level} > \alpha$, tai kintamųjų vidurkių skirtumas statistiškai nereikšmingas. Nagrinėjamu atveju

p-level = 0,006 < α = *0,05*, skirtumas 0,5 Lt (stulpelis *Mean Difference*) yra statistiškai reikšmingas.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
A_kaina	60	3,0000	1,35296	,17467

One-Sample Test

	Test Value = 3.5					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
A_kaina	-2,863	59	,006	-,50000	-,8495	-,1505

7.2 pav. ***T* kriterijaus skaičiavimo rezultatas**

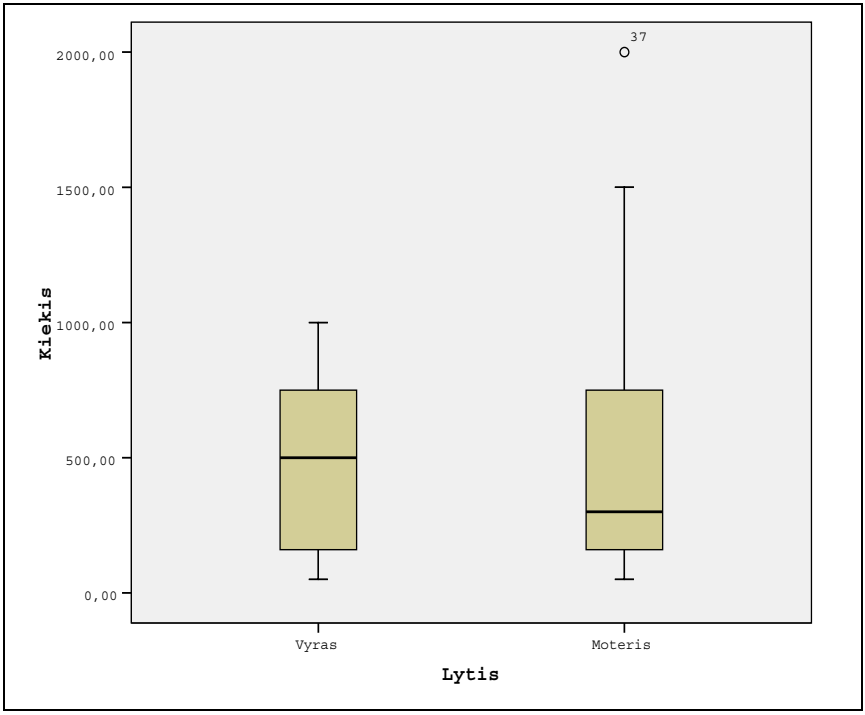
Atliktos procedūros sintaksė:

```
T-TEST
  /TESTVAL = 3.5
  /MISSING = ANALYSIS
  /VARIABLES = A_kaina
  /CRITERIA = CI(.95) .
```

7.3. *T* kriterijus dviem nepriklausomoms imtims

***T* kriterijus dviem nepriklausomoms imtims** (*Independent–samples t test*) yra taikomas tada, kai tiriamos dvi atskiros imtys ir norima sužinoti, ar šių imčių vidurkiai reikšmingai skiriasi. Šiuo atveju nepriklausomos imtys suprantamos kaip imtys, atrinktos pagal binarinę (dichotominę kategoriją), pavyzdžiui, vyrai ir moterys, žmonės, kurie

perka ir neperka kokio nors produkto, t. y. kai egzistuoja du variantai priklausomojo (grupuojančio) kintamojo ir keletas nepriklausomųjų kintamųjų, kurių analizė ir parodys šių grupių skirtumą. Kai tiriamos dvi *atskiros* imtys ir norima sužinoti, ar šių imčių vidurkiai reikšmingai skiriasi, taikomas *t* kriterijus, skirtas dviem nepriklausomoms imtims. Imtys gaunamos stebint du nepriklausomuosius atsitiktinius dydžius, turinčius normalųjį skirstinį. Šių dydžių vidurkiai yra nežinomi. Tikrinama *hipotezė apie dviejų imčių vidurkių lygybę*, *kai dispersija nežinoma*.



7.3 pav. **Dviejų grupių palyginimas taikant „dėžės ir ūsų“ diagramą**

Atliekant tyrimą kompiuteriu, grupės gali būti atskiros arba joms atskirti galima taikyti grupavimo kintamąjį (tarkim, *Lytis: vyrai/moterys*). Taikant grupavimo kintamąjį tiriama, ar priklausomojo kintamojo (pavyzdžiui, *Kiekis*) vidurkiai yra lygūs. Skirtumus tarp grupių vidur-

kių ir dispersijų vizualiai galima stebėti taikant „dėžės ir ūsų“ diagramas (žr. 7.3 pav.). Stačiakampis („dėžė“) vaizduoja medianą (vidurinė linija), 25-ą ir 75-ą procentilius, taigi stačiakampio viduje yra 50 procentų reikšmių. Reikšmės, nutolusios nuo stačiakampio kraštų daugiau nei per tris „dėžės“ ilgius, vadinamos išskirtimis (*extreme values*). Jos žymimos žvaigždutėmis (*). Reikšmės, nutolusios nuo 1,5 iki 3 „dėžės“ ilgių, vadinamos sąlyginėmis išskirtimis (*outliers*) ir žymimos rutuliukais. Didžiausia ir mažiausia reikšmės, kurios nėra išskirtys, brėžiamos kaip „dėžės ūsai“.

Taikant *t* kriterijų dviem imtims, būtina, kad abi imtys būtų atrinktos iš populiacijų, turinčių normaliuosius skirstinius ir turėtų vienodas dispersijas. Dažniausiai dispersijų lygybei tikrinti taikomas *F* arba *Livyno* kriterijus. *F* kriterijus labai jautriai reaguoja į normalumo sąlygos netenkinimą. Jei populiacijos neturi normaliojo skirstinio, *F* kriterijus gali atmesti nulinę hipotezę teigiančią dispersijų lygybę ir tais atvejais, kai populiacijų dispersijos iš tiesų yra lygios. Normalumo prielaidos pažeidimui mažiau jautrus yra *Livyno* kriterijus.

***F* kriterijus** nustatomas apskaičiavus imčių dispersijas s_1^2 ir s_2^2:

$$F = \frac{s_1^2}{s_2^2}. \qquad (7.2)$$

Hipotezė H_0 atmetama (t. y. teigiama, kad dispersijos reikšmingai skiriasi), jei gauta reikšmė *F* yra didesnė už kritinę reikšmę $F > F_{\alpha/2}(n - 1, m - 1)$, čia α – pasirinktas reikšmingumo lygmuo, *n* ir *m* – pirmosios ir antrosios imties dydžiai.

Suformuluokime statistines hipotezes:

H_0: $\mu_1 = \mu_2$,
H_1: $\mu_1 \neq \mu_2$.

Įvertinę dispersijų lygybę, renkamės vieną iš dviejų galimų *t* kriterijaus skaičiavimo būdų.

Stjudento *t* kriterijus, kai populiacijų dispersijos yra lygios

Kadangi populiacijos dispersijos nežinomos, jos keičiamos dispersijų svertiniu vidurkiu:

$$s_p^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}. \qquad (7.3)$$

T kriterijus skaičiuojamas pagal formulę:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{m}}}. \qquad (7.4)$$

Kritinė sritis sudaroma remiantis tuo, kad statistika (7.4) turi Stjudento skirstinį su ($n + m - 2$) laisvės laipsniais, kai $\bar{x}_1 = \bar{x}_2$. Hipotezė H_0 atmetama, jei apskaičiuota $|t| > t_{\alpha/2}(n + m - 2)$.

Stjudento *t* kriterijus, kai populiacijų dispersijos nelygios

Jei jungtinę dispersiją taikysime tada, kai dispersijos nelygios, apskaičiuota *t* kriterijaus reikšmė gali būti klaidinga. Klaidos dydis priklauso nuo imčių dydžių ir dispersijų skirtumų. Didelėms imtims neatitikimai tarp rezultatų, gautų šiais dviem metodais, yra maži. Kai dispersijos nelygios hipotezei tikrinti taikomas *t* kriterijus:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}. \qquad (7.5)$$

Kadangi alternatyva dvipusė, sprendimui priimti reikia taikyti $t_{\alpha/2}(k)$ reikšmę. Laisvės laipsnių skaičius *k* yra mažiausias sveikasis skaičius, tenkinantis sąlygą:

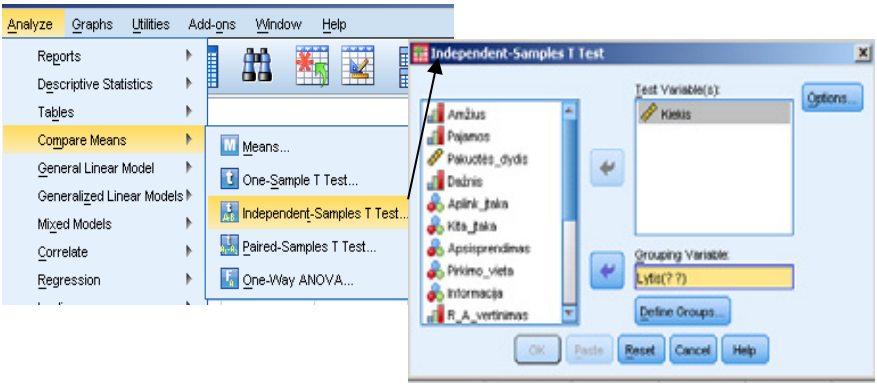
$$k \le \frac{(s_1^2/n + s_2^2/m)^2}{s_1^4/n^3 + s_2^4/m^3}. \qquad (7.6)$$

Hipotezė H_0 atmetama, jeigu apskaičiuota $|t| > t_{\alpha/2}(k)$.

7.3.1. *T* kriterijaus dviem nepriklausomoms imtims skaičiavimas SPSS terpėje

Pavyzdžiui, analizuojamas vyrų ir moterų suvartojamo per mėnesį kavos kiekio skirtumas.

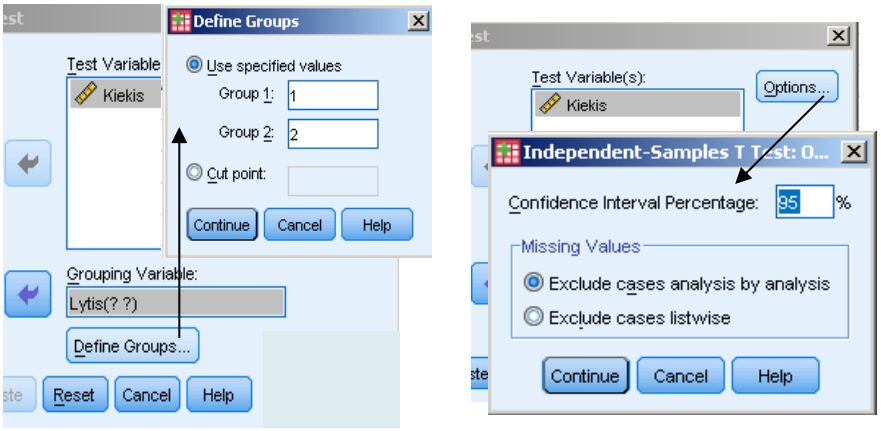
Meniu *Analyze* pasirenkame komandas *Compare Means* → *Independent-Samples T Test*. Atsiveria dialogo langas *Independent-Samples T Test* (žr. 7.4 pav.).



7.4 pav. **Dialogo langas *Independent-Samples T Test***

Nepriklausomąjį kintamąjį *Kiekis* perkeliame į *Test Variable*(*s*) lauką, o priklausomąjį (grupuojantį) kintamąjį *Lytis* – į *Grouping Variable* lauką. Šis kintamasis gali įgyti dvi užkoduotas reikšmes: 1 – vyras, 2 – moteris. Jas reikia nurodyti dialogo lange, kuris atsiveria paspaudus mygtuką *Define Groups*… (žr. 7.5 pav.).

Mygtuku *Options* atveriamas dialogo langas *Independent-Samples T Test: Options* (žr. 7.5 pav.). Šiame lange galima nustatyti *t* kriterijaus skaičiavimo patikimumo lygmenį (*Confidence Interval Percentage*). Nustatytasis yra 95 procentų lygmuo. Mygtuko *Continue* spragtelėjimu grąžinamas *Independent-Samples T Test* langas. Procedūra baigiama mygtuko *OK* paspaudimu. Lange *Output-SPSS Viewer* rodomi analizės rezultatai (žr. 7.6 pav.).



7.5 pav. **Dialogo langai *Define Groups* ir *Independent-Samples T Test: Options***

Group Statistics

	Lytis	N	Mean	Std. Deviation	Std. Error Mean
Kiekis	Vyras	30	498,0000	356,33576	65,05771
	Moteris	30	473,0000	443,83245	81,03235

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Kiekis	Equal variances assumed	,011	,916	,241	58	,811	25,00000	103,91702	-183,01252	233,012
	Equal variances not assumed			,241	55,412	,811	25,00000	103,91702	-183,21956	233,219

7.6 pav. ***T* kriterijaus skaičiavimo rezultatas**

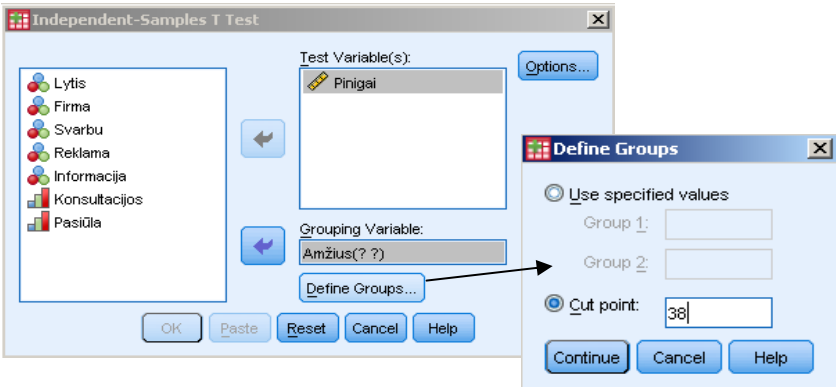
Pirmoje lentelėje *Group Statistics* pateikti kintamojo *Kiekis* vidurkiai (*Mean*) respondentų grupėms *Vyras* ir *Moteris.* Lentelė *Independent Samples Test* leidžia nustatyti statistinius grupių skirtumus. Antros len-

telės *Independent Samples Test* analizę reikia pradėti nuo Livyno kriterijaus (*Levene's Test*) analizės. Šis kriterijus taikomas hipotezės dėl analizuojamų kintamųjų dispersijų lygybei tikrinti. Jeigu stebimojo reikšmingumo lygmens *p-level* reikšmė *Sig.* stulpelyje *Levene's Test for Equality of Variances* rodo, kad rezultatas statistiškai nereikšmingas (*p-level* > α, kur α – pasirenkamas reikšmingumo lygmuo, pavyzdžiui, α = *0,05*) (nagrinėjamu atveju *p-level = 0,916*), tai kintamųjų dispersijos statistiškai lygios ir vidurkių skirtumas nustatomas iš eilutės *Equal variances assumed* stulpelio *Sig.* (*2-tailed*). Jei rezultatas statistiškai reikšmingas, tai kintamųjų dispersijos statistiškai skiriasi ir vidurkių skirtumas nustatomas iš eilutės *Equal variances not assumed* stulpelio *Sig.* (*2-tailed*). Nagrinėjamu atveju Livyno kriterijus rodo statistinį reikšmingumą, o stulpelio *Sig.* (*2-tailed*) reikšmė *0,811* (*p-level* > α) rodo, kad vyrų ir moterų suvartojamos kavos kiekio skirtumas statistiškai nereikšmingas.

Atliktos procedūros sintaksė:

```
T-TEST
  GROUPS = Lytis(1 2)
  /MISSING = ANALYSIS
  /VARIABLES = Kiekis
  /CRITERIA = CI(.95) .
```

Išnagrinėsime atvejį, kai grupuojantis kintamasis ne dichotominis, o kiekybinis. Pavyzdžiui, norint analizuoti skirtingo amžiaus respondentų sumokamų už kavą pinigų kiekio skirtumą (žr. 7.7 pav.).



7.7 pav. **Dialogo langai *Independent-Samples T Test* ir *Define Groups***

Group Statistics

	Amžius	N	Mean	Std. Deviation	Std. Error Mean
Pinigai	>= 38,00	18	8,0556	3,26248	,76897
	< 38,00	18	10,1667	3,20386	,75516

Independent Samples Test

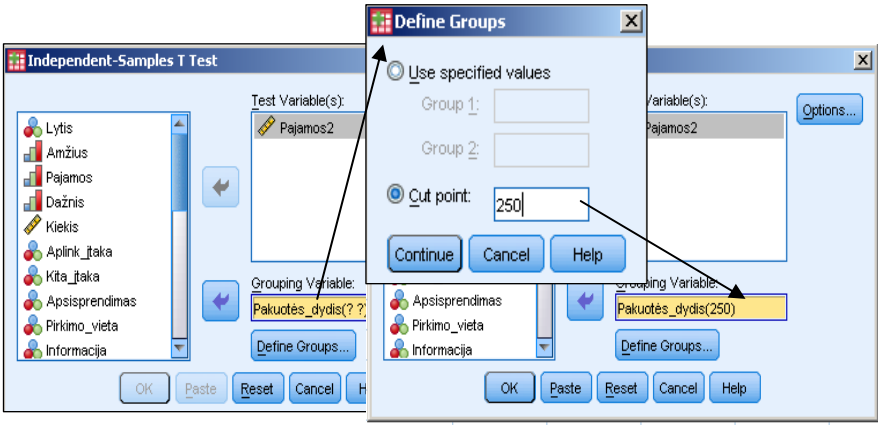
		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Pinigai	Equal variances assumed	,416	,524	-1,959	34	,058	-2,11111	1,07777	-4,30140	,07918
	Equal variances not assumed			-1,959	33,989	,058	-2,11111	1,07777	-4,30143	,07920

7.8 pav. ***T* kriterijaus skaičiavimo rezultatas**

Šiuo atveju lange *Define Groups* galima nustatyti atskirties tašką (*Cut point*), kuris suskirstys visas galimas kintamojo *Amžius* reikšmes į dvi grupes. Daugiau analizės procedūra nesiskiria. Gautas rezultatas pateiktas 7.8 pav. Sprendimų priėmimas analogiškas anksčiau pateiktam pavyzdžiui. Pirmoje lentelėje *Group Statistics* pateikti kintamojo *Pinigai* vidurkiai (*Mean*) skirtingo amžiaus respondentų grupėms >= *38,00* ir < *38,00*. Lentelė *Independent Samples Test* leidžia nustatyti statistinius grupių skirtumus. Antros lentelės *Independent Samples Test* analizę reikia pradėti nuo Livyno kriterijaus (*Levene's Test*) analizės. Stebimojo reikšmingumo lygmens *p-level* reikšmė *Sig*. stulpelyje *Levene's Test for Equality of Variances* rodo, kad $p\text{-level} = 0{,}524 > \alpha = 0{,}05$, kur α – pasirenkamas reikšmingumo lygmuo, kintamųjų dispersijos statistiškai lygios ir vidurkių skirtumas nustatomas iš eilutės *Equal variances assumed* stulpelio *Sig.* (*2-tailed*). Nagrinėjamu atveju stulpelio *Sig.* (*2-tailed*)

reikšmė *0,058* (*p-level* > *α*) rodo, kad skirtingo amžiaus respondentų sumokamų už kavą pinigų kiekio skirtumas statistiškai nereikšmingas.

Išnagrinėsime atvejį, kaip skiriasi respondentų pajamos, atsižvelgiant į perkamos kavos pakuotės dydį (žr. 7.9 pav.). Gautas rezultatas pateiktas 7.10 pav.



7.9 pav. **Dialogo langai *Independent-Samples T Test* ir *Define Groups***

Lentelės *Independent Samples Test* stebimojo reikšmingumo lygmens *p-level* reikšmė *Sig*. stulpelyje *Levene's Test for Equality of Variances* rodo, kad *p-level* = *0,211* > *α* = *0,05*, kur *α* – pasirenkamas reikšmingumo lygmuo, kintamųjų dispersijos statistiškai lygios ir vidurkių skirtumas nustatomas iš eilutės *Equal variances assumed* stulpelio *Sig*. (*2-tailed*). Nagrinėjamu atveju stulpelio *Sig*. (*2-tailed*) reikšmė *0,005* (*p-level* < *α*) rodo, kad skirtingo dydžio pakuotes perkančiųjų respondentų pajamų skirtumas statistiškai reikšmingas.

Atliktos procedūros sintaksė:

```
T-TEST
  GROUPS = Pakuotės_dydis(250)
  /MISSING = ANALYSIS
  /VARIABLES = Pajamos2
  /CRITERIA = CI(.95) .
```

Group Statistics

	Pakuotės_dydis	N	Mean	Std. Deviation	Std. Error Mean
Pajamos2	>= 250,00	37	1636,4865	426,16648	70,06134
	< 250,00	23	1304,3478	444,64313	92,71450

Independent Samples Test

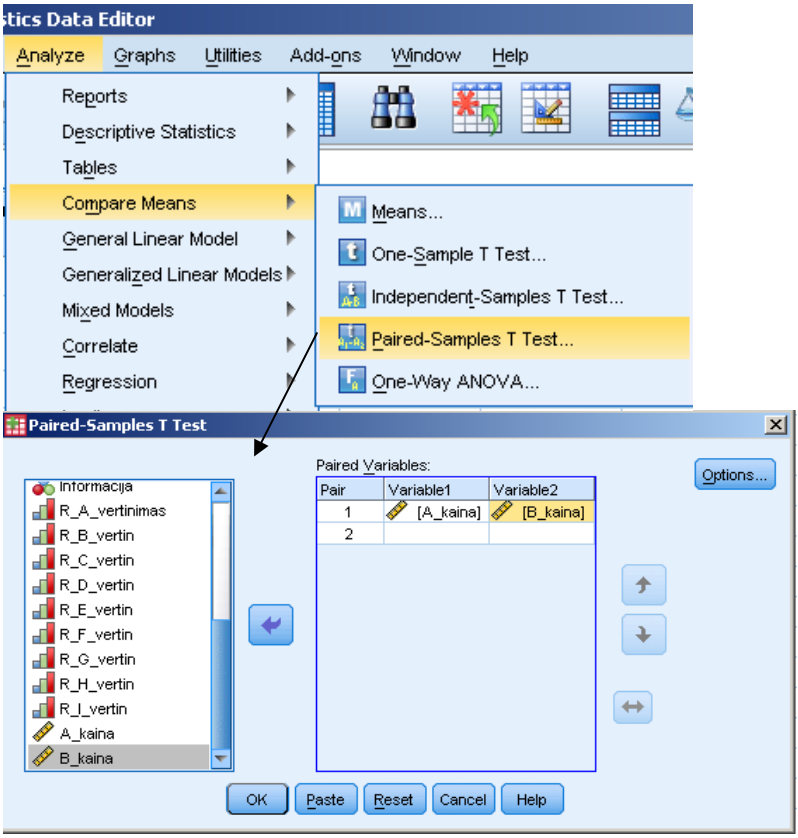
		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Pajamos2	Equal variances assumed	1,597	,211	2,887	58	,005	332,13866	115,04482	101,85142	562,42590
	Equal variances not assumed			2,858	45,277	,006	332,13866	116,20916	98,12089	566,15643

7.10 pav. ***T* kriterijaus skaičiavimo rezultatas**

7.4. *T* kriterijaus dviem priklausomoms (porinėms) imtims skaičiavimas SPSS terpėje

***T* kriterijus dviem priklausomoms (porinėms) imtims** (*Paired sample t test*) taikomas tada, kai dvi duomenų grupės sudaro poras. Tai reiškia, kad kiekvieną vienos grupės reikšmę atitinka kitos grupės reikšmė. Šio tipo testai taikomi, kai į skirtingus klausimus atsako ta pati respondentų grupė. Pavyzdžiui, vartotojai nurodo A ir B jogurtų priimtiną 500 gr pakuotės kainą.

Norėdami nustatyti, ar kainos įvertinimas statistiškai reikšmingas, meniu *Analyze* pasirenkame komandą *Compare Means* → *Paired-Samples T Test*. Atsiveria dialogo langas *Paired-Samples T Test.* Kintamieji *A_kaina* ir *B_kaina* iš eilės perkeliami į langelį *Paired Variables* (žr. 7.11 pav.). Procedūra baigiama mygtuko *OK* paspaudimu. Lange *Output-SPSS Viewer* rodomi analizės rezultatai (žr. 7.12 pav.).



7.11 pav. **Dialogo langas *Paired-Samples T Test***

Pirmoje lentelėje *Paired Samples Statistics* pateikti apskaičiuoti kintamųjų vidurkiai. Lentelėje *Paired Samples Correlations* pateiktas analizuojamų kintamųjų Pirsono koreliacijos koeficientas. Lentelė *Paired Samples Test* leidžia įvertinti statistiškai reikšmingus kintamųjų skirtumus. Išvada daroma vertinant stebimojo reikšmingumo lygmens *p-level* (*Sig. (2 tailed)*) reikšmę. Jei *p-level* > α, tai kintamųjų vidurkių skirtumas statistiškai nereikšmingas. Nagrinėjamu atveju *p-level* = *0,129* > α = *0,05*, skirtumas 0,4 Lt (žr. *Paired Differences* → *Mean*) yra statistiškai nereikšmingas.

Atliktos procedūros sintaksė:

```
T-TEST
  PAIRS = A_kaina WITH B_kaina (PAIRED)
  /CRITERIA = CI(.95)
  /MISSING = ANALYSIS.
```

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	A_kaina	3,0000	60	1,35296	,17467
	B_kaina	3,4000	60	1,13794	,14691

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	A_kaina & B_kaina	60	-,297	,021

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	A_kaina - B_kaina	-,40000	2,01014	,25951	-,91928	,11928	-1,541	59	,129

7.12 pav. ***T* kriterijaus skaičiavimo rezultatas**

7.5. Dispersinė analizė

Atliekant rinkos tyrimus dažnai reikia palyginti daugiau kaip dvi respondentų grupes, pavyzdžiui, ar kelių skirtingo amžiaus grupių atstovai išleidžia pirkiniams vienodą pinigų kiekį. Tuomet tenka lyginti kelių nepriklausomų imčių vidurkius ir, remiantis tyrimu, daryti išvadas apie populiacijas, iš kurių atrinktos imtys. Šiems uždaviniams spręsti taikoma dispersinė analizė. Į dispersinę analizę galima žiūrėti kaip į *t* kriterijaus apibendrinimą.

Dispersinėje analizėje vieni kintamieji nagrinėjami kaip priežastys, o kiti – kaip šių priežasčių veikimo pasekmės. Pirmojo tipo kintamieji yra *nepriklausomi* ir vadinami veiksniais, o jų veikimą atspindintys antrojo tipo kintamieji – *priklausomais* kintamaisiais. Pavyzdžiui, tiriamųjų asmenų amžius arba pirkimo vieta yra veiksniai, o sumokėti pinigai – priklausomasis kintamasis. Dispersinės analizės atveju priklausomasis kintamasis tas, kuris nustato skirtingų tiriamųjų grupių vidurkių skirtumą. Kintamasis, kuris apibūdina grupių išskyrimą, vadinamas nepriklausomuoju.

Vienfaktorinė dispersinė analizė (*One-Way ANOVA*) taikoma tada, kai populiacijos viena nuo kitos skiriamos tik pagal vieną kintamąjį (veiksnį). Dispersinės analizės tikslas – nuspręsti, ar priklausomojo kintamojo, išmatuoto skirtingose populiacijose, vidurkiai statistiškai reikšmingai skiriasi. Metodas leidžia atsakyti į klausimą, ar iš kelių nepriklausomų imčių bent dviejų vidurkiai statistiškai reikšmingai skiriasi. Veiksnio parinkimas dažnai yra subjektyvus ir specialistų nuomonės apie tai, kuris kintamasis yra veiksnys, o kuris priklausomasis kintamasis, gali skirtis.

Dispersinės analizės procedūra ANOVA reikalauja, kad būtų patenkintos šios prielaidos:

- kiekviena tiriama imtis turi būti nepriklausoma nuo kitų;
- tiriamas kintamasis kiekvienoje populiacijoje pasiskirstęs pagal normalųjį dėsnį;
- imčių dispersijos lygios.

Ar patenkinami šie reikalavimai, galima patikrinti pradinės duomenų analizės procedūromis (nubraižant „dėžės ir ūsų“ diagramą, apskaičiuojant dispersijas) ir taikant statistinius kriterijus, skirtus duomenų normaliajam skirstiniui bei dispersijų lygybei tikrinti. Taikant vienfaktorinę dispersinę analizę labai svarbu, kad visos imtys būtų panašaus dydžio. Kuo daugiau stebėjimų kiekvienoje imtyje, tuo statistiškai reikšmingesni tampa vidurkių skirtumai. Kai imtys yra panašaus dydžio, ANOVA gana atspari nukrypimams nuo normaliojo skirstinio ir dispersijų lygybės.

Normalumo prielaida. Kadangi ANOVA labai atspari normalumo prielaidos pažeidimams, ji taikoma net ir tuo atveju, kai kintamieji matuojami pagal rangų skalę. Svarbu, kad rangų būtų gana daug ir vidurkis turėtų aiškią prasmę. Nėra griežtų taisyklių, reglamentuojančių įgyjamų rangų skaičių. Dažniausiai manoma, kad turėtų būti bent 7 skirtingi ran-

gai. Kartais ANOVA taikoma ir kai duomenys išmatuoti 5 skirtingų rangų skale.

Dispersijų lygybės prielaida. Dispersijų lygybę galima patikrinti įvairiais būdais. Pats paprasčiausias yra reikalavimas, kad didžiausioji iš imčių dispersijų būtų didesnė už mažiausiąją ne daugiau kaip tris kartus. Griežčiau tikrinant dispersijų lygybę taikomi *F* ir *Livyno* (*Levene*) kriterijai. Jei šie kriterijai atmeta dispersijų lygybę, daromas paklaidos dydis labai priklauso nuo imčių dydžio. Jei didesnė imtis atrinkta iš populiacijos su didesne dispersija, apskaičiuotas stebimasis reikšmingumo lygmuo *p-level* gaunamas daug didesnis už tikrąją reikšmę. Todėl vidurkių skirtumo reikšmingumas pripažįstamas daug rečiau nei turėtų būti. Jei iš populiacijos su didesne dispersija atrinkta maža imtis apskaičiuota *p-level* reikšmė gaunama mažesnė už tikrąją ir skirtumo reikšmingumas pripažįstamas dažniau nei reikėtų. Kuo mažesni imčių dydžių skirtumai, tuo mažiau ANOVA jautri dispersijų skirtumams.

Tikrinamos hipotezės:

$$\begin{cases} H_0: \mu_1 = \mu_2 = \ldots = \mu_k, \\ H_1: \textit{bent du vidurkiai skiriasi}, \end{cases}$$

čia: μ_j – nepriklausomųjų kintamųjų vidurkiai,

k – imčių skaičius.

Dispersinė analizė remiasi faktu, kad dispersija gali būti padalyta. Dispersija yra nuokrypių nuo vidurkio kvadratų sumos funkcija. Visa kvadratų suma *SST* (*sum of squares total*), kuri apima visų duomenų ir bendrojo vidurkio skirtumus, yra skaidoma į dviejų komponenčių – vidinės kvadratų sumos *SSW* (*sum of squares within*), įvertinančios kiekvienos imties duomenų sklaidą apie imties vidurkį, ir grupių kvadratų sumos *SSB* (*sum of squares between*), įvertinančios imčių vidurkių ir bendrojo vidurkio skirtumus, sumą:

$$SST = SSW + SSB. \qquad (7.7)$$

Nulinei hipotezei $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$ (visų populiacijų vidurkiai lygūs) tikrinti taikomas *F* kriterijus:

$$F = \frac{MSB}{MSW}, \qquad (7.8)$$

čia: $MSW = \frac{SSW}{N-k}$,

$MSB = \frac{SSB}{k-1}$,

N – bendras visų stebėjimų skaičius,

k – imčių skaičius.

Skirtumai (N-k) ir (k-1) vadinami laisvės laipsniais.

Jei gauta F reikšmė daug didesnė už vienetą, tai tikėtina, kad vidurkiai skiriasi, jei artima vienetui – vidurkių skirtumai maži. Kai nulinė hipotezė teisinga, tai kriterijus F turi Fišerio skirstinį su (k-1) ir (N-k) laisvės laipsniais. SPSS paketu apskaičiuotą F reikšmę atitinkantis stebimasis reikšmingumo lygmuo *p-level* pateikiamas rezultatų lentelėje. Jei stebimasis reikšmingumo lygmuo *p-level* mažesnis už pasirinktą reikšmingumo lygmenį α, hipotezė apie vidurkių lygybę atmetama.

ANOVA procedūra gauta F kriterijaus reikšmė rodo, ar imčių vidurkiai yra lygūs, tačiau negali pasakyti, kurios imties vidurkis skiriasi nuo kitų. Nulinė hipotezė apie vidurkių lygybę atmetama, jei *bet kurie du* vidurkiai yra nelygūs. Kad nustatytume, kurie vidurkiai reikšmingai skiriasi nuo kitų, taikomos *daugkartinio lyginimo procedūros* (*post hoc* kriterijai). Daugkartinio lyginimo procedūros apsaugo nuo klaidingų vidurkių skirtumų nustatymo. Daugkartinio lyginimo procedūrų yra gana daug. Skirtingos procedūros taikomos esant lygioms ir nelygioms dispersijoms.

Fišerio LSD kriterijus (*Fischer LSD test*) – liberaliausias *post hoc* kriterijus, t. y. jis dažniausiai randa statistiškai reikšmingus vidurkių skirtumus. Po to, kai ANOVA nustato reikšmingą vidurkių skirtumą, visos imčių poros lyginamos taikant t kriterijų.

Pagal *Bonferonio* (*Bonferroni*) kriterijų visos imčių poros lyginamos taikant Stjudento kriterijų, esant reikšmingumo lygmeniui $\alpha = \alpha_E/C$; čia α_E – eksperimento reikšmingumo lygmuo (t. y. tikimybė lyginant visas įmanomas poras nors kartą neteisingai nustatyti statistiškai reikšmingą dviejų imčių vidurkių skirtumą), α – reikšmingumo lygmuo lygi-

nant vieną imčių porą, $C = k(k-1)/2$, k – imčių skaičius. *Bonferonio* kriterijus netaikytinas, kai imčių yra daug, nes labai sumažėja α ir nefiksuojamas statistiškai reikšmingas vidurkių skirtumas, nors tikrieji populiacijų vidurkiai ir skiriasi (t. y. labai padidėja antrosios rūšies klaidos tikimybė).

Tjukio HSD kriterijus (*Tukey HSD test*) siūlo lyginant *i*-tosios ir *j*-tosios imčių vidurkius skaičiuoti statistiką:

$$Q(i,j) = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{MSW/n}}. \qquad (7.9)$$

Vidurkiai $\overline{x_i}$ ir $\overline{x_j}$ reikšmingai skiriasi, jei

$$\left|\mathrm{Q(i,j)}\right| > \mathrm{Q}_{\alpha}(\mathrm{nk-k,k}).$$

Čia: $Q_{\alpha}(nk-k,k)$ – Q statistikos α lygmens kritinė reikšmė,
k – imčių skaičius.

Visų imčių dydžiai vienodi ir lygūs n. Šis kriterijus taikomas tada, kai imčių daug. *Tjukio HSD* kriterijus yra labai konservatyvus, t. y. labiausiai nelinkęs atmesti nulinės hipotezės.

Stjudento–Njumeno–Keulso (*SNK*) *kriterijus* (*Student–Newman–Keuls test*). Šis kriterijus skiriasi nuo *Tjukio HSD* kriterijaus tuo, kad atsižvelgia į skaičių imčių, kurių vidurkiai yra tarp $\overline{x_i}$ ir $\overline{x_j}$. Šis kriterijus, kaip ir *LSD* kriterijus, statistiškai reikšmingus vidurkių skirtumus linkęs rasti ir tada, kai populiacijų vidurkiai nesiskiria. Kai kurie statistikai siūlo jį taikyti tik tuo atveju, kai yra trys imtys.

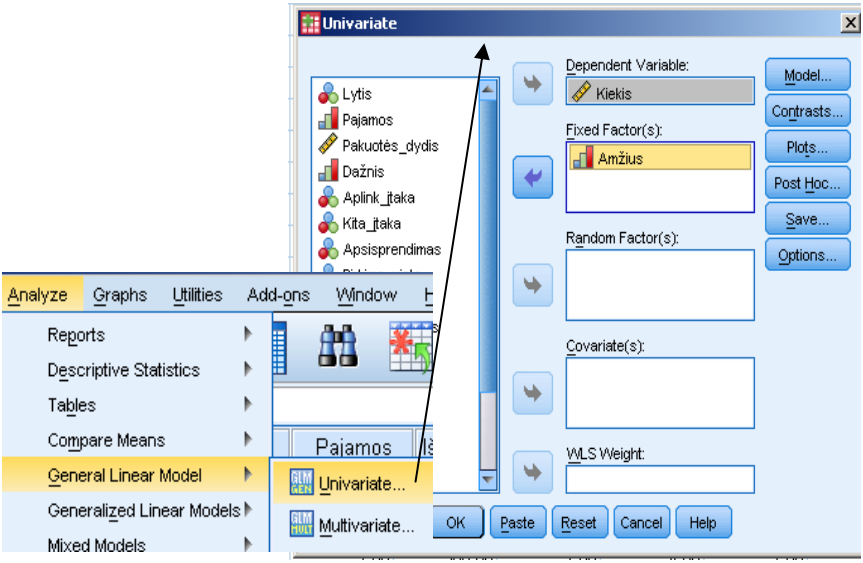
Šefė (*Scheffe*) *kriterijus*. Šis kriterijus yra konservatyvus ir statistiškai reikšmingam vidurkių skirtumo pripažinimui reikalauja didesnio vidurkių skirtumo nei kiti kriterijai.

Daugelis kitų *post hoc* kriterijų yra panašūs į aptartuosius. Dažnai tai įvairios šių kriterijų atmainos.

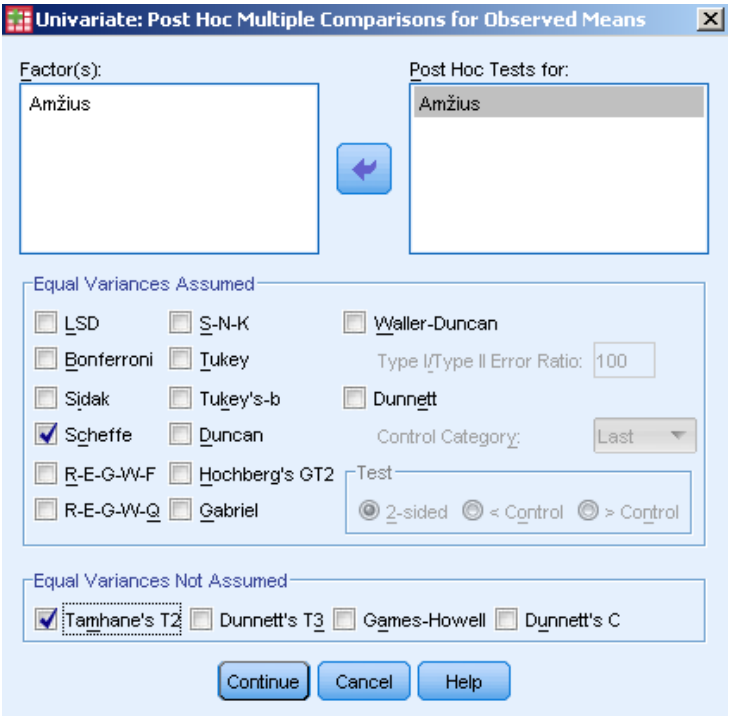
7.5.1. Vienfaktorinė dispersinė analizė SPSS terpėje

Vienfaktorinę dispersinę analizę galima atlikti taikant procedūrą *Analyze → Compare Means → One-Way ANOVA* arba apibendrintą tiesinį modelį *Analyze → General Linear Model→ Univariate...* Antrasis būdas yra tinkamesnis.

Pavyzdžiui, išnagrinėsime amžiaus poveikį perkamos kavos kiekiui. Respondentai pagal amžių suskirstyti į keturias grupes: 15–25 m., 26–35 m., 36–45 m. ir daugiau kaip 45 m. Reikia nustatyti, ar vienodą kavos kiekį jie perka. Dispersinės analizės dialogo langas atveriamas komanda *Analyze → General Linear Model → Univariate...* (žr. 7.13 pav.). Priklausomąjį kintamąjį *Kiekis* perkeliame į langelį *Dependent Variable*, o kintamąjį *Amžius* – į nepriklausomųjų kintamųjų langelį *Fixed Factor*(*s*). Jeigu dialogo lange *Univariate* nenurodyti papildomi parametrai, tai gautas rezultatas suteiks galimybę tik įvertinti kintamojo *Kiekis* vidurkių lygybę esant skirtingoms amžiaus grupėms. Nustatyti, kurios grupės skiriasi tarpusavyje, galima taikant daugkartinio lyginimo (aposteriorinius *post hoc*) kriterijus.

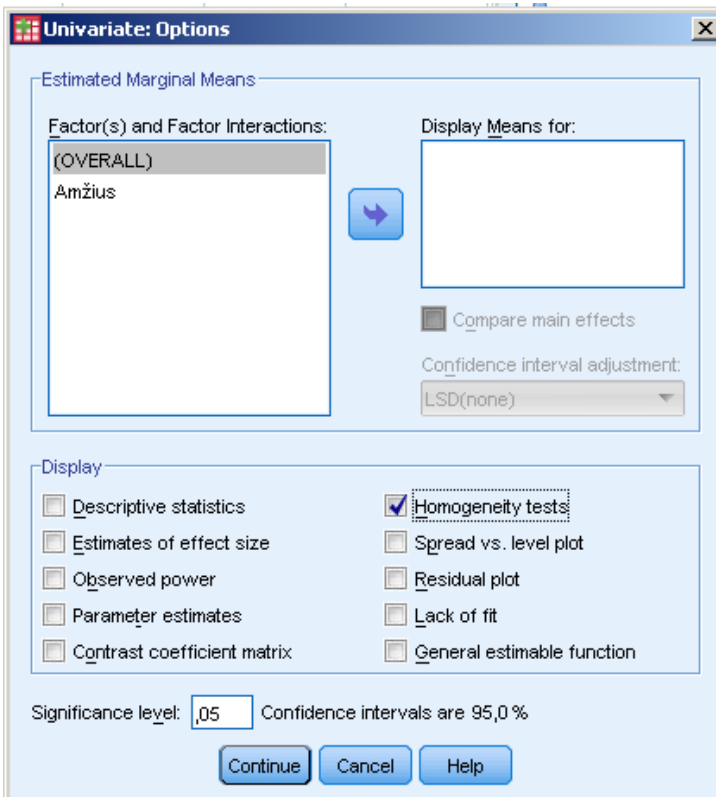


7.13 pav. **Dialogo langas *Univariate***



7.14 pav. **Dialogo langas *Univariate: Post Hoc Multiple Comparisons for Observed Means***

Paspaudus mygtuką *Post Hoc* atveriamas dialogo langas *Univariate: Post Hoc Multiple Comparisons for Observed Means* (žr. 7.14 pav.). Perkeliame nepriklausomąjį kintamąjį *Amžius* iš langelio *Factor*(*s*) į langelį *Post Hoc Tests for.* SPSS taiko skirtingus vidurkių lygybės nustatymo kriterijus esant lygiai ir nelygiai kintamųjų dispersijai. Bendruoju atveju dispersijų lygybė iš anksto nežinoma, todėl rekomenduojama nurodyti abiejų tipų kriterijų nustatymą. Populiariausi yra *Scheffe* kriterijus dėl lygių dispersijų ir *Tamhane's T2* kriterijus dėl nelygių dispersijų. Reikia pažymėti atitinkamus langelius. Paspaudę mygtuką *Continue* grįžtame į pradinį langą.



7.15 pav. **Dialogo langas *Univariate: Options***

Dispersijų lygybė nustatoma *Livyno* kriterijumi. Dispersijų tikrinimo parametrai užduodami lange *Univariate: Options*, kuris atsiveria paspaudus mygtuką *Options* (žr. 7.15 pav.). Tikrinant *Livyno* kriterijų reikia pažymėti langelį *Homogeneity tests.* Paspaudę mygtuką *Continue* grįžtame į pradinį langą.

Pastaba: jei nepriklausomasis kintamasis turi tik 2 kategorijas, *post hoc* kriterijai netaikomi. Kintamųjų vidurkių skirtumams nustatyti reikia perkelti analizuojamą priklausomąjį kintamąjį iš langelio *Factor*(*s*) *and Factor Interactions* į langelį *Display Means for.* Analizuojamu atveju tai nereikalinga, nes nepriklausomasis kintamasis turi daugiau kaip dvi kategorijas. Paspaudę mygtuką *Continue* vėl grįžtame į pradinį langą. Pro-

cedūra baigiama mygtuko *OK* paspaudimu. Lange *Output-SPSS Viewer* rodomi analizės rezultatai (žr. 7.16 pav.).

Levene's Test of Equality of Error Variances[a]

Dependent Variable:Kiekis

F	df1	df2	Sig.
1,997	3	56	,125

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + Amžius

Tests of Between-Subjects Effects

Dependent Variable:Kiekis

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	965003,511[a]	3	321667,837	2,134	,106
Intercept	3788413,858	1	3788413,858	25,139	,000
Amžius	965003,511	3	321667,837	2,134	,106
Error	8439281,489	56	150701,455		
Total	2,355E7	60			
Corrected Total	9404285,000	59			

7.16 pav. **Vienfaktorinės dispersinės analizės rezultatas**

Lentelėje *Levene's Test of Equality of Error Variances p-level* (stulpelis *Sig.*) rodo *F* kriterijaus reikšmingumą. Jei *p-level* > $\alpha = 0,05$ (*F* kriterijus nereikšmingas), dispersijos lygios, skirtumų analizei toliau taikomas *Scheffe* kriterijus, priešingu atveju – *Tamhane's T2* kriterijus. Nepriklausomųjų kintamųjų grupių skirtumų reikšmingumą rodo nepriklausomojo kintamojo *Amžius F* kriterijaus *p-level* (žr. stulpelį *Sig.*). Jei *Livyno* kriterijus parodė dispersijų skirtumą, tai patikimumo lygmenį α reikia padidinti nuo *0,05* iki *0,01.* Nagrinėjamu atveju dispersijos lygios, *p-level* = *0,106* > $\alpha = 0,05$, skirtumas nereikšmingas.

Lentelė *Post Hoc Tests*, *Multiple Comparisons* (žr. 7.17 pav.) leidžia patikrinti, kaip viena grupė skiriasi nuo kitos ir į kokią pusę. Reikia

žiūrėti rezultatus, kurie atitinka *Livyno* testu nustatytą kriterijų (nagrinėjamu atveju – *Scheffe* kriterijus). *Mean Difference* stulpelio reikšmės rodo grupių vidurkių skirtumą, o *p-level* (*Sig*. stulpelis) reikšmė – šių skirtumų reikšmingumą. Nagrinėjamu atveju skirtumai nereikšmingi.

Post Hoc Tests

Amžius

Multiple Comparisons

Dependent Variable:Kiekis

	(I) Amžius	(J) Amžius	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Scheffe	15-25 m.	26-35 m.	-110,8857	129,58568	,865	-484,4049	262,6335
		36-45 m.	-230,0211	118,15123	,296	-570,5815	110,5394
		virš 45 m.	348,4000	285,26967	,686	-473,8644	1170,6644
	26-35 m.	15-25 m.	110,8857	129,58568	,865	-262,6335	484,4049
		36-45 m.	-119,1353	136,73348	,859	-513,2574	274,9867
		virš 45 m.	459,2857	293,45377	,490	-386,5687	1305,1401
	36-45 m.	15-25 m.	230,0211	118,15123	,296	-110,5394	570,5815
		26-35 m.	119,1353	136,73348	,859	-274,9867	513,2574
		virš 45 m.	578,4211	288,58687	,271	-253,4049	1410,2470
	virš 45 m.	15-25 m.	-348,4000	285,26967	,686	-1170,6644	473,8644
		26-35 m.	-459,2857	293,45377	,490	-1305,1401	386,5687
		36-45 m.	-578,4211	288,58687	,271	-1410,2470	253,4049
Tamhane	15-25 m.	26-35 m.	-110,8857	145,63796	,974	-540,6070	318,8356
		36-45 m.	-230,0211	112,37726	,262	-545,9867	85,9446
		virš 45 m.	348,4000*	58,48840	,000	180,7713	516,0287
	26-35 m.	15-25 m.	110,8857	145,63796	,974	-318,8356	540,6070
		36-45 m.	-119,1353	164,30848	,979	-588,2456	349,9749
		virš 45 m.	459,2857*	133,37736	,026	46,3526	872,2189
	36-45 m.	15-25 m.	230,0211	112,37726	,262	-85,9446	545,9867
		26-35 m.	119,1353	164,30848	,979	-349,9749	588,2456
		virš 45 m.	578,4211*	95,95705	,000	295,0668	861,7754
	virš 45 m.	15-25 m.	-348,4000*	58,48840	,000	-516,0287	-180,7713
		26-35 m.	-459,2857*	133,37736	,026	-872,2189	-46,3526
		36-45 m.	-578,4211*	95,95705	,000	-861,7754	-295,0668

Based on observed means.
The error term is Mean Square(Error) = 150701,455.

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

7.17 pav. **Grupių lygybės tikrinimas**

Lentelėje *Homogeneous Subsets* (žr. 7.18 pav.) pateikti nepriklausomojo kintamojo grupių skirtumai. Išskiriamos grupių kategorijos (*Subset*). Nagrinėjamu atveju visos amžiaus grupės priklauso vienai kategorijai.

Homogeneous Subsets

Kiekis

	Amžius	N	Subset 1
Scheffe[a,b,c]	Daugiau kaip 45 m.	2	50,0000
	15–25 m.	25	398,4000
	26–35 m.	14	509,2857
	36–45 m.	19	628,4211
	Sig.		,095

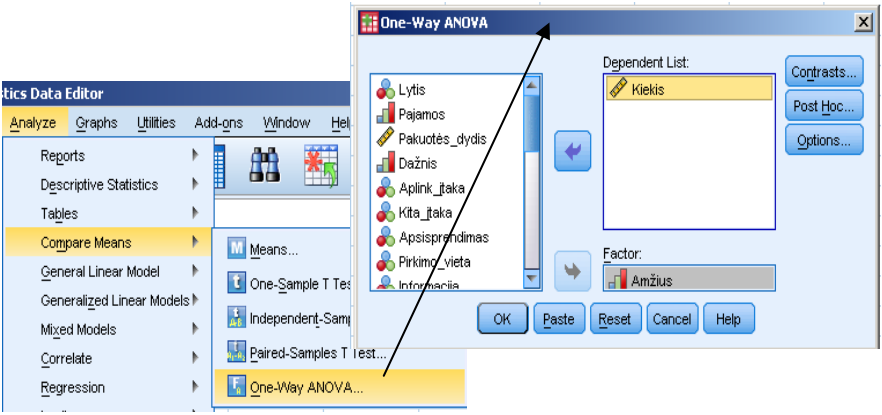
Means for groups in homogeneous subsets are displayed.
Based on observed means.
The error term is Mean Square(Error) = 150701,455.
a. Uses Harmonic Mean Sample Size = 6,024.
b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

7.18 pav. **Grupių lygybės tikrinimas**

Atliktos procedūros sintaksė:

```
UNIANOVA
  Kiekis BY Amžius
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /POSTHOC = Amžius (SCHEFFE T2)
  /EMMEANS = TABLES(Amžius)
  /PRINT = HOMOGENEITY
  /CRITERIA = ALPHA(.05)
  /DESIGN = Amžius .
```

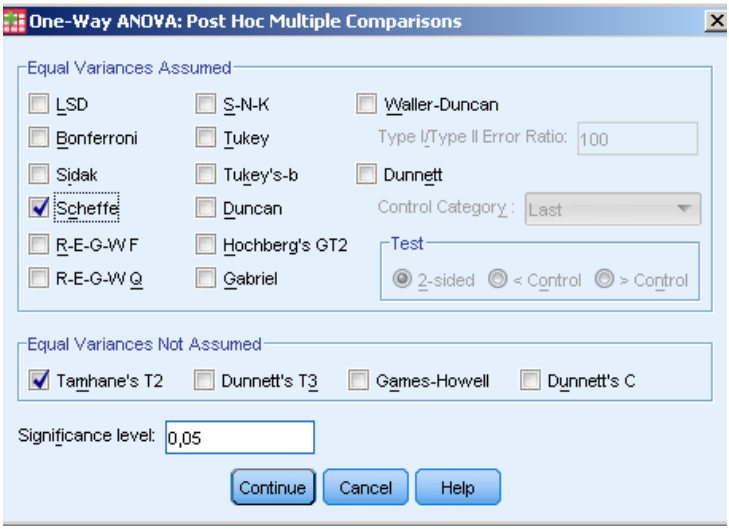
Sprendžiant tą pačią užduotį galima pritaikyti komandą *Analyze → Compare Means → One-Way ANOVA*. Dispersinės analizės dialogo langas atveriamas komanda *Analyze → Compare Means → One-Way ANOVA…* (žr. 7.19 pav.). Priklausomąjį kintamąjį *Kiekis* perkeliame į langelį *Dependent List*, o kintamąjį *Amžius* – į nepriklausomųjų kintamųjų langelį *Factor*.



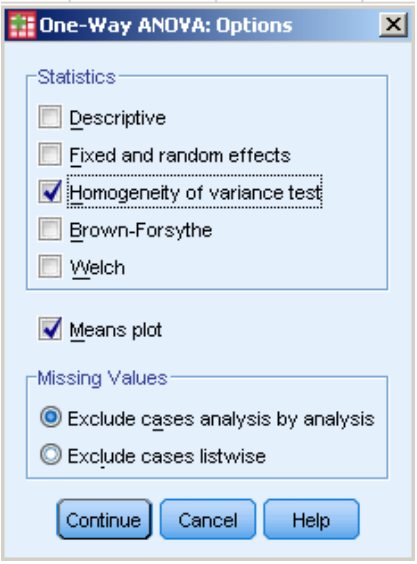
7.19 pav. **Dialogo langas *One-Way ANOWA***

Nustatyti, kurios grupės skiriasi tarpusavyje, galima taikant daugkartinio lyginimo (aposteriorinius *post hoc*) kriterijus. Paspaudus mygtuką *Post Hoc* atveriamas dialogo langas *One-Way ANOVA: Post Hoc Multiple Comparisons* (žr. 7.20 pav.). Kriterijų pasirinkimas analogiškas 7.14 pav. Paspaudę mygtuką *Continue* grįžtame į pradinį langą.

Dispersijų lygybė nustatoma *Livyno* kriterijumi. Dispersijų tikrinimo parametrai užduodami lange *One-Way ANOVA: Options*, kuris atsiveria paspaudus mygtuką *Options* (žr. 7.21 pav.). *Livyno* kriterijui tikrinti reikia pažymėti langelį *Homogeneity of variance tests.* Tame pačiame lange reikia pažymėti ir langelį *Means plot*, kad būtų parodytas vidurkių grafikas.



7.20 pav. **Dialogo langas *One-Way ANOVA: Post Hoc Multiple Comparisons***



7.21 pav. **Dialogo langas *One-Way ANOVA: Options***

Paspaudę mygtuką *Continue* grįžtame į pradinį langą. Procedūra baigiama mygtuko *OK* paspaudimu. Lange *Output-SPSS Viewer* rodomi analizės rezultatai (žr. 7.22 pav.).

Test of Homogeneity of Variances

Kiekis

Levene Statistic	df1	df2	Sig.
1,997	3	56	,125

ANOVA

Kiekis

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	965003,511	3	321667,837	2,134	,106
Within Groups	8439281,489	56	150701,455		
Total	9404285,000	59			

7.22 pav. **Vienfaktorinės dispersinės analizės rezultatas**

Lentelėje *Test of Homogeneity of Variances p-level* (stulpelis *Sig.*) rodo Livyno kriterijaus (*Levene Statistic*) reikšmingumą. Jei *p-level* > α = *0,05* (kriterijus nereikšmingas), dispersijos lygios, skirtumų analizei toliau taikomas *Scheffe* kriterijus, priešingu atveju – *Tamhane's T2* kriterijus. Nepriklausomųjų kintamųjų grupių skirtumų reikšmingumą rodo *F* kriterijaus *p-level* (žr. stulpelį *Sig.*) lentelėje *ANOVA*. Jei *Livyno* kriterijus parodė dispersijų skirtumą, tai patikimumo lygmenį α reikia padidinti nuo *0,05* iki *0,01*. Nagrinėjamu atveju dispersijos lygios, *p-level* = *0,106* > α = *0,05*, skirtumas nereikšmingas.

Lentelė *Post Hoc Tests*, *Multiple Comparisons* (žr. 7.23 pav.) leidžia patikrinti, kaip viena grupė skiriasi nuo kitos ir į kokią pusę. Reikia žiūrėti rezultatus, kurie atitinka *Livyno* testu nustatytą kriterijų (nagrinėjamu atveju – *Scheffe* kriterijus). *Mean Difference* stulpelio reikšmės rodo grupių vidurkių skirtumą, o *p-level* (*Sig.* stulpelis) reikšmė – šių skirtumų reikšmingumą. Nagrinėjamu atveju skirtumai nereikšmingi. Lentelė analogiška 7.17 pav.

Post Hoc Tests

Multiple Comparisons

Dependent Variable:Kiekis

	(I) Amžius	(J) Amžius	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Scheffe	15-25 m.	26-35 m.	-110,88571	129,58568	,865	-484,4049	262,6335
		36-45 m.	-230,02105	118,15123	,296	-570,5815	110,5394
		virš 45 m.	348,40000	285,26967	,686	-473,8644	1170,6644
	26-35 m.	15-25 m.	110,88571	129,58568	,865	-262,6335	484,4049
		36-45 m.	-119,13534	136,73348	,859	-513,2574	274,9867
		virš 45 m.	459,28571	293,45377	,490	-386,5687	1305,1401
	36-45 m.	15-25 m.	230,02105	118,15123	,296	-110,5394	570,5815
		26-35 m.	119,13534	136,73348	,859	-274,9867	513,2574
		virš 45 m.	578,42105	288,58687	,271	-253,4049	1410,2470
	virš 45 m.	15-25 m.	-348,40000	285,26967	,686	-1170,6644	473,8644
		26-35 m.	-459,28571	293,45377	,490	-1305,1401	386,5687
		36-45 m.	-578,42105	288,58687	,271	-1410,2470	253,4049
Tamhane	15-25 m.	26-35 m.	-110,88571	145,63796	,974	-540,6070	318,8356
		36-45 m.	-230,02105	112,37726	,262	-545,9867	85,9446
		virš 45 m.	348,40000*	58,48840	,000	180,7713	516,0287
	26-35 m.	15-25 m.	110,88571	145,63796	,974	-318,8356	540,6070
		36-45 m.	-119,13534	164,30848	,979	-588,2456	349,9749
		virš 45 m.	459,28571*	133,37736	,026	46,3526	872,2189
	36-45 m.	15-25 m.	230,02105	112,37726	,262	-85,9446	545,9867
		26-35 m.	119,13534	164,30848	,979	-349,9749	588,2456
		virš 45 m.	578,42105*	95,95705	,000	295,0668	861,7754
	virš 45 m.	15-25 m.	-348,40000*	58,48840	,000	-516,0287	-180,7713
		26-35 m.	-459,28571*	133,37736	,026	-872,2189	-46,3526
		36-45 m.	-578,42105*	95,95705	,000	-861,7754	-295,0668

*. The mean difference is significant at the 0.05 level.

7.23 pav. **Grupių lygybės tikrinimas**

Lentelėje *Homogeneous Subsets* (žr. 7.24 pav.) pateikti nepriklausomojo kintamojo grupių skirtumai. Išskiriamos grupių kategorijos (*Subset*). Nagrinėjamu atveju visos amžiaus grupės priklauso vienai kategorijai. Pateikiama grupių vidurkių diagrama (*Means Plots*) (žr. 7.25 pav.). Ji rodo, kokį kavos kiekį vidutiniškai perka kiekvienos amžiaus grupės atstovai.

Homogeneous Subsets

Kiekis

	Amžius	N	Subset for alpha = 0.05
			1
Scheffe[a,b]	virš 45 m.	2	50,0000
	15-25 m.	25	398,4000
	26-35 m.	14	509,2857
	36-45 m.	19	628,4211
	Sig.		,095

Means for groups in homogeneous subsets are displayed.

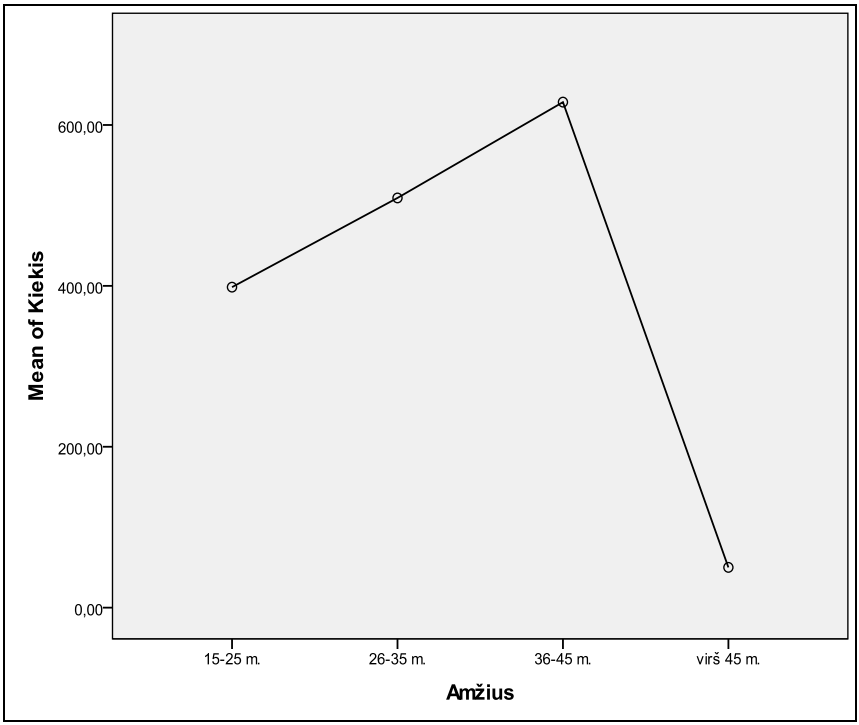
a. Uses Harmonic Mean Sample Size = 6,024.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

7.24 pav. **Grupių lygybės tikrinimas**

Atliktos procedūros sintaksė:

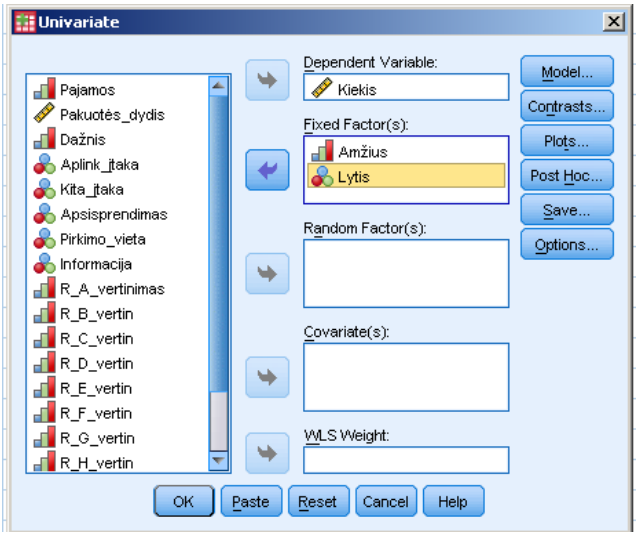
```
ONEWAY Kiekis BY Amžius
  /STATISTICS HOMOGENEITY
  /PLOT MEANS
  /MISSING ANALYSIS
  /POSTHOC=SCHEFFE T2 ALPHA(0.05).
```



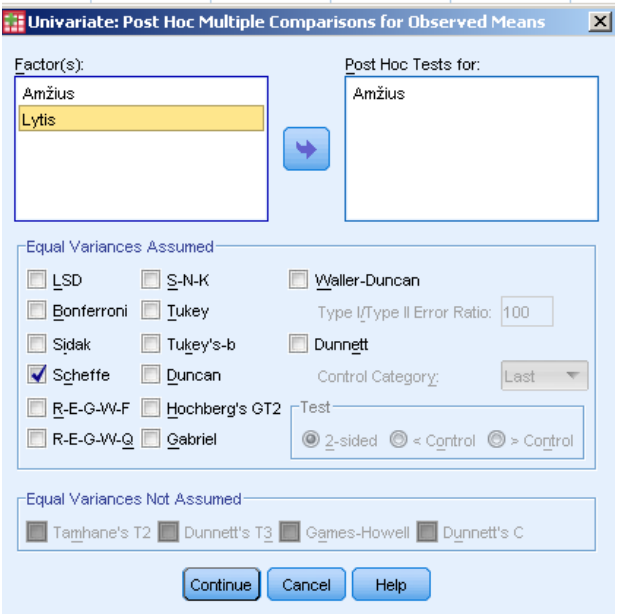
7.25 pav. **Grupių vidurkių diagrama**

7.5.2. Dviejų veiksnių dispersinė analizė SPSS terpėje

Kaip pavyzdį išnagrinėsime amžiaus ir lyties poveikį perkamos kavos kiekiui. Respondentai pagal amžių suskirstyti į keturias grupes: 15–25 m., 26–35 m., 36–45 m. ir daugiau kaip 45 m.; pagal lytį – į dvi grupes: vyrai ir moterys. Reikia nustatyti, ar vienodą kavos kiekį jie perka. Dispersinės analizės dialogo langas atveriamas komanda *Analyze → General Linear Model → Univariate* (žr. 7.26 pav.). Į langelį *Fixed Factor*(*s*) reikia papildomai perkelti kintamąjį *Lytis* (žr. 7.26 pav.). Paspaudus mygtuką *Post Hoc* atveriamas dialogo langas *Univariate: Post Hoc Multiple Comparisons for Observed Means* (žr. 7.27 pav.).

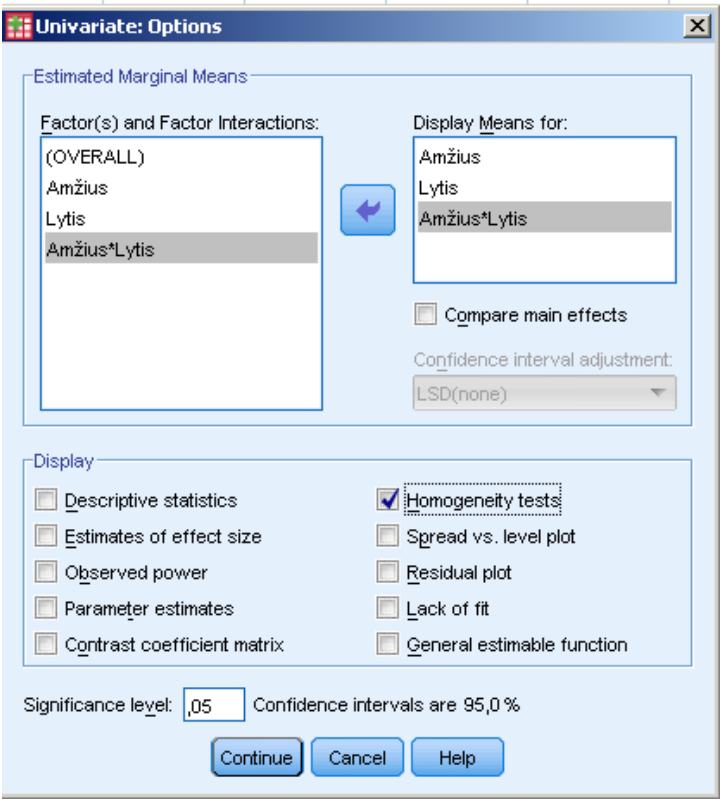


7.26 pav. **Dialogo langas *Univariate***

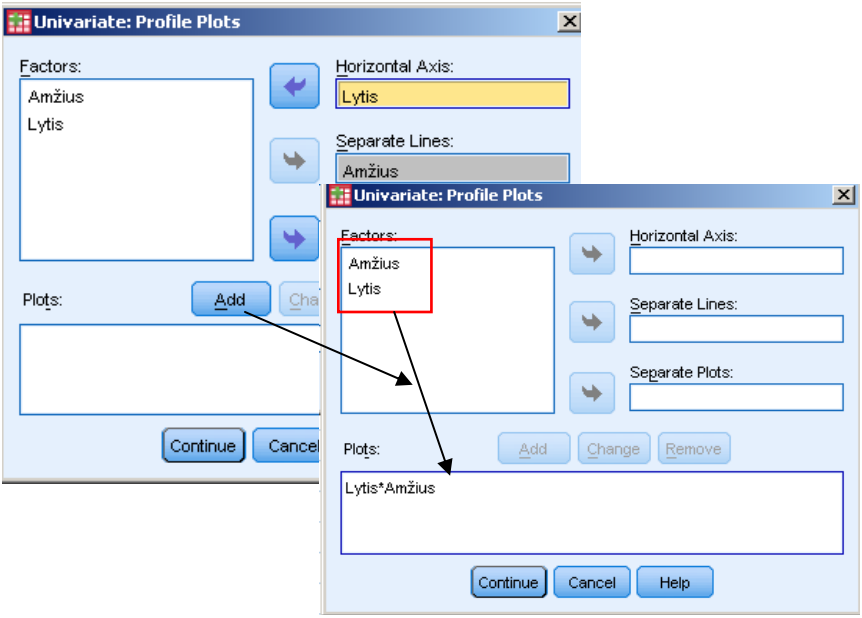


7.27 pav. **Dialogo langas *Univariate: Post Hoc Multiple Comparisons for Observed Means***

Į langelį *Post Hoc Tests for* reikia perkelti visus nepriklausomuosius kintamuosius, išskyrus kintamąjį *Lytis.* Nepriklausomąjį kintamąjį *Amžius* iš langelio *Factor*(*s*) reikia perkelti į langelį *Post Hoc Tests for. Post hoc* kriterijus pasirenkame kaip ir vienfaktorinės dispersijos atveju. Dialogo lange *Univariate: Options* (žr. 7.28 pav.) analizuojamus priklausomuosius kintamuosius *Amžius*, *Lytis* bei jų sąveiką *Amžius*Lytis* reikia perkelti iš langelio *Factor*(*s*) *and Factor Interactions* į langelį *Display Means for*. Paspaudę mygtuką *Continue* grįžtame į pradinį langą.



7.28 pav. **Dialogo langas *Univariate: Options***



7.29 pav. **Dialogo langas *Univariate: Profile Plots***

Mygtuku *Plots* atveriamas dialogo langas *Univariate: Profile Plots* (žr. 7.29 pav.), kuriame galima nurodyti grafikų parametrus.

Dialogo lange *Univariate: Profile Plots* nepriklausomąjį kintamąjį *Lytis* perkeliame į langelį *Horizontal Axis*, o kitą nepriklausomąjį kintamąjį *Amžius* – į langelį *Separate Lines.* Tai leidžia nubrėžti keletą linijų. Mygtuko *Add* paspaudimu patvirtinamas pasirinkimas. Paspaudę mygtuką *Continue* grįžtame į pradinį langą. Procedūra baigiama mygtuko *OK* paspaudimu. Lange *Output-SPSS Viewer* rodomi analizės rezultatai. Lentelės *Levene's Test of Equality of Error Variances*, *Tests of Between-Subjects Effects*, *Post Hoc Tests*, *Multiple Comparisons* ir *Homogeneous Subsets* analizuojamos taip pat, kaip dėl vienfaktorinės dispersinės analizės, tik lentelėje *Tests of Between-Subjects Effects* (žr. 7.30 pav.) rodomas vidurkių skirtumas dėl kintamųjų *Amžius*, *Lytis* bei jų sąveikos *Amžius*Lytis* grupių.

Papildomai rodoma lentelė *Estimated Marginal Means*, kurioje pateikti visų grupių priklausomojo kintamojo vidurkiai (žr. 7.31 pav.). Išvedamas kintamųjų grupių vidurkių grafikas (žr. 7.32 pav.). Informaciją galima taikyti rinkai segmentuoti.

Levene's Test of Equality of Error Variances[a]

Dependent Variable:Kiekis

F	df1	df2	Sig.
2,177	7	52	,051

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + Amžius + Lytis + Amžius * Lytis

Tests of Between-Subjects Effects

Dependent Variable:Kiekis

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1,371E6	7	195824,908	1,268	,285
Intercept	3413470,825	1	3413470,825	22,095	,000
Amžius	967090,489	3	322363,496	2,087	,113
Lytis	6965,857	1	6965,857	,045	,833
Amžius * Lytis	404516,481	3	134838,827	,873	,461
Error	8033510,641	52	154490,589		
Total	2,355E7	60			
Corrected Total	9404285,000	59			

a. R Squared = ,146 (Adjusted R Squared = ,031)

7.30 pav. **Dispersinės analizės rezultatas**

Atliktos procedūros sintaksė:

```
UNIANOVA
  Kiekis BY Amžius Lytis
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /POSTHOC = Amžius (SCHEFFE T2)
  /EMMEANS = TABLES(Amžius)
  /EMMEANS = TABLES(Lytis)
  /EMMEANS = TABLES(Amžius*Lytis)
  /PRINT = HOMOGENEITY
  /CRITERIA = ALPHA(.05)
  /DESIGN = Amžius Lytis Amžius*Lytis .
```

Estimated Marginal Means

1. Amžius

Dependent Variable:Kiekis

Amžius	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
15-25 m.	394,968	78,674	237,098	552,838
26-35 m.	450,250	116,267	216,944	683,556
36-45 m.	632,917	93,467	445,361	820,472
virš 45 m.	50,000	277,930	-507,708	607,708

2. Lytis

Dependent Variable:Kiekis

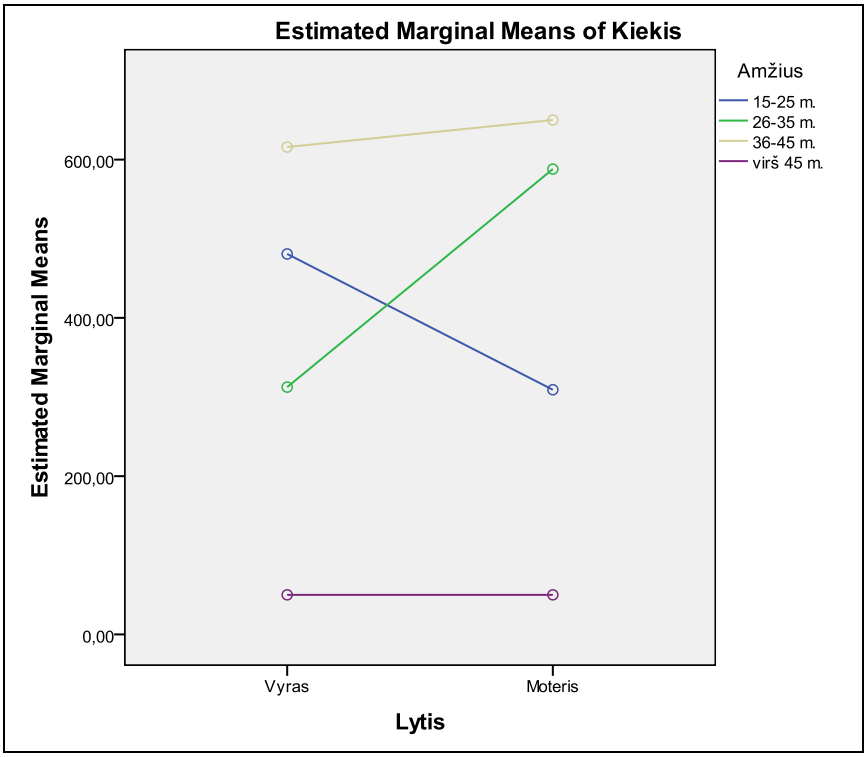
Lytis	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Vyras	364,776	116,692	130,617	598,935
Moteris	399,292	113,160	172,219	626,364

3. Amžius * Lytis

Dependent Variable:Kiekis

Amžius	Lytis	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
15-25 m.	Vyras	480,769	109,013	262,018	699,520
	Moteris	309,167	113,465	81,483	536,850
26-35 m.	Vyras	312,500	196,526	-81,859	706,859
	Moteris	588,000	124,294	338,585	837,415
36-45 m.	Vyras	615,833	113,465	388,150	843,517
	Moteris	650,000	148,560	351,892	948,108
virš 45 m.	Vyras	50,000	393,053	-738,718	838,718
	Moteris	50,000	393,053	-738,718	838,718

7.31 pav. **Dispersinės analizės rezultatas. Vidurkių skaičiavimas**



7.32 pav. **Kintamųjų grupių vidurkių grafikas *Profile Plots***

KONTROLINIAI KLAUSIMAI

7.1. Išvardykite pagrindinius *t* kriterijaus tipus.
7.2. Kokie kintamieji analizuojami *t* kriterijaus metodu?
7.3. Koks kriterijus nustato imčių dispersijų lygybę?
7.4. Kokiais atvejais taikoma dispersinė analizė?
7.5. Kokios dispersinės analizės prielaidos?
7.6. Kaip dispersinės analizės atlikimo metu parenkamas priklausomasis ir nepriklausomasis kintamasis?
7.7. Kokie kintamieji vadinami veiksniais?
7.8. Kam reikalingos daugkartinio lyginimo procedūros (*post hoc* kriterijai)?

UŽDUOTYS

7.1 užduotis

Prekybos centre akcijos lipdukas duodamas už pirkiniams išleistą 30 Lt pinigų sumą. Atlikta analizė, norint nustatyti, ar iš tikrųjų per vieną kartą išleidžiamų pinigų vidurkis reikšmingai skiriasi nuo šios sumos. Skaičiavimo rezultatas pateiktas 7.33 pav. Paaiškinkite gautą rezultatą.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Pinigai	30	32,0000	26,83282	4,89898

One-Sample Test

	Test Value = 30					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Pinigai	,408	29	,686	2,00000	-8,0195	12,0195

7.33 pav. ***T* kriterijaus skaičiavimo rezultatas**

7.2 užduotis

Atlikta analizė, ar skirtingo amžiaus respondentai per vieną kartą pirkiniams išleidžia vienodai pinigų. Analizės rezultatas pateiktas 7.34 pav. Paaiškinkite gautą rezultatą.

Group Statistics

	Amžius	N	Mean	Std. Deviation	Std. Error Mean
Pinigai	>= 35,00	42	33,5714	23,45951	3,61988
	< 35,00	48	30,6250	29,12966	4,20450

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Pinigai	Equal variances assumed	,906	,344	,523	88	,602	2,94643	5,62841	-8,23884	14,13170
	Equal variances not assumed			,531	87,432	,597	2,94643	5,54810	-8,08025	13,97311

7.34 pav. ***T* kriterijaus skaičiavimo rezultatas**

7.3 užduotis

Nagrinėjant amžiaus poveikį pinigų kiekiui, išleidžiamų per mėnesį šampūnui pirkti, gauti rezultatai pateikti 7.35 ir 7.36 pav. Paaiškinkite gautus rezultatus.

Levene's Test of Equality of Error Variances[a]

Dependent Variable: Pinigai

F	df1	df2	Sig.
2,456	4	31	,066

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept+Amzius

Tests of Between-Subjects Effects

Dependent Variable: Pinigai

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	115,280[a]	4	28,820	3,427	,020
Intercept	2176,211	1	2176,211	258,755	,000
Amzius	115,280	4	28,820	3,427	,020
Error	260,720	31	8,410		
Total	3292,000	36			
Corrected Total	376,000	35			

a. R Squared = ,307 (Adjusted R Squared = ,217)

7.35 pav. **Vienfaktorinės dispersinės analizės rezultatas**

Multiple Comparisons

Dependent Variable: Pinigai

	(I) Amzius	(J) Amzius	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Scheffe	maziau 31 m.	31-40 m.	3,4524	1,37925	,208	-1,0624	7,9671
		41-50 m.	3,1667	1,54367	,397	-1,8863	8,2196
		51-60 m.	2,0417	1,32369	,669	-2,2912	6,3745
		61-70 m.	5,4167	1,67435	,054	-,0640	10,8974
	31-40 m.	maziau 31 m.	-3,4524	1,37925	,208	-7,9671	1,0624
		41-50 m.	-,2857	1,69810	1,000	-5,8441	5,2727
		51-60 m.	-1,4107	1,50092	,925	-6,3237	3,5023
		61-70 m.	1,9643	1,81771	,881	-3,9857	7,9142
	41-50 m.	maziau 31 m.	-3,1667	1,54367	,397	-8,2196	1,8863
		31-40 m.	,2857	1,69810	1,000	-5,2727	5,8441
		51-60 m.	-1,1250	1,65329	,976	-6,5367	4,2867
		61-70 m.	2,2500	1,94542	,853	-4,1180	8,6180
	51-60 m.	maziau 31 m.	-2,0417	1,32369	,669	-6,3745	2,2912
		31-40 m.	1,4107	1,50092	,925	-3,5023	6,3237
		41-50 m.	1,1250	1,65329	,976	-4,2867	6,5367
		61-70 m.	3,3750	1,77591	,474	-2,4381	9,1881
	61-70 m.	maziau 31 m.	-5,4167	1,67435	,054	-10,8974	,0640
		31-40 m.	-1,9643	1,81771	,881	-7,9142	3,9857
		41-50 m.	-2,2500	1,94542	,853	-8,6180	4,1180
		51-60 m.	-3,3750	1,77591	,474	-9,1881	2,4381
Tamhane	maziau 31 m.	31-40 m.	3,4524	1,26178	,135	-,6198	7,5245
		41-50 m.	3,1667	1,57842	,550	-2,6909	9,0243
		51-60 m.	2,0417	1,52827	,895	-2,9781	7,0615
		61-70 m.	5,4167*	,97669	,001	2,1018	8,7316
	31-40 m.	maziau 31 m.	-3,4524	1,26178	,135	-7,5245	,6198
		41-50 m.	-,2857	1,51680	1,000	-6,2599	5,6885
		51-60 m.	-1,4107	1,46454	,987	-6,4036	3,5821
		61-70 m.	1,9643	,87360	,457	-1,5386	5,4672
	41-50 m.	maziau 31 m.	-3,1667	1,57842	,550	-9,0243	2,6909
		31-40 m.	,2857	1,51680	1,000	-5,6885	6,2599
		51-60 m.	-1,1250	1,74476	1,000	-7,3696	5,1196
		61-70 m.	2,2500	1,28938	,805	-4,5172	9,0172
	51-60 m.	maziau 31 m.	-2,0417	1,52827	,895	-7,0615	2,9781
		31-40 m.	1,4107	1,46454	,987	-3,5821	6,4036
		41-50 m.	1,1250	1,74476	1,000	-5,1196	7,3696
		61-70 m.	3,3750	1,22748	,235	-1,3987	8,1487
	61-70 m.	maziau 31 m.	-5,4167*	,97669	,001	-8,7316	-2,1018
		31-40 m.	-1,9643	,87360	,457	-5,4672	1,5386
		41-50 m.	-2,2500	1,28938	,805	-9,0172	4,5172
		51-60 m.	-3,3750	1,22748	,235	-8,1487	1,3987

Based on observed means.

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

7.36 pav. **Grupių lygybės tikrinimas**

7.4 užduotis

Nagrinėjant amžiaus ir lyties poveikį pinigų kiekiui, išleidžiamų per mėnesį šampūnui pirkti, gauti rezultatai pateikti 7.37 ir 7.38 pav. Paaiškinkite gautus rezultatus.

Levene's Test of Equality of Error Variances[a]

Dependent Variable: Pinigai

F	df1	df2	Sig.
6,149	9	26	,000

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept+Amzius+Lytis+Amzius * Lytis

Tests of Between-Subjects Effects

Dependent Variable: Pinigai

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	126,550[a]	9	14,061	1,466	,213
Intercept	1751,483	1	1751,483	182,556	,000
Amzius	102,761	4	25,690	2,678	,054
Lytis	,110	1	,110	,012	,915
Amzius * Lytis	8,995	4	2,249	,234	,916
Error	249,450	26	9,594		
Total	3292,000	36			
Corrected Total	376,000	35			

a. R Squared = ,337 (Adjusted R Squared = ,107)

7.37 pav. **Dispersinės analizės rezultatas**

Estimated Marginal Means

1. Amzius

Dependent Variable: Pinigai

Amzius	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
maziau 31 m.	11,000	,948	9,051	12,949
31-40 m.	7,800	1,296	5,137	10,463
41-50 m.	7,250	1,732	3,691	10,809
51-60 m.	9,125	1,095	6,874	11,376
61-70 m.	5,750	1,549	2,567	8,933

2. Lytis

Dependent Variable: Pinigai

Lytis	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
vyras	8,250	,955	6,288	10,212
moteris	8,120	,746	6,587	9,653

3. Amzius * Lytis

Dependent Variable: Pinigai

Amzius	Lytis	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
maziau 31 m.	vyras	11,500	1,095	9,249	13,751
	moteris	10,500	1,549	7,317	13,683
31-40 m.	vyras	8,000	2,190	3,498	12,502
	moteris	7,600	1,385	4,753	10,447
41-50 m.	vyras	6,000	3,097	-,367	12,367
	moteris	8,500	1,549	5,317	11,683
51-60 m.	vyras	9,750	1,549	6,567	12,933
	moteris	8,500	1,549	5,317	11,683
61-70 m.	vyras	6,000	2,190	1,498	10,502
	moteris	5,500	2,190	,998	10,002

7.38 pav. **Dispersinės analizės rezultatas. Vidurkių skaičiavimas**

8. FAKTORINĖ ANALIZĖ

Faktorinė analizė leidžia kintamųjų masyve išskirti keletą grupių, kurios vadinamos veiksniais. Grupavimas atliekamas apskaičiuojant kintamųjų koreliaciją. Į vieną veiksnį jungiami kintamieji, stipriai koreliuojantys tarpusavyje ir silpnai koreliuojantys arba nekoreliuojantys su kitais kintamaisiais, kurie sudaro kitus veiksnius. Taip sumažinamas kintamųjų skaičius.

Faktorinė analizė plačiai taikoma ekonomikoje, rinkodaroje ir kituose taikomuosiuose moksluose, kurie nagrinėja didelius duomenų kiekius. Ji pasižymi sudėtingais matematiniais skaičiavimais. Pateikiant faktorinės analizės aprašus, tenka įvertinti ir pagrįsti šių metodų taikymo galimybes ir ypatybes, matematinių modelių veikimo principus, o rezultatų interpretacija nėra paprasta ir vienareikšmė.

Faktorinė analizė neturi vienintelio sprendinio. Koreliacinę matricą galima pavaizduoti veiksniais ir skirtingais būdais. Be to, faktorinę analizę sudaro skirtingi modeliai, kurie paremti veiksnių svorių matricos apskaičiavimu. Dažniausiai faktorinės analizės programos iškart siūlo keletą skirtingų metodų. Pavyzdžiui, SPSS paketas siūlo šiuos metodus: pagrindinių komponenčių metodą, maksimalaus tikėtinumo metodą, bendrąją faktorinę analizę, nesvertinį mažiausiųjų kvadratų metodą, alfa metodą ir vaizdų veiksnius. Geriausio metodo pasirinkimą lemia skaičiavimo patogumas, taip pat tyrėjo nuomonė.

Faktorinės analizės metodais sprendžiami dviejų tipų uždaviniai: tiriamieji ir patvirtinamieji. Pagal šiuos uždavinius ir faktorinė analizė vadinama tiriamąja ir patvirtinamąja.

Tiriamoji faktorinė analizė taikoma tada, kai turima kintamųjų aibė ir neaišku, nei kiek gali būti veiksnių, nei kokie kintamieji juos sudaro. Ir netgi neaišku, ar apskritai kintamieji yra kolinearūs.

Patvirtinamoji faktorinė analizė taikoma tada, kai iš anksto žinomi teoriniai ar anksčiau atlikti empiriniai tyrimai ir norima jau turimas žinias patvirtinti ar išplėsti.

Tiriamoji faktorinė analizė taikoma dažniau nei patvirtinamoji. Pagrindinė tiriamosios faktorinės analizės taikymo sritis – kintamųjų tarpusavio ryšių ir šablonų išskyrimas bei apibrėžimas.

Faktorinė analizė leidžia sumažinti duomenų kiekį, kuris dažnai apsunkina rinkos tyrimų analizę. Taikant faktorinės analizės metodus išskiriama keletas bendrųjų veiksnių, kuriuose sutelkta informacija be nuostolių pakeičia informaciją, užfiksuotą dešimtyse kintamųjų.

Faktorinės analizės metodai leidžia priskirti stebimus kintamuosius veiksniams, o kiekvienas veiksnys atvaizduoja kintamųjų empirinius ryšius. Be to, veiksniai kiekvienam kintamajam suteikia svorį, kuris gali būti naudojamas jungiant kintamuosius į integruotus rodiklius.

Analizuojant socialinių grupių ar atskirų asmenų požiūrius, elgesį ar konfliktiškumą, dažnai formuluojamos pradinės hipotezės apie šių savybių tarpusavio ryšius. Faktorinės analizės metodai leidžia įvertinti vienai ar kitai grupei būdingas savybes ir jų ryšius.

Faktorinė analizė gali būti pritaikyta pradiniams duomenims transformuoti taip, kad jie atitiktų kitų statistikos metodų keliamus reikalavimus. Pavyzdžiui, taikant daugiamatės regresijos metodus, reikalaujama, kad nepriklausomieji kintamieji tarpusavyje būtų nekoreliuoti. Kai ši sąlyga netenkinama, pradžioje gali būti taikomi faktorinės analizės metodai, kuriais nepriklausomųjų kintamųjų skaičius sumažinamas iki mažesnės veiksnių reikšmių imties, kurią galima panaudoti regresinei analizei, neprarandant duomenyse esančios informacijos.

Ryšių šablonai gali būti nagrinėjami dviem būdais.

Taikant pirmąjį būdą, tarp kintamųjų ieškomi latentiniai (lot. *latens* (kilm. *latentis*) – paslėptas, nematomas) veiksniai, kurie lemia kintamųjų pokyčius. Galima stebėti šių požymių kitimo reguliarumą ir aprašyti jį kaip kitimo šabloną. Faktorinė analizė, kuri aprašo kintamųjų kitimo šablonus, vadinama *R* faktorine analize.

Tačiau galima išnagrinėti kintamųjų pokyčius ir stebėjimus sugrupuoti pagal jų panašumą. Pritaikę pastebėtus panašumo šablonus individams, grupėms, miestams ar valstybėms gautume vadinamąją *Q* faktorinę analizę. Šios analizės rezultatas – remiantis nustatytų veiksnių skaičiumi sudarytos stebėjimų grupės (klasteriai). Pastaruoju metu šis metodas retai taikomas, nes jau sukurta nemažai specialių daugiamačių kintamųjų sudalijimo į klasterius metodų.

Faktorinės analizės taikymo principai: absoliutus ir euristinis mąstymo būdai

Dažnai statistiniai metodai taikomi siekiant nustatyti ryšį tarp priklausomojo ir nepriklausomųjų kintamųjų. ***Faktorinės analizės tikslas kitas: ištirti daugelio priklausomųjų kintamųjų ryšius. Šio tyrimo tikslas – atrasti ir paaiškinti juos veikiančio nepriklausomojo kintamojo (kuris nėra matuojamas tiesiogiai) prigimtį***. Todėl atsakymai, gauti taikant faktorinės analizės metodus, visuomet yra negalutiniai ir daugiau nuspėjami, palyginti su tais, kurie gaunami tada, kai nepriklausomasis kintamasis yra matuojamas tiesiogiai. Numanomi nepriklausomieji kintamieji yra vadinami veiksniais (*faktoriais*). Taikant faktorinę analizę siekiama nustatyti skirtingų veiksnių, kurie paaiškina tiriamų priklausomųjų kintamųjų ryšius, skaičių bei šių veiksnių prigimtį. Faktorinė analizė leidžia patikrinti, kaip gerai šie hipotetiniai veiksniai paaiškina turimus stebėjimus (duomenis) ir kurią dalį dispersijos paaiškina kiekvienas veiksnys.

Jeigu mes teigiame, kad tik konkretus veiksnių skaičius gali paaiškinti tiriamą reiškinį, tai gauname absoliutaus mąstymo būdo pavyzdį. Jeigu manome, kad geriausiai reiškinį paaiškina du (trys, keturi ar penki) veiksniai, nors jų gali būti ir daugiau, tai gauname euristinio mąstymo būdo pavyzdį.

Euristinio mąstymo būdo samprata padeda suprasti faktorinės duomenų analizės savybes. Keletas mokslininkų gali pritaikyti faktorinę analizę tiems patiems arba labai panašiems kintamiesiems ir nustatyti skirtingą veiksnių skaičių. Toks nevienareikšmiškumas yra būdingas daugeliui sričių. Jeigu norime gauti paprastesnius teorinius samprotavimus, renkamės mažesnį veiksnių skaičių; jei norime išsamesnių rezultatų – veiksnių skaičių tenka padidinti. Faktorinė analizė skiriasi nuo kitų metodų tuo, kad nėra tikslaus kriterijaus, leidžiančio vienareikšmiškai nustatyti veiksnių skaičių. Taikomi kriterijai dažniausiai pateikia daugiau veiksnių, nei galima tinkamai interpretuoti.

Apribojimai: kintamųjų ir stebėjimų (atvejų) skaičius

Stebėjimų skaičius turi būti toks, kad būtų galima gauti patikimus kintamųjų koreliacijos įverčius. Jeigu imtis nedidelė ir dar yra nutolusių reikšmių, koreliacijos koeficientų įverčiai gaunami netikslūs. Be to, reikiamas stebėjimų skaičius priklauso nuo tikrojo veiksnio struktūros. Kuo aiškesnė ir paprastesnė veiksnio struktūra, tuo mažesnės duomenų

imties užtenka šiai struktūrai atskleisti. Pagal bendrą nuomonę netikslinga taikyti faktorinės analizės metodus, jeigu stebėjimų yra mažiau nei 50.

Taisyklė, nusakanti ***kintamųjų*** skaičių, yra priešinga regresinės analizės nuostatoms. Akivaizdu, kad vienas kintamasis negali apibūdinti bendrojo veiksnio. Tuo tikslu reikia turėti ne mažiau kaip tris kintamuosius. Apskritai kalbant, kuo daugiau kintamųjų, tuo geriau, jeigu šie kintamieji yra susiję su juos veikiančiais veiksniais. Kartais kintamųjų gali būti daugiau nei stebėjimų (atvejų). Tačiau bendrai rekomenduojama, kad imties dydis būtų dukart didesnis nei kintamųjų skaičius, t. y. kad stebėjimų būtų dukart daugiau nei kintamųjų, ir stebėjimai būtų heterogeniniai taikomų matų atžvilgiu. Faktorinei analizei tinka kiekybiniai ir dichotominiai kintamieji (jei dichotominis kintamasis koduojamas skaičiais 1 ir 2). Dėl ranginių kintamųjų galimi tik euristiniai faktoriniai modeliai be rezultatų statistinės interpretacijos. Tai reiškia, kad galima atlikti ranginių kintamųjų faktorinę analizę, tačiau negalima apskaičiuoti veiksnių reikšmių ir nustatyti veiksnių reikšmingumo.

Be šių apribojimų, faktorinės analizės metodais gaunami patikimesni rezultatai, jei tenkinamos šios prielaidos:

- kintamieji turi daugiamatį normalųjį skirstinį;
- kintamieji susiję tiesinės formos ryšiais;
- nėra ryškiai išsiskiriančių (nutolusių) reikšmių.

Faktorinės analizės etapai:

1. Apskaičiuoti stebimų duomenų tarpusavio koreliacinę matricą ir nustatyti, ar duomenys tinka faktorinei analizei.
2. Rasti pradinį galimų veiksnių sprendinį.
3. Iš pradinio sprendinio nustatyti tinkamiausią veiksnių kiekį.
4. Jeigu reikia išryškinti veiksnių šabloną, pritaikyti veiksnių sukimo metodą.
5. Atsižvelgiant į sprendžiamą uždavinį, kiekvienam tiriamam objektui apskaičiuoti veiksnių reikšmes (*factor scores*).

Faktorinės analizės duomenimis gali būti pradiniai stebėjimai arba jų koreliacinė matrica. Jeigu skaičiuoti naudojami pradiniai stebėjimai, tai pirmuoju faktorinės analizės žingsniu šie stebėjimai standartizuojami (z transformacija) ir tarp jų reikšmių apskaičiuojami Pirsono koreliacijos koeficientai.

Pagrindiniu tolesnių skaičiavimų elementu tampa koreliacinė matrica. Jeigu koreliacinė matrica gali būti paaiškinta bendruoju veiksniu g,

tai tada turi būti tokia stebimų kintamųjų koreliacijos su kintamuoju *g* aibė, kad bet kurių dviejų šių koreliacijos koeficientų sandauga būtų lygi koreliacijai tarp stebimų kintamųjų.

Jeigu mes įvertinome veiksnio koreliaciją su stebimais kintamaisiais, tai galime pasakyti su kuriais kintamaisiais veiksnys koreliuoja stipriai, o su kuriais – silpnai. Taikant faktorinės analizės metodus negalima pasakyti, kokią įtaką veiksniui turi atskiras stebimas asmuo (respondentas). Tačiau, jei norima įvertinti atskiro asmens įtaką veiksniui, tai galima padaryti taikant daugiamatės regresinės analizės metodus.

Koreliacinės matricos gali nesiderinti su vieno bendrojo veiksnio hipoteze, tačiau gali pasitvirtinti hipotezė apie du, tris ar dar daugiau bendrųjų veiksnių. Kuo mažiau veiksnių, tuo paprastesnės hipotezės. Kadangi paprastos hipotezės turi loginį mokslinį pranašumą prieš sudėtingesnes hipotezes, mažiau veiksnių turinčios hipotezės yra vertesnės nei apimančios daug veiksnių.

Nesigilinant į matematines išraiškas, galima pasakyti, kad faktorinė analizė mėgina išreikšti kiekvieną kintamąjį kaip dviejų dalių – bendrosios ir specifinės – sumą. Bendrąją kintamųjų dalį išsamiai paaiškina bendrieji veiksniai, o specifinės dalys yra visiškai nekoreliuotos tarpusavyje. Kokiu laipsniu turimi duomenys atitinka šią sąlygą, sprendžiama iš vadinamosios liekanų kovariacinės matricos. Šios matricos įstrižainės elementai nėra lygūs vienetui, tačiau tipiniais metodais ši matrica gali būti transformuota į koreliacinę matricą (kiekvieną nediagonalųjį elementą dalijant iš kvadratinės šaknies iš jį atitinkančių diagonaliųjų elementų sandaugos). Iš taip apskaičiuotų koreliacinės matricos elementų sprendžiama, ar tiriami kintamieji suderinami su *m* bendrųjų veiksnių hipoteze. Jeigu apskaičiuotos koreliacijų reikšmės yra didelės, jos yra nesuderinamos su hipoteze, kad populiacijoje jos lygios nuliui, ir tada hipotezę, kad yra *m* bendrųjų veiksnių, tenka atmesti. Didinant veiksnių skaičių *m*, šios koreliacijos mažėja ir hipotezė apie veiksnių skaičių geriau derinasi su duomenimis.

Visada norima rasti mažiausią veiksnių skaičių *m*, kuris derintųsi su duomenimis. Tuo tikslu skaičiuoti pradedama nuo mažiausio veiksnių skaičiaus (paprastai $m = 1$) ir tikrinama hipotezė, kad šis skaičius atitinka turimus duomenis. Jei reikia, skaičius *m* didinamas. Kiekvienas veiksnių skaičiaus *m* didinimas daro hipotezę sudėtingesnę, bet geriau atitinka duomenis. Sustojama, kai randama, kad hipotezė suderinama su duomenimis.

Centrinė faktorinės analizės teorema yra kiek panaši į dispersinės analizės centrinę teoremą, kuri teigia, kad kintamojo *X* dispersija gali būti išskirstyta arba, kitaip sakant, padalyta į atskirus komponentus.

Analogiškai dispersinei analizei, faktorinės analizės kovariacijų matrica *R* gali būti padalyta į bendrąją dalį *C*, kurią paaiškina veiksnių aibė, ir specifinę dalį *U*, kurios šie veiksniai nepaaiškina. Kalbant matricų terminais, $R = C + U$. Kaip ir dispersinėje analizėje, paaiškinta dalis *C* gali būti dalijama toliau į komponentų matricas c_1, c_2, …, kurias paaiškina atskiri veiksniai. Kiekvienas iš šių vieno veiksnio komponentų c_j lygus veiksnių svorių stulpelio sandaugai. Kiekvienas diagonalusis šios matricos elementas yra veiksnio paaiškintas dispersijos kiekis.

Liekanų matrica *U* yra kovariancinė matrica, kurios reikšmės rodo veiksnio nepaaiškintą kiekvieno kintamojo dispersijos dalį. Dažnai *C* yra kelių matricų c_j suma, o ne atskira matrica. Komponentų matricų c_1, c_2, …, kurių suma sudaro matricą *C*, skaičius yra vadinamas matricos *C* rangu. Matricos *C* rangas yra modelio bendrųjų veiksnių skaičius. Jeigu apibrėžiamas veiksnių skaičius *m*, matricos *C* ir *U* (kurių suma yra pradinė kovariacijų matrica *R*) sudaromos taip, kad matricos *C* rangas būtų *m*. Kuo didesnis *m*, tuo geriau *C* aproksimuoja *R*. Jeigu $m = k$, kur *k* yra matricos kintamųjų skaičius, tai kiekvienas matricos *C* elementas bus lygus matricos *R* elementui, o matricos *U* elementai bus nuliai. Tikslas – rasti kuo mažesnę veiksnių aibę *m* ir pasiekti, kad *C* vis dar tinkamai aproksimuotų *R*.

Veiksnių skaičius. Kartais veiksnių skaičius *m* gali būti spėjamas iš anksto. Jeigu taip nėra, tai taikomi matematiniai metodai. Dažniausiai taikomi metodai pagrįsti koreliacinės matricos tikrinėmis reikšmėmis. Populiariausi iš jų – *Kaizerio* (*Kaiser*) tikrinių reikšmių taisyklė ir *Ketelo* (*Cattell*) *scree* kriterijus. *H. Kaizeris* pasiūlė tokią taisyklę veiksnių skaičiui nustatyti: veiksnių skaičius *m* turėtų būti lygus matricos tikrinių reikšmių skaičiui minus vienetas. Bendra šios taisyklės taikymo logika yra tokia: tikrinė reikšmė parodo dispersijos kiekį, kurį paaiškina vienas veiksnys, todėl nėra prasmės įtraukti veiksnį, kuris paaiškina mažiau dispersijos, negu jos yra viename kintamajame. Ši taisyklė buvo sukurta tais laikais, kai faktorinei analizei nebuvo taikomos kompiuterinės programos. Šiuo metu galima nesunkiai atlikti ir palyginti analizės rezultatus su skirtingais veiksnių rinkiniais.

H. Kaizerio metodo alternatyva – *scree* kriterijus. Taikant šį metodą, iš eilės ir nuosekliai braižoma matricos tikrinių reikšmių diagrama.

Tada pagal brėžinį sprendžiama, kurioje vietoje brėžinys išsilygina (nusistovi). Šio metodo trūkumas tas, kad iš brėžinio galima gauti skirtingas išvadas.

Kitas taikomas metodas panašus į *scree* kriterijų, tačiau remiasi papildomais skaičiavimais. Kiekvienai tikrinei reikšmei T yra apskaičiuojama visų tolesniųjų tikrinių reikšmių suma S, kartu įtraukiant ir pačią reikšmę T. Santykis T/S yra T nepaaiškintos variacijos dalis. Pavyzdžiui, tarkime, kad devynių kintamųjų modelyje gauname keturias paskutiniąsias tikrines reikšmes, lygias 0,7, 0,3, 0,15 ir 0,1. Jų suma yra 1,25, o santykis $T/S = 0,7/1,25 = 0,56$. Taigi, pridėjus dar vieną veiksnį, bendras modelis paaiškintų 0,56, arba 56 procentus, anksčiau nepaaiškintos dispersijos. Skaičiuodami toliau, gautume, kad didinant veiksnių skaičių, dar kitas veiksnys paaiškintų $T/S = 0,3/(0,3 + 0,15 + 0,1) = 0,54$. Vadinasi, modelis su dar vienu veiksniu paaiškintų 54 procentus anksčiau nepaaiškintos dispersijos.

Be šių taisyklių, bendrųjų veiksnių skaičiui nustatyti taikomi ir kiti metodai. Kadangi, taikant šiuolaikines programas, skaičiavimai yra greiti, kartu su kitais metodais galima taikyti ir patį paprasčiausią logiškai pagrįstą metodą. Skaičiavimai atliekami įvairiems veiksnių skaičiams m ir parenkamas toks jų skaičius, kuris geriausiai paaiškina tiriamą reiškinį.

Nustatant maksimalų veiksnių skaičių derėtų laikytis šios taisyklės: jeigu kintamųjų yra p, tai *maksimalų* galimą veiksnių skaičių m galima rasti iš formulės:

$$(p + m) \leq (p + m)^2. \qquad (8.1)$$

Nesilaikant šios taisyklės uždavinys gali išsigimti ir tapti trivialus.

Taikant įvairias taisykles, visada būtina atminti, kad faktorinės analizės tikslas – paaiškinti daugiamačius kintamuosius kuo mažesniu veiksnių skaičiumi.

Sukimas. Pradinius kintamuosius galima pakeisti jų tiesinėmis funkcijomis neprarandant informacijos, nes tiesinės funkcijos turi tą patį informacijos kiekį kaip ir pradiniai kintamieji. Tiesinės kintamųjų funkcijos gali užtikrinti paprastesnę ir suprantamesnę modelio struktūrą nei pradiniai kintamieji.

Šį metodą galima pritaikyti faktorinei analizei. Faktorinėje analizėje manoma, kad stebėti kintamieji X_1, X_2,..., X_k sudaro priklausomųjų kin-

tamųjų aibę, o veiksniai F_1, F_2,..., F_m yra nepriklausomieji kintamieji. Tada gauname k lygčių sistemą:

$$X_1 = \lambda_{11}F_1 + \lambda_{12}F_2 + ... + \lambda_{1m}F_m + \varepsilon_1,$$
$$X_2 = \lambda_{21}F_1 + \lambda_{22}F_2 + ... + \lambda_{2m}F_m + \varepsilon_2,$$
$$.....$$
$$X_k = \lambda_{k1}F_1 + \lambda_{k2}F_2 + ... + \lambda_{km}F_m + \varepsilon_k. \quad (8.2)$$

Daugikliai λ_{ij} vadinami veiksnių svoriais, ε_i – specifinis (charakteringasis) veiksnys.

Reikia atsiminti, kad *veiksniai F_1, F_2,..., F_m yra kintamųjų funkcija*, *o ne kintamieji*. Nežinomi kintamieji, kurie sudaro kiekvieną funkciją F, yra susieti nežinomais būdais. Žinodama kintamuosius X_1, X_2,..., X_k faktorinė analizė nustato nežinomas funkcijas F_1, F_2,..., F_m. Standartizuoti šios algebrinės lygties koeficientai (konstantos) sudaro matricą $k \times m$, kuri vadinama veiksnių svorių (*factor loadings*) matrica. Be to, faktorinė analizė suskaičiuoja veiksnių reikšmes (*factor scores*) ir veiksnių reikšmių koeficientus.

Jeigu originalūs veiksniai bus pakeisti jų tiesinėmis funkcijomis, gausime tuos pačius priklausomųjų kintamųjų įverčius kaip ir anksčiau, tačiau veiksnių svorių matricos skirsis. Išnagrinėję daugelį galimų tiesinių funkcijų, rastume vieną, kurios veiksnių svorių matrica būtų paprasčiausia. Paprastumas suprantamas kaip matricos nulinių elementų kiekis: kuo daugiau nulių, tuo paprastesnė struktūra. Sukimo būdas, kuris reikalauja, kad veiksniai liktų nekoreliuoti, vadinamas ortogonaliuoju sukimu, o kiti sukimo būdai – pasviruoju sukimu. Tipiniai ortogonalaus sukimo metodai yra *Varimax*, *Quartimax*, *Equamax* ir *Orthomax*. Dažniausiai taikomas *Varimax* sukimo būdas. Pasvirieji sukimai duoda paprastesnę struktūrą, tačiau, interpretuojant rezultatus, kartu būtina nagrinėti veiksnių tarpusavio koreliacijas. Reikia pasakyti, kad matematinė veiksnių sukimo procedūra paremta iteraciniais skaičiavimais, todėl kartais veiksnių gali nepavykti pasukti.

Pagrindinių komponenčių analizė (nagrinėjama suminė dispersija) ir bendrųjų veiksnių analizė (nagrinėjama bendroji dispersija) – du svarbiausi faktorinės analizės metodai. Tarp šių dviejų metodų yra konceptualus skirtumas: bendrųjų veiksnių modelius galima įvertinti statisti-

niais kriterijais, o pagrindinių komponenčių modeliams to padaryti negalima. Taip yra dėl to, kad bendrųjų veiksnių modelių specifiniai (charakteringieji) veiksniai privalo būti nekoreliuoti, o pagrindinių komponenčių modelių liekanos tarpusavyje koreliuoja. Pagaliau komponentės nėra latentiniai kintamieji, jie yra tiesinės pradinių kintamųjų kombinacijos, taigi determinuoti. Tuo tarpu veiksniai yra latentiniai kintamieji, taigi nedeterminuoti.

Nors pagrindinių komponenčių analizė (PKA) sprendžia panašias problemas kaip ir bendrųjų veiksnių analizė, tačiau skirtingomis priemonėmis. PKA formuluoja ir sprendžia tiksliai statistiškai apibrėžtas problemas ir gaunamas sprendinys, išskyrus keletą specialių atvejų, yra vienintelis su aiškiomis matematinėmis savybėmis. Tačiau, taikant PKA praktiškai, susiduriama su problemomis. Todėl dažnai PKA taikoma kaip bendrųjų veiksnių analizės aproksimacija. Jeigu analizės tikslas yra sumažinti duomenų kiekį, reikia taikyti pagrindinių komponenčių analizės metodiką. Jeigu norima paaiškinti pradinių duomenų tarpusavio koreliacijas ir sudaryti testuojamą modelį, reikia taikyti faktorinės analizės metodus.

Pilnoji koreliacinė matrica pateikiama tais atvejais, kai kintamųjų skaičius nėra labai didelis. Paprastai koreliacinė matrica pateikiama be komentarų, nes ji nėra analizės tikslas. Tačiau koreliacinė matrica apima daug naudingos informacijos, padedančios suvokti ryšius tarp kintamųjų porų.

Kad *duomenys tiktų faktorinei analizei*, jie turėtų būti koreliuoti (susiję) tarpusavyje. Jeigu duomenys nesusiję, tai nėra ir bendrumų. Todėl prieš pradedant faktorinę analizę reikia nustatyti, ar šio metodo taikymas turi prasmę. Ar kintamieji susiję, tarpusavyje galima spręsti iš pradinės koreliacinės matricos koeficientų. Jeigu kintamieji nesusiję, jie negali būti grupuojami ir gali sudaryti tik atskirus veiksnius. Tokius kintamuosius iš pradinių kintamųjų sąrašo derėtų pašalinti.

Ar duomenys (kintamieji) tinka faktorinei analizei, tikrinama Kaizerio–Mejerio–Olkino (*KMO*) matu ir *Bartleto* sferiškumo kriterijumi. Šie du metodai sudaro minimalų standartą, kuris būtinas prieš atliekant faktorinę ar pagrindinių komponenčių analizę.

KMO matas tikrina, ar kintamųjų daliniai koreliacijos koeficientai yra maži. *KMO* reikšmės kinta nuo 0 iki 1 ir kuo reikšmės artimesnės vienetui, tuo geriau. *KMO* matas skaičiuojamas pagal formulę:

$$KMO = \frac{\sum\sum_{i \neq j} r_{ij}}{\sum\sum_{i \neq j} r_{ij} + \sum\sum_{i \neq j} \tilde{r}_{ij}}, \qquad (8.3)$$

čia: r_{ij} – koreliacijos koeficientas,

$\tilde{r}_{ij}$ – dalinės koreliacijos koeficientas.

Jei *KMO* reikšmė maža, tai nagrinėjami rodikliai koreliacinei analizei netinka. *KMO* reikšmių interpretacija pateikta 8.1 lentelėje.

8.1 lentelė. **Kaizerio–Mejerio–Olkino mato interpretacija**

KMO reikšmė	*Interpretacija*
$0,9 < KMO \leq 1,0$	Faktorinė analizė tinka idealiai
$0,8 < KMO \leq 0,9$	Faktorinė analizė tinka puikiai
$0,7 < KMO \leq 0,8$	Faktorinė analizė tinka vidutiniškai
$0,6 < KMO \leq 0,7$	Faktorinė analizė tinka pakenčiamai
$0,5 < KMO \leq 0,6$	Faktorinė analizė tinka blogai
$0,0 \leq KMO \leq 0,5$	Faktorinė analizė netinka

Kitas kintamųjų tarpusavio ryšio stiprumo rodiklis yra *Bartleto* (*Bartlett*) *sferiškumo testas*. Šis testas tikrina nulinę hipotezę, kad populiacijos koreliacinės matricos kintamieji yra nekoreliuoti, t. y. kad koreliacinė matrica yra vienetinė, jos įstrižainės elementai yra vienetai, o visi kiti – nuliai (kitaip sakant, kintamieji yra nekolinearūs).

Skaičiuojant pateikiami standartiniai pagrindinių komponenčių analizės rezultatai. Įvedamos naujos komponentės paaiškina kaskart vis mažiau ir mažiau dispersijos. Visa matricos dispersija priklauso nuo kintamųjų skaičiaus.

Pasirinkus veiksnių skaičių, galima apskaičiuoti veiksnių svorius. Veiksnio svoris – tai koreliacija tarp kintamojo ir veiksnio, kuris buvo gautas (kitaip ištrauktas) iš stebėtų duomenų. Skaičiavimo programos pateikia keletą veiksnių matricos variantų. Pirmasis – nepasukta veiksnių matrica, antrasis – pasuktų veiksnių svorių matrica.

Veiksnių interpretavimas ir sukimas

Vienas iš pagrindinių faktorinės analizės uždavinių – suprasti ir aprašyti, ką matuoja veiksniai, o svarbiausias rodiklis, padedantis suprasti veiksnių prigimtį, yra ***veiksnių svoriai***. Idealiu atveju vieno veiksnio kintamojo svoris turėtų būti didelis (arti vieneto), o kitų veiksnių arti nulio. Praktiškai pirmiausia reiktų nustatyti, kuris kintamojo svoris veiksniui yra didžiausias.

Interpretuojant veiksnius galima taikyti žymių metodiką, t. y. nustatytiems veiksniams priskirti atitinkamas žymes. Šios žymės palengvina rezultatų aptarimą ir tolesnę jų analizę. Interpretacijai taikomos trijų tipų žymės: *simbolinės*, *aprašomosios* ir *priežastinės*.

Simbolinės žymės yra paprasčiausi simboliai be jokių savarankiškų prasmių. Jų tikslas – pažymėti šablonus. Trijų veiksnių šablonai gali būti žymimi, pavyzdžiui, A, B ir C arba A1, A2 ir A3. Šios žymės tiesiog priskiriamos veiksniams be jokios papildomos prasmės ar interpretacijos.

Nors simbolinės žymės padeda išvengti nesusipratimų, jos nepadeda aiškinti tyrimų rezultatų ir lyginti tyrimų. Tokių trūkumų neturi ***aprašomosios*** žymės, kai veiksniams priskiriami reikšminiai žodžiai. Aprašomosioms žymėms vartojamos sąvokos turi atspindėti tiriamo reiškinio esmę ir padėti rezultatus suskirstyti į grupes.

Priežastinių žymių tikslas – atskleisti šablonus sukeliančias priežastis. Priežastinė žymė yra glaustas paaiškinimas, kodėl šablonas įtraukia tam tikrus kintamuosius.

Veiksnių šablonų interpretavimo būdas priklauso nuo asmeninio požiūrio ir patirties. Tyrėjas gali vartoti sąvokas, kurios skatina mąstyti apie gautus rezultatus ir padaro juos lengviau suprantamus. Tačiau šiuo atveju kyla klaidingos interpretacijos ar konkretumo praradimo pavojus. Kartais suteikta simbolinė etiketė yra tikslesnė ir aiškesnė nei išsamus aprašas to, ką ji reiškia.

Jeigu vienas ar keli kintamieji turi beveik vienodus svorius veiksniuose, kartu veiksnių aiškinimą padarydami nevienareikšmį, skaičiuojant veiksnių svorius taikomos įvairios, dažniausiai ortogonalios, veiksnių sukimo strategijos. Šių strategijų tikslas – gauti aiškius veiksnių svorius, t. y. kad vieni kintamieji turėtų didelius veiksnių svorius (idealiu atveju – vienetus), o kiti – mažus (idealiu atveju – nulius). Taip būtų išryškinti svarbiausi šablonai.

Veiksnių reikšmių (*factor scores*) skaičiavimas yra vienas iš pagrindinių tyrimo etapų. Iš tikrųjų neužtenka nustatyti vien veiksnių, paaiškinančių tiriamų požymių tarpusavio koreliacijas ir pagrindinę jų dispersijos dalį, egzistavimą. Pageidautina tiesiogiai apibrėžti šiuos bendrus veiksnius, aprašyti juos taikant kintamuosius ir pateikti patogią jų interpretaciją. Tam taikomi įvairūs metodai, pavyzdžiui, *Bartleto*, *Tomsono* (*Thomson*) arba *Andersono–Rubino*. Skaičiuojant veiksnių reikšmes, dažniausiai veiksniai interpretuojami kaip nežinomi regresijos lygties koeficientai.

Praktiškai taikantys faktorinę analizę tyrėjai teigia, kad sąlygos, kurias teoriškai tenkina veiksnių reikšmės (pavyzdžiui, jų dispersija lygi vienetui, jeigu veiksniai yra ortogonalūs), ne visada tenkinamos. Apskaičiuotos veiksnių reikšmės retai kada turi vienetinę dispersiją ir jos dažnai koreliuoja tarpusavyje, nors, atliekant analizę, veiksniai buvo ortogonalūs.

Veiksnių reikšmėms apskaičiuoti taikomi metodai, kuriuos galima suskirstyti į dvi grupes. Pirmąją grupę sudaro vadinamieji *tikslūs* metodai (žr. 8.2 lentelę), pasižymintys skirtingomis savybėmis. Vienas iš pirmųjų, *Terstouno* (*Thurstone*) pasiūlytų, regresijos metodų minimizuoja neapibrėžtumą, tačiau juo apskaičiuoti veiksnių reikšmių įverčiai koreliuoja tarpusavyje net ir tada, kai veiksniai yra ortogonalūs. *Andersono–Rubino* metodu apskaičiuoti veiksnių reikšmių įverčiai yra visiškai ortogonalūs, tačiau neminimizuoja neapibrėžtumo.

Vienareikšmiškumas, *esant ortogonaliems veiksniams*, parodo, ar veiksnių reikšmių įverčiai gali perimti („užsikrėsti“) dispersiją nuo kitų ortogonalių veiksnių. Koreliacijos išsaugojimas parodo, ar koreliacija tarp veiksnių reikšmių įverčių atitinka koreliaciją tarp pačių veiksnių. Nors šių metodų skaičiavimo algoritmai sudėtingi, juos galima rasti populiariose statistinėse programose. Pavyzdžiui, SPSS paketas siūlo regresijos, *Andersono–Rubino* (tinkamo tik ortogonaliems veiksniams), *Berge*, *Krijneno*, *Vansbyko* ir *Bartleto* metodus.

Antrąją metodų grupę sudaro vadinamieji *netikslūs*, *apytiksliai* metodai. Skaičiuojant veiksnių reikšmes šiais metodais, pradinės kintamųjų reikšmės yra standartizuojamos ir dauginamos iš atitinkamų kintamųjų veiksnių svorių, pateikiamų veiksnių svorių matricoje. Šios sandaugos susumuojamos. Veiksnio reikšmės parodo objekto (valstybės, miesto ar atskiro asmens) sąsają su nagrinėjamu veiksniu. Tai sudėtinis matas, kuris gali būti apskaičiuotas kiekvienam tiriamam objektui. Kintamųjų,

menkai susietų su konkrečiu veiksniu, svoris artimas nuliui. Šie veiksnių reikšmių koeficientai gali būti pavaizduoti grafiškai ir taikomi tolesnei analizei. Taip apskaičiuotos veiksnių reikšmės dažnai pasitaiko literatūroje ir vadinamos suminiais koeficientais (*total*), indeksais (*index*), domenais (*domain*), aspektais (*facet*), skalės (*scale*) ar poskalės (*subscale*) reikšmėmis. Šios *apytikslės* veiksnių reikšmės tarpusavyje koreliuoja net ir tada, kai veiksniai yra ortogonalūs, jų dispersija nelygi vienetui, ir beveik visada šios reikšmės silpniau koreliuos su veiksniais nei gautos vadinamaisiais tiksliais metodais.

8.2 lentelė. **Veiksnių reikšmių įverčių skaičiavimo metodų savybės**

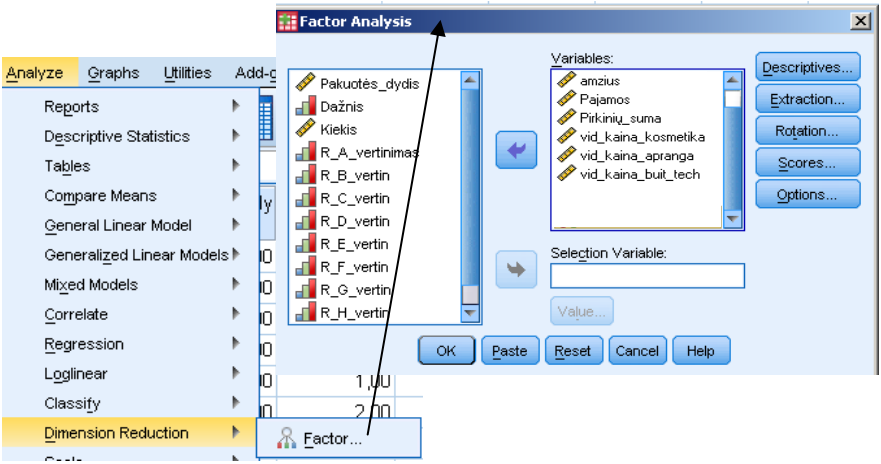
Metodas	*Maksimizuoja patikimumą (validity)*	*Vienareikšmiškumas esant ortogonaliems veiksniams*	*Koreliacijos išsaugojimas*
Terstouno (*Thurstone*) regresijos metodas (1935)	Taip	Ne	Ne
Berge, Krijneno, Vansbyko (*Wansbeek*) ir Šapiro metodas (1999)	Ne	Ne	Taip
Andersono–Rubino metodas (1956)	Ne	Ne	Taip
Bartleto metodas (1937)	Ne	Taip	Ne
Harmano metodas (1976)	Ne	Taip	Ne

Pagerinti metodai gali užtikrinti tam tikras statistines savybes: patikimumo maksimizavimą ar veiksnių reikšmių įverčių ortogonalumą. Apytiksliais metodais įverčiai lengviau apskaičiuojami ir apskritai teigiama, kad jie stabilesni. Pradžioje buvo manoma, kad abiejų šių grupių metodai smarkiai koreliuoja tarpusavyje ir kad jų skirtumai nėra esminiai. Tačiau po gausių tyrimų pasirodė, kad šie skirtumai iš esmės gali būti reikšmingi. Kol kas yra neaišku, ar tiksliųjų metodų nestabilumas

yra didelis. Todėl pasirinkti vieną iš skirtingų veiksnių reikšmių skaičiavimo procedūrų nėra taip paprasta. Tai turėtų priklausyti nuo konkrečios tyrimo programos konteksto.

Galutinis faktorinės analizės rezultatas – tai veiksnių įverčiai. Užuot kūrus žemėlapius su atskirais erdviniais kintamųjų pavyzdžiais, galima pateikti diagramą su nustatytųjų veiksnių reikšme. Kiekvieno veiksnio įvertis gali būti pateikiamas linijine pradinių kintamųjų diagrama.

8.1. Faktorinė analizė SPSS terpėje



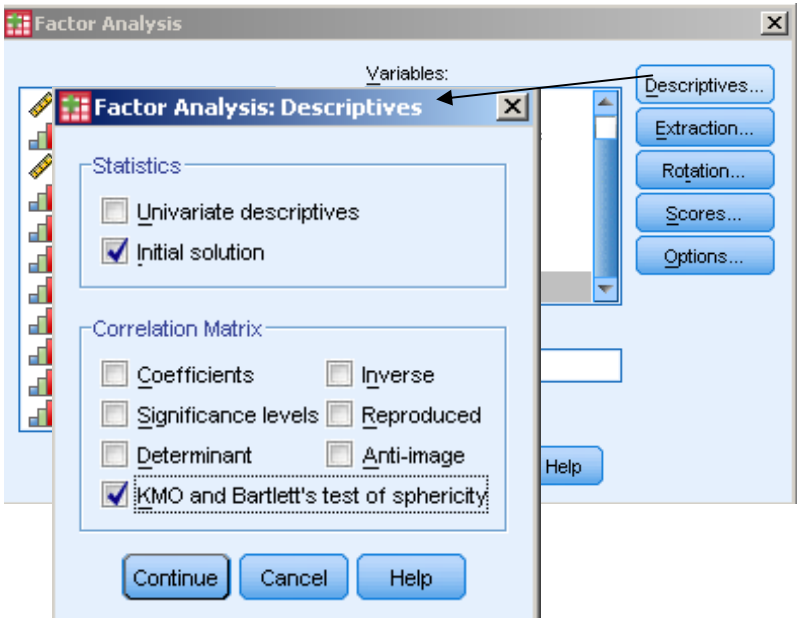
8.1 pav. **Dialogo langas *Factor Analysis***

Faktorinės analizės dialogo langas *Factor Analysis* atveriamas komanda *Analyze* → *Dimension Reduction* → *Factor...* (žr. 8.1 pav.).

Faktorinei analizei parinksime kintamuosius: respondentų amžius (*amžius*), respondentų pajamos per mėnesį (*Pajamos*), suma, skirta per mėnesį pirkiniams (*Pirkinių_suma*), vidutinė kaina, kurią respondentas sutinka mokėti už kosmetiką (*vid_kaina_kosmetika*), vidutinė kaina, kurią respondentas sutinka mokėti už aprangą (*vid_kaina_apranga*), vidutinė kaina, kurią respondentas sutinka mokėti už buitinę techniką

(*vid_kaina_techn*). Analizuojami kintamieji perkeliami į langelį *Variables.*

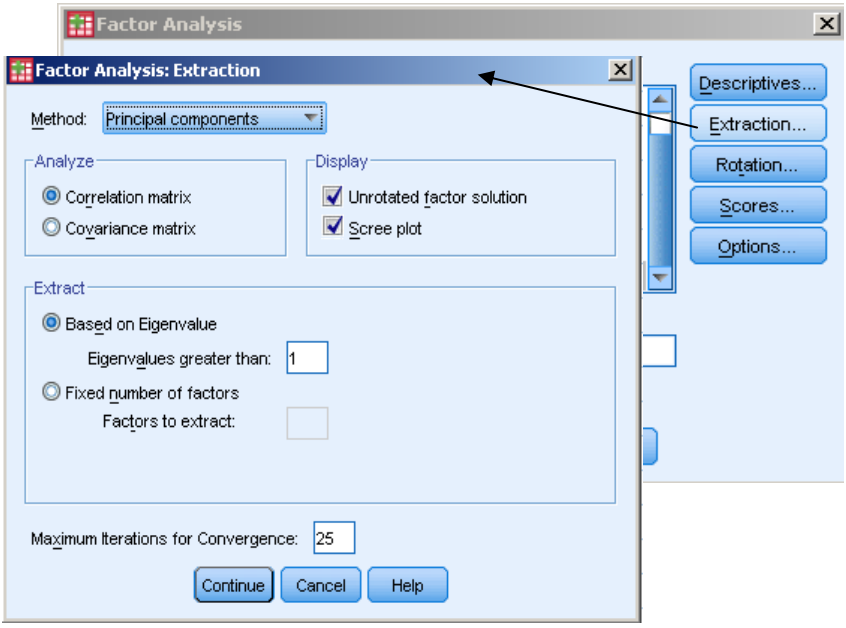
Mygtuko *Descriptives…* paspaudimu atveriamas dialogo langas *Factor Analysis: Descriptives* (žr. 8.2 pav.).



8.2 pav. **Dialogo langas *Factor Analysis: Descriptives***

Šiame lange reikia pažymėti langelį *KMO and Bartlett's test of sphericity.* Šis kriterijus leidžia nustatyti kintamųjų tinkamumą faktorinei analizei. Kitos nurodytos šiame lange statistikos rinkos tyrimuose dažniausiai nenaudojamos. Paspaudę mygtuką *Continue* grįžtame į pradinį langą. Mygtuko *Extraction...* paspaudimu atveriamas dialogo langas *Factor Analysis: Extraction* (žr. 8.3 pav.). Nustatytasis veiksnių nustatymo metodas (*Method → Principal components*) geriausiai tinka sprendžiant daugumos rinkos tyrimų užduotis. Veiksnių skaičius nustatomas srityje *Extract.* Nustatytas yra veiksnių skaičiaus apskaičiavimas pagal tikrines reikšmes (*Based on Eigenvalue*). Tačiau galima užduoti tam tikrą veiksnių skaičių (*Fixed number of factors*). Langelio *Scree*

plot žymėjimas leidžia nubraižyti tikrinių įverčių (*Scree plot*) diagramą. Paspaudę mygtuką *Continue* grįžtame į pradinį langą.



8.3 pav. **Dialogo langas *Factor Analysis: Extraction***

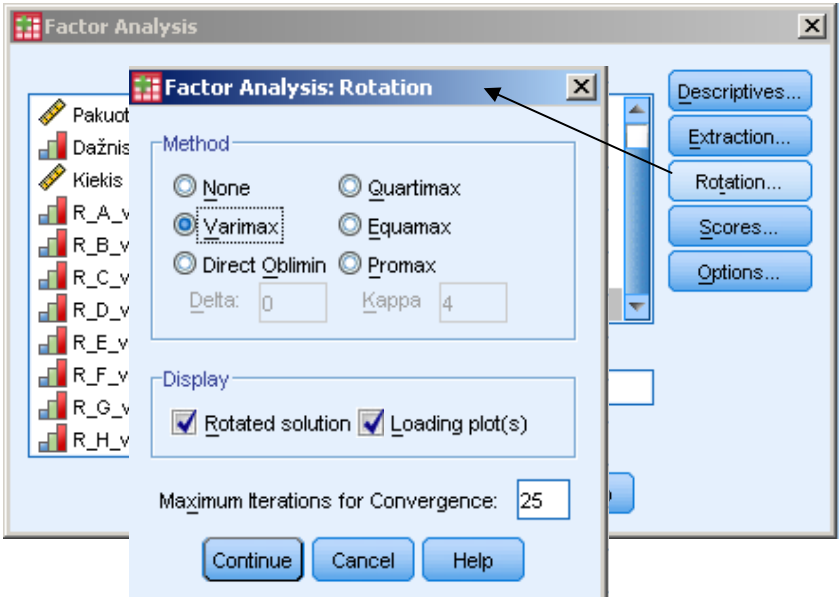
Mygtuko *Rotation*… paspaudimu atveriamas dialogo langas *Factor Analysis: Rotation* (žr. 8.4 pav.). Šiame lange reikia užduoti koreliacijos koeficientų matricos sukimo būdą (dažniausiai pasirenkamas *Varimax* metodas). Langelio *Loading plot*(*s*) žymėjimas leidžia nubraižyti veiksnių – kintamųjų žemėlapį. Paspaudę mygtuką *Continue* grįžtame į pradinį langą.

Mygtuko *Scores*… paspaudimu atveriamas dialogo langas *Factor Analysis: Factor Scores* (žr. 8.5 pav.). Šiame lange reikia pažymėti langelį *Display factor score coefficient matrix.* Tai suteikia galimybę išvesti veiksnių įverčius. Paspaudę mygtuką *Continue* grįžtame į pradinį langą.

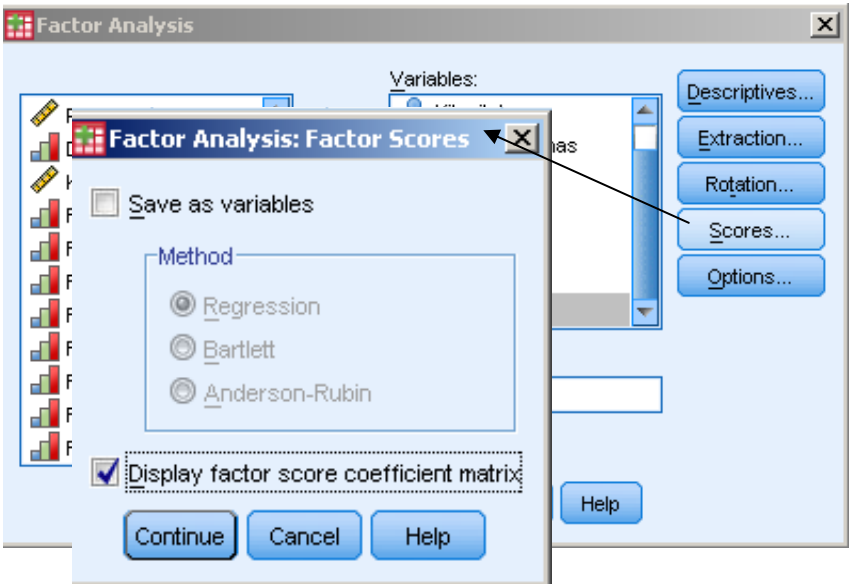
Procedūra baigiama mygtuko *OK* paspaudimu. Lange *Output-SPSS Viewer* rodomi analizės rezultatai.

Atliktos procedūros sintaksė:

```
FACTOR
  /VARIABLES amzius Pajamos Pirkiniu_suma
vid_kaina_kosmetika
  vid_kaina_apranga vid_kaina_buit_tech /MISSING LISTWISE
/ANALYSIS amzius
  Pajamos Pirkiniu_suma vid_kaina_kosmetika vid_kaina_apranga
  vid_kaina_buit_tech
  /PRINT INITIAL KMO EXTRACTION ROTATION FSCORE
  /PLOT EIGEN ROTATION
  /CRITERIA MINEIGEN(1) ITERATE(25)
  /EXTRACTION PC
  /CRITERIA ITERATE(25)
  /ROTATION VARIMAX
  /METHOD=CORRELATION .
```



8.4 pav. **Dialogo langas *Factor Analysis: Rotation***



8.5 pav. **Dialogo langas *Factor Analysis: Factor Scores***

Ar duomenys (kintamieji) tinka faktorinei analizei, tikrinama Kaizerio–Mejerio–Olkino (*KMO*) matu ir *Bartleto* sferiškumo kriterijumi (žr. 8.6 pav.). Lentelėje *KMO and Bartlett's Test KMO = 0,618* reikšmė rodo, kad duomenys faktorinei analizei tinka patenkinamai. *Bartleto* sferiškumo kriterijus (*Bartlett's test of signification*) rodo, ar tarp analizuojamų kintamųjų yra statistiškai reikšmingai koreliuojančių, t. y. tikrina tokią hipotezę:

H_0: visi kintamieji yra nekoreliuoti,
H_1: kintamieji koreliuoti.

Pasirenkame reikšmingumo lygmenį $\alpha = 0,05$, *p-level* $= 0,000 < \alpha$. Patvirtinta hipotezė H_1. Turimiems duomenims faktorinė analizė turi prasmę.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,618
Bartlett's Test of Sphericity	Approx. Chi-Square	134,728
	df	15
	Sig.	,000

8.6 pav. **Faktorinės analizės rezultatas**

Total Variance Explained

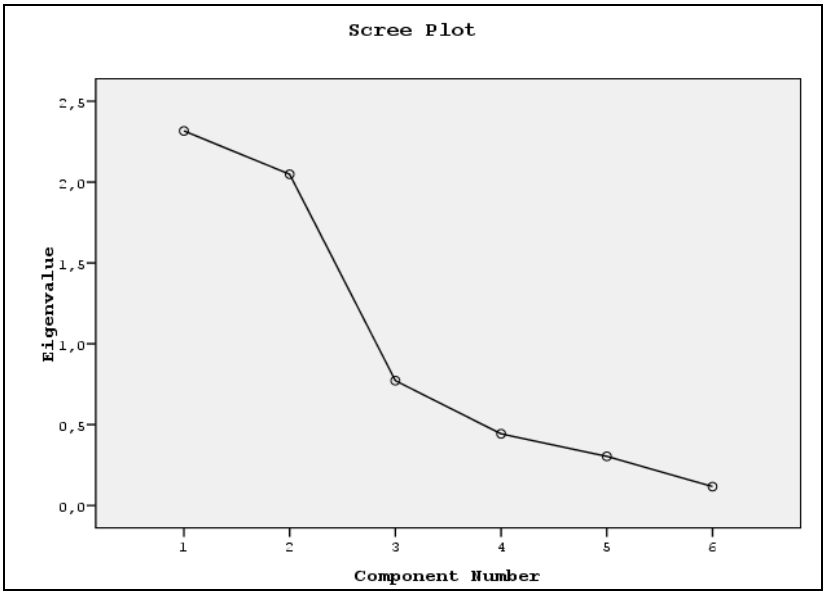
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,316	38,603	38,603	2,316	38,603	38,603	2,291	38,182	38,182
2	2,049	34,142	72,745	2,049	34,142	72,745	2,074	34,563	72,745
3	,772	12,865	85,610						
4	,443	7,384	92,994						
5	,304	5,062	98,057						
6	,117	1,943	100,000						

Extraction Method: Principal Component Analysis.

8.7 pav. **Faktorinės analizės rezultatas**

Lentelėje *Total Variance Explained* pateikiami standartiniai pagrindinių komponenčių (veiksnių) analizės rezultatai (žr. 8.7 pav.). Įvedamos naujos komponentės paaiškina kaskart vis mažiau ir mažiau dispersijos. Visa matricos dispersija priklauso nuo kintamųjų skaičiaus. Pavyzdyje nagrinėjami šeši kintamieji, taigi visa dispersija gali būti išgauta per šešis kartus. Stulpelyje *Initial Eigenvalues* pateiktos tikrinės reikšmės (*Total*) parodo kiekvieno vertinamo veiksnio dispersiją. Stulpelyje *% of Variance* yra apskaičiuoti bendros dispersijos procentai. Matome, kad pirmojo veiksnio tikrinė reikšmė yra 2,316. Kadangi ši reikšmė didesnė už 1,0, tai reiškia, kad veiksnys paaiškina daugiau dispersijos nei atskiras kintamasis, jei tiksliau, 2,316 karto daugiau. Skaičiuojant procentinę išraišką, gautume, kad pirmasis veiksnys paaiškina 2,316/6 = 0,38603 arba 38,603 procentus dispersijos. Antrojo veiksnio tikrinė

reikšmė yra 2,049 ir ji irgi didesnė už vienetą. Kitų veiksnių tikrinės reikšmės mažesnės už vienetą ir jie paaiškina mažiau dispersijos nei atskiri kintamieji. Stulpelyje *Cumulative %* pateikta sukauptoji dispersija. Pirmieji du veiksniai paaiškina 72,745 procentus visos šešių veiksnių dispersijos. Reikia nutarti, kokį veiksnių skaičių reiktų nagrinėti. Iš esmės tai sutartinis pasirinkimas, tačiau pagal *Kaizerio* kriterijų, siūloma nagrinėti tuos veiksnius, kurių tikrinės vertės viršija vienetą. Pagal *Ketelo scree* kriterijų (žr. 8.8 pav.), reiktų žiūrėti, kur tikrinių verčių lygis nusistovi. Abiem atvejais priimtiniausias veiksnių skaičius – du veiksniai.



8.8 pav. **Tikrinių verčių diagrama (*Cattell scree* kriterijus)**

Lentelėje *Component Matrix* (žr. 8.9 pav.) pateikti nepasuktų veiksnių svoriai, t. y. kintamųjų ir veiksnių tarpusavio ryšių stiprumas bei kryptis. Pirmasis kintamasis vidutiniškai koreliuoja su antruoju veiksniu, tačiau nekoreliuoja su pirmuoju veiksniu. Antrasis kintamasis labai gerai koreliuoja su antruoju veiksniu, tačiau nekoreliuoja su pirmuoju veiksniu ir t. t. Lentelėje *Rotated Component Matrix* pateikti veiksnių

svoriai, gauti pritaikius *Varimax* sukimo metodą. Galima pasakyti, kad šiuo atveju koreliaciniai ryšiai išryškinti geriau. Po sukimo matome, kad veiksnys F_1 turi stiprų ryšį su vidutine kaina, kurią respondentas sutinka mokėti už kosmetiką (*vid_kaina_kosmetika*), vidutine kaina, kurią respondentas sutinka mokėti už aprangą (*vid_kaina_apranga*), vidutine kaina, kurią respondentas sutinka mokėti už buitinę techniką (*vid_kaina_techn*). Veiksnys F_2 turi stiprų ryšį su respondentų amžiumi (a*mžius*), respondentų pajamomis per mėnesį (*Pajamos*), suma, skirta per mėnesį pirkiniams (*Pirkinių_suma*). Veiksnį F_1 galima apibūdinti kaip *tinkamų kainų pasirinkimą*, o veiksnį F_2 – kaip *asmenines respondento savybes.*

Component Matrix[a]

	Component	
	1	2
amzius	,232	,645
Pajamos	,223	,848
Pirkiniu_suma	,344	,796
vid_kaina_kosmetika	,889	-,322
vid_kaina_apranga	,828	-,419
vid_kaina_buit_tech	,786	,026

Extraction Method: Principal Component Analysis.
a. 2 components extracted.

Rotated Component Matrix[a]

	Component	
	1	2
amzius	,023	,685
Pajamos	-,048	,875
Pirkiniu_suma	,083	,864
vid_kaina_kosmetika	,945	-,034
vid_kaina_apranga	,917	-,144
vid_kaina_buit_tech	,740	,266

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.

8.9 pav. **Nepasuktų ir pasuktų veiksnių svorių matrica**

Žinant veiksnių svorius, galima apskaičiuoti, kiek kiekvieno kintamojo dispersijos paaiškina parinkti veiksniai. Tuo tikslu veiksnių svoriai pakeliami kvadratu ir sudedami. Tai ir vadinama kintamojo bendrumu (*communality*). Kintamųjų bendrumai pateikiami atskiroje matricoje *Communalities* (žr. 8.10 pav.). Du veiksniai paaiškina skirtingą kiekvieno kintamojo dispersijos kiekį (*Extraction*).

Esant dviem veiksniams veiksnių ir kintamųjų ryšį gerai iliustruoja veiksnių – kintamųjų žemėlapis (žr. 8.11 pav.).

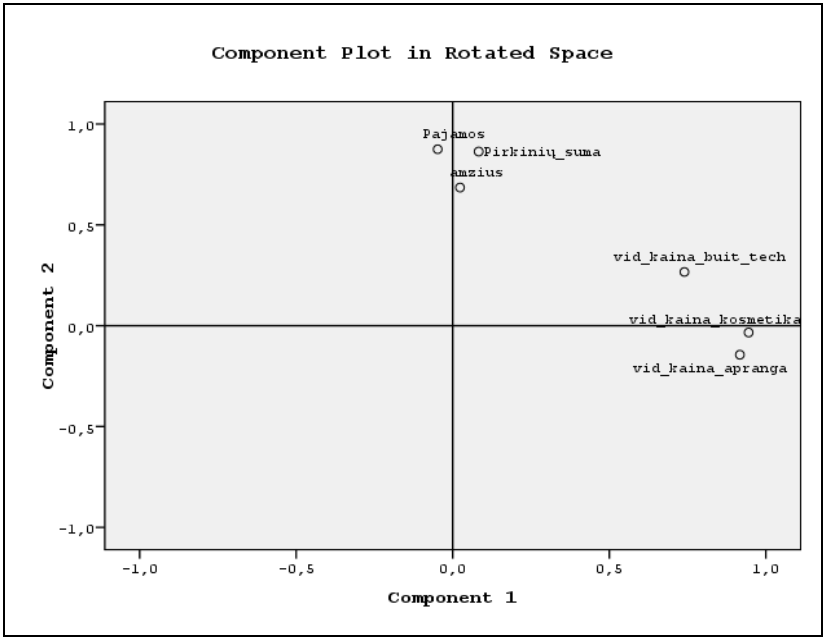
Lentelėje *Component Score Coefficient Matrix* pateikti veiksnių reikšmių skaičiavimo koeficientai (žr. 8.12 pav.).

Communalities

	Initial	Extraction
amžius	1,000	,470
Pajamos	1,000	,768
Pirkinių_suma	1,000	,753
vid_kaina_kosmetika	1,000	,894
vid_kaina_apranga	1,000	,862
vid_kaina_buit_tech	1,000	,618

Extraction Method: Principal Component Analysis.

8.10 pav. **Dviem veiksniais paaiškinta bendros dispersijos dalis**



8.11 pav. **Faktorinės analizės rezultatai**

Component Score Coefficient Matrix

	Component	
	1	2
amzius	-,001	,330
Pajamos	-,035	,423
Pirkiniu_suma	,022	,416
vid_kaina_kosmetika	,414	-,032
vid_kaina_apranga	,403	-,085
vid_kaina_buit_tech	,319	,116

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

8.12 pav. **Faktorinės analizės rezultatai**

KONTROLINIAI KLAUSIMAI

8.1. Kokiais atvejais taikoma faktorinė analizė?
8.2. Kokie uždaviniai sprendžiami faktorinės analizės metodais?
8.3. Koks faktorinės analizės tikslas?
8.4. Kokie kintamieji vadinami veiksniais?
8.5. Koks kintamųjų ir stebėjimų skaičius tinkamas faktorinei analizei?
8.6. Išvardykite faktorinės analizės prielaidas.
8.7. Kaip nustatomas veiksnių skaičius?
8.8. Kam reikalingi sukimo metodai?
8.9. Kaip nustatoma, ar duomenys tinka faktorinei analizei?

UŽDUOTYS

8.1 užduotis

Atliekant faktorinę analizę gautas *KMO* ir *Bartleto* sferiškumo kriterijų skaičiavimo rezultatas (žr. 8.13 pav.). Ar faktorinė analizė turi prasmę?

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,383
Bartlett's Test of Sphericity	Approx. Chi-Square	176,580
	df	21
	Sig.	,000

8.13 pav. **Faktorinės analizės rezultatas**

8.2 užduotis

Atliekant faktorinę analizę gautas pagrindinių komponenčių (veiksnių) analizės rezultatas (žr. 8.14 pav.). Nustatykite, kiek veiksnių reikia pasirinkti ir kokią bendros dispersijos dalį jie paaiškina.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,380	34,000	34,000	2,380	34,000	34,000
2	1,519	21,707	55,707	1,519	21,707	55,707
3	1,132	16,173	71,880	1,132	16,173	71,880
4	,887	12,677	84,557			
5	,613	8,756	93,313			
6	,423	6,046	99,359			
7	,045	,641	100,000			

Extraction Method: Principal Component Analysis.

8.14 pav. **Faktorinės analizės rezultatas**

8.3 užduotis

Atliekant gyvybės draudimo paslaugų tyrimą faktorinei analizei buvo pasirinkti kintamieji: *draudimo sutarties laikotarpis*, *mėnesinė draudimo įmoka*, *gyvybės draudimo suma*, *pajamos per mėnesį*, *apdraustų šeimos narių skaičius.* Analizės rezultatai pateikti 8.15, 8.16 ir 8.17 pav. Paaiškinkite gautus rezultatus.

KMO and Bartlett's test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,535
Bartlett's Test of Sphericity	Approx. Chi-Square	18,980
	df	10
	Sig.	,041

8.15 pav. **Faktorinės analizės rezultatas**

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	1,663	33,254	33,254	1,663	33,254	33,254	1,566	31,329	31,329
2	1,203	24,068	57,322	1,203	24,068	57,322	1,300	25,994	57,322
3	,841	16,827	74,149						
4	,797	15,941	90,090						
5	,496	9,910	100,000						

Extraction Method: Principal Component Analysis.

8.16 pav. **Faktorinės analizės rezultatas**

	Component	
	1	2
draudimo sutarties laikotarpis	,554	,645
mėnesinė draudimo įmoka	,823	,169
gyvybės draudimo suma	,477	-,542
pajamos per mėnesį	-,300	,679
apdraustų šeimos narių skaičius	,601	-,057

Extraction Method: Principal Component Analysis.
a 2 components extracted.

8.17 pav. **Faktorinės analizės rezultatas. Pasuktų veiksnių svorių matrica**

9. KLASTERINĖ ANALIZĖ

Klasterine analize vadinama didelė įvairovė skaitinių duomenų analizės metodų, skirtų objektų grupėms arba *klasteriams*, kurie yra homogeniniai ir skirtingi, nustatyti ir išaiškinti. Daugelis dabartiniu metu taikomų klasterizavimo metodų atsirado dėl psichologų, sociologų, biologų ir matematikų bendro darbo. Sprendžiamų problemų įvairovė nulėmė ir klasterizacijos metodų gausą. Pastaraisiais metais klasterinė analizė tapo sudėtine duomenų gavybos technologijų dalimi ir kartu su kitais metodais taikoma nustatant grupes su panašiais šablonais.

Klasterinės analizės metodai sėkmingai taikomi analizuojant socialines ekonomines problemas. Kai požymių daug, derinami klasterinės ir faktorinės analizės metodai.

Panašumo ir skirtingumo matai

Pagrindinės klasterinės analizės sąvokos yra ***panašumas*** ir ***skirtingumas*** (atstumas): atstumas nurodo, kiek objektai yra nutolę vienas nuo kito (skirtingi), o panašumas rodo objektų artumą. Panašūs objektai priklauso tam pačiam klasteriui, nutolę objektai – skirtingiems klasteriams.

Esminis klasterinės analizės momentas yra objektų artumo mato parinkimas, nes būtent šis matas lemia galutinį objektų suskirstymą į grupes. Kiekvienu konkrečiu atveju pasirenkama atsižvelgiant į tyrimo tikslą, į tiriamų duomenų fizinę prigimtį ir jų savybes. Artumo (panašumo) matai parenkami įvertinant duomenų tipą. Populiariausi yra šie: Euklido, Euklido atstumo kvadratas, Pirsono koreliacijos koeficiento, Čebyšovo, ilgio skirtumo ir kiti.

Panašumo matai. Kai objektus apibūdinantys požymiai matuojami pagal kiekybinę (*intervalų arba santykių*) skalę, taikomi metriniai atstumo matai. Šiuos matus būtų galima vadinti atvirkštiniais matais – kuo didesnė reikšmė, tuo objektai mažiau panašūs. Nevienodai skirtingose skalėse išmatuotų požymių įtakai išvengti taikomos šių kintamųjų standartizuotosios reikšmės (z reikšmės). Intervalinių kintamųjų panašumo matas gali būti taip pat koreliacijos koeficientas.

Euklido atstumas. Euklido atstumas tarp taško *x* ir taško *y* – tai trumpiausias atstumas tarp šių taškų. Plokštumoje arba trimatėje erdvėje tai taškus jungianti tiesė.

Euklido atstumo kvadratas. Šiuo matu tiksliau įvertinami dideli skirtumai. Euklido atstumo kvadratas turi būti taikomas pasirinkus centrų, medianos arba *Vordo* klasterizacijos metodą.

Kosinuso panašumo matas. Šio panašumo mato reikšmių diapazonas yra tarp –1 ir +1.

Čebyšovo atstumas. Atstumu tarp dviejų stebėjimų yra laikoma didžiausia absoliuti skirtumo tarp visų šių stebėjimų porų reikšmė

Bloko atstumo matas. Tai absoliučių skirtumų tarp visų stebėjimų porų reikšmių suma.

Pirsono koreliacijos panašumo matas. Netaikomas tuo atveju, kai yra tik du kintamieji.

Stebėjimų dažniais išreikštų kintamųjų atstumo matai

Atstumas tarp dviejų stebėjimų nustatomas lyginant šiems stebėjimams priklausančių kintamųjų pasirodymo dažnius: χ^2 *(Chi square) matas* ir Φ^2 *(Phi square) matas.*

Kai objektus apibūdinantys požymiai matuojami pagal ***binarinę*** skalę, paprastai jie koduojami: 1 – įvykis įvyko, 0 – įvykis neįvyko. Remiantis šiais dažniais yra apskaičiuojama įvairių atstumo ir panašumo matų: *Euklido atstumas*, *Euklido atstumo kvadratas*, *Didumų skirtumas*, *Pavyzdinis skirtumas*, *Sklaida*, *Formos koeficientas*, *Lenso* (*Lance*) *ir Viljamso* (*Williams'o*) *atstumas.*

Skirtingumo matai

Antras svarbus momentas – atstumo tarp objektų grupių parinkimas. Du labiausiai paplitę atstumų tarp grupių matai – *artimiausio ir tolimiausio kaimyno*. Dažnai taikomas ir *Vordo* metodas (*Ward's method*), tačiau taikant šį metodą gaunama daugiau (mažesnių) klasterių, o metodo rezultatams didelę įtaką daro objektai su požymių išskirtimis.

Artimiausio kaimyno matas – tai atstumas tarp klasterių *artimiausių* objektų. Svorio centrų matas skaičiuoja atstumą tarp klasterių centrinių taškų.

Tolimiausio kaimyno matas – atstumas tarp klasterių *tolimiausių* objektų. Vieno ar kito atstumo mato parinkimas turi didelės įtakos objektų grupavimo geometriniam pavidalui. Algoritmai, paremti artimiau-

sio kaimyno matu, duoda gerus rezultatus tais grupavimo atvejais, kai klasteriai pasižymi sudėtinga, pavyzdžiui, grandinine, struktūra. Tolimiausio kaimyno metodas gerai tinka tada, kai požymių erdvėje išskiriamos grupės yra rutulio formos.

Tarpinę vietą užima algoritmai, taikantys *svorio centrų* atstumo matą. Šie algoritmai geriausiai tinka tada, kai klasterių forma yra elipsinio pavidalo.

Taikomi ir kiti atstumo tarp objektų grupių matai: ryšių tarp grupių, *Vordo* ir kiti.

Ryšys tarp grupių. Atstumas tarp klasterių yra lygus atstumo tarp visų galimų stebėjimų porų, kai vienas stebėjimas imamas iš vieno klasterio, o kitas – iš antro, vidurkiui.

Ryšys grupių viduje. Atstumas tarp klasterių yra apskaičiuojamas remiantis visomis abiejų klasterių stebėjimų poromis, tarp jų ir esančiomis klasterio viduje.

Medianos klasterizacija. Metodas panašus į centrų klasterizacijos metodą, tačiau abudu klasteriai turi vienodą svorį (nepriklausantį nuo klasterio stebėjimų skaičiaus).

Vordo metodas. Taikant *Vordo* metodą, imama klasterių pora ir apskaičiuojama požymių vektorių nuokrypių nuo jungtinio klasterio centro kvadratų suma. Analogiškai daroma su kita klasterių pora ir t. t. Jungiami tie klasteriai, kuriuos sujungus nuokrypių suma mažiausia.

Iš visų aprašytųjų metodų dažniausiai taikomas SPSS pakete nustatytasis variantas – *Between-groups linkage* (ryšys tarp grupių). Apskritai išskirtys iškraipo visų klasterizavimo metodų rezultatus, todėl prieš klasterizuojant pageidautina jas pašalinti. Kintamieji taip pat neturi būti koreliuoti, visose grupėse pageidautinas toks pat elementų skaičius.

Automatinio grupavimo savybių kriterijai

Norint pasirinkti konkretų atstumo matą, pradžioje iškeliama hipotezė apie grupavimo tipą. Jeigu ši hipotezė pasirodo esanti neteisinga, tai galima gauti netinkančius ar net neteisingus rezultatus. Todėl tyrėjas, įvertinantis apriorinį neapibrėžtumą, dažniausiai pritaiko daug klasterinės analizės algoritmų ir, kompleksiškai įvertinęs gautus rezultatus, pasirenka kokį nors vieną algoritmą ir gautas išvadas.

Klasterinės analizės metodų įvairovę lemia kriterijų, atspindinčių vienas ar kitas automatinio grupavimo savybes, gausa. Paprasčiausias grupavimo kokybės kriterijus – atstumo tarp klasterių dydis. Tačiau toks kriterijus neatspindi klasterio „tankumo“, t. y. santykinio objektų pasi-

skirstymo tankio klasterio viduje. Todėl kita kriterijų grupė skaičiuoja vidutinius atstumus tarp klasterių viduje esančių objektų. Tačiau visų dažniausiai taikomi algoritmai, kurie vertina klasterių „tankumo“ ir atstumo tarp klasterių santykį. Tai gali būti, pavyzdžiui, atstumų tarp klasterių sumos santykis su atstumų tarp objektų, esančių klasterio viduje, suma. Kitas kriterijus skaičiuoja bendros duomenų dispersijos santykį su tarpklasterinės ir klasterių centrų dispersijos suma.

Klasterinės analizės prielaidos. Nors klasterinės analizės metodų yra daug, daugelio jų skaičiavimo algoritmai sudaryti iš dviejų etapų. Pirmuoju etapu atliekamas pradinis (gali būti dirbtinis arba laisvas) objektų suskirstymas į grupes ir nustatomas matematinis automatinio klasifikavimo kokybės kriterijus. Antruoju etapu objektai kilnojami iš vienos klasės į kitą iki tol, kol kriterijaus savybės nenustoja gerėti.

Klasterinės analizės metodai taikomi įvairių tipų duomenims. Pagal duomenų tipą parenkamas klasifikavimo metodas ir (atstumo) matas. Kai duomenys kiekybiniai ir matuojami intervalų skale, jiems taikomi šie pagrindiniai atstumo matai: *Euklido*, *Euklido atstumo kvadratas*, *kosinusas*, *Pirsono koreliacijos koeficientas*, *Čebyšovo*, *Manheteno* (*Manhattan*) ir *Minkovskio* (*Minkowski*) *atstumo matai.* Binariniams (dvejetainiams) duomenims taikomi *Euklido*, *Euklido atstumo kvadratas*, *ilgio skirtumo*, *formos koeficientas* ir kiti. Jeigu dalis kintamųjų yra kokybinių, o dalis kiekybinių (binarinių ar ranginių), tai, prieš taikant klasterinę analizę, rekomenduojama kiekybinius duomenis transformuoti į kokybinę rangų skalę. Kintamieji turi būti nepriklausomi.

Nors kintamųjų standartizuoti nereikalaujama, ši procedūra dažnai rekomenduojama tam, kad visi kintamieji turėtų vienodą įtaką atstumui skaičiuoti. Kai kurie kompiuteriuose naudojami skaičiavimo algoritmai standartizavimo procedūrą taiko pagal nustatytuosius parametrus.

K vidurkių algoritmas taikomas tada, kai turima didelė duomenų aibė (pavyzdžiui, $n > 200$). *K vidurkių* algoritmas yra labai jautrus labai nutolusioms reikšmėms. Prieš taikant šį algoritmą, tokios reikšmės turi būti pašalintos. Kartais ši operacija numatoma skaičiavimo algoritme. Pavyzdžiui, SPSS programoje galima pasirinkti *Outlier Treatment*, ir labai nutolusios reikšmės automatiškai išskiriamos į atskirą klasterį.

Klasterizavimo metodai

Klasterinės analizės metodų įvairovė duomenis leidžia klasifikuoti įvairiais būdais: hierarchiniais; vienas kitą išskiriančiais; persidengiančiais; tikimybiniais.

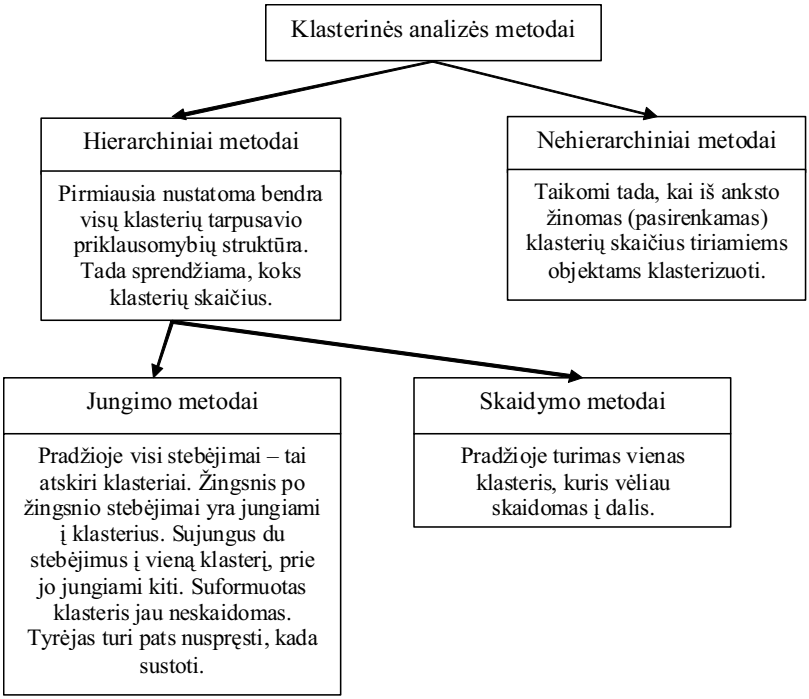
Dvi stambiausios klasterinės analizės metodų klasės – hierarchiniai ir nehierarchiniai metodai.

Kai elementai skirstomi taikant vienas kitą išskiriančius metodus, vienas elementas gali būti priskirtas tik vienam iš klasterių. Klasifikuojant persidengiančiais metodais, priešingai nei išskirtiniais, tas pats elementas gali priklausyti keliems klasteriams. Hierarchiniai klasterizavimo algoritmai yra įgyvendinami nuosekliai jungiant elementus ir naudojantis jau sudarytais klasteriais. Tikimybiniai klasterizavimo metodai remiasi klasterių tikimybių skirstiniais.

Dažniausiai taikomi hierarchiniai ir išskirtiniai klasifikavimo metodai (žr. 9.1 pav.).

Hierarchiniais klasterizavimo metodais grupuojama sudarant klasterių medį. Medis nėra klasterių aibė, tai greičiau daugelio lygių hierarchija, kur vieno lygio klasteriai yra toliau jungiami kitu lygiu. Hierarchiniais metodais nustatoma bendra visų klasterių tarpusavio priklausomybių struktūra ir tik tada sprendžiama, koks klasterių skaičius optimalus. Tai leidžia nutarti, kuris lygis ar klasterizavimo skalė yra priimtiniausia.

Hierarchiniai algoritmai klasterius skaičiuoja nuosekliai, naudodamiesi jau sudarytais klasteriais. Hierarchiniai algoritmai skirstomi į jungimo ir skaidymo. Taikant jungimo metodus pradinį klasterį sudaro vienas elementas, jungiant kitus elementus daromi vis stambesni klasteriai, kol galiausiai lieka vienas klasteris. Skaidymo metodai veikia priešingai: pradžioje visa duomenų aibė sudaro vieną klasterį, o vėliau šis klasteris nuosekliai dalijamas į smulkesnius klasterius. Tiek vienu, tiek kitu metodu gaunama hierarchinė struktūra. Jungimo metodai, vaizduojantys šią struktūrą, pradedami nuo medžio viršūnės, o skaidymo – nuo apačios. Jungimo metodu vadinamas klasifikavimo kriterijus, pagal kurį klasterizavimo algoritmas apskaičiuoja atstumą tarp klasterių ir atskirą elementą priskiria kažkuriam iš klasterių. Skaidymo metodai yra loginė jungimo metodų priešingybė – vienintelis klasteris nuosekliai skaidomas į dalis.



9.1 pav. **Klasterinės analizės metodų klasifikavimas**

Klasterizuojant hierarchiniais metodais, atliekami šie pagrindiniai žingsniai:

1. Kiekvienos duomenų aibės objektų poros panašumui ar skirtingumui įvertinti apskaičiuojami atstumai tarp šių objektų.
2. Pagal pirmuoju žingsniu apskaičiuotą atstumą objektai grupuojami į dvejetainį hierarchinės struktūros medį. Jungiami panašiausi objektai, tarp kurių yra mažiausias atstumas. Kai objektai suporuojami į dvejetainius klasterius, jie vėl grupuojami į didesnius klasterius, kol suformuojamas hierarchinis medis.
3. Nustatomas sutartinis lygis, kuriame hierarchinis medis „nukertamas“, išskaidant jį į „šakas“. Visi objektai, esantys žemiau nukirstos „šakos“, priskiriami vienam klasteriui.

Tradiciškai gauta hierarchinė struktūra vaizduojama medžio pavidalo diagrama (kitaip vadinama *dendrograma*). Nukirtę šį medį kokiame

nors aukštyje, gauname tam tikro tikslumo klasterius. Kuo apytikslesnis klasterizavimas, tuo mažiau klasterių.

Hierarchija pradedama nuo atskirų elementų ir toliau šie elementai jungiami į klasterius. Tuo tikslu reikia nustatyti, kuriuos elementus galima sujungti į klasterį. Dažniausiai imami du arčiausiai esantys elementai, todėl reikia nustatyti atstumą tarp elementų. Atstumas yra skaitinė išraiška, įvertinanti, kaip elementai nutolę vienas nuo kito. Fizikoje atstumu gali būti fizinis ilgis, laiko periodas ir panašiai. Matematikoje taikomi griežtesni ir sudėtingesni atstumo matai. Kadangi klasifikavimo rezultatai priklauso nuo parinkto atstumo mato, jo parinkimas yra labai svarbus klasterinės analizės žingsnis. Klasterinėje analizėje nėra tokio rezultato, kuris galėtų būti pavadintas vieninteliu teisingu klasifikavimu. Vietoj to bandoma vartoti sąvoką *optimalus klasifikavimas*.

Atstumo reikšmė yra labai svarbi, nes pagal ją nustatoma, kada reikia nutraukti klasterių jungimo procedūrą. Jungimo procesą rekomenduojama nutraukti tada, kai atstumas šuoliškai padidėja, nes priešingu atveju į klasterius bus sujungti vienas nuo kito labai nutolę elementai.

Vienas iš hierarchinių klasterinės analizės metodų trūkumų – skaičiuoti naudojama atstumų matrica. Todėl dideliems ($n > 200$) objektų masyvams klasterizuoti dažnai taikomi nehierarchiniai klasterizavimo metodai. Šių algoritmų trūkumas tas, kad būtina iš anksto nurodyti klasterių skaičių.

Optimalaus klasterių skaičiaus problema nėra sprendžiama analitiškai – „teisingas“ skaičius labai priklauso nuo interpretacijos. Koks klasterių skaičius yra optimalus, paprastai sprendžiama vadovaujantis lentele, kuri reiškia atstumą tarp dviejų klasterių pagal pasirinktą atstumo matą. Pasiekus etapą, kur šis atstumas šuoliškai padidėja, sukaupimo į klasterius vyksmą reikėtų laikyti užbaigtą, nes priešingu atveju gali būti sujungti toli vienas nuo kito esantys klasteriai. Kaip minėta, šis suskirstymas yra sąlygiškas.

Hierarchiniai klasterinės analizės metodai. Hierarchiniai jungimo metodai apima didelę metodų grupę. Žinomiausi yra keli hierarchinio jungimo metodai:

- vienetinės jungties arba artimiausio kaimyno (*single linkage*);
- pilnosios jungties arba tolimiausio kaimyno (*complete linkage*);
- vidutinės jungties (*average linkage*);
- Vordo jungties (*Ward's linkage*);

- svertinis vidutinės jungties (*weighted average linkage*);
- centroidų jungties (*centroid linkage*);
- medianų jungties (*median linkage*).

Metodo pavadinimas priklauso nuo atstumo tarp klasterių skaičiavimo būdo. Kai atstumas yra vienetinės jungties, jungimo metodas vadinamas vienetinės jungties metodu ir panašiai. Paprasčiausi ir dažniausiai taikomi yra šie: *vidutinės jungties*, *pilnosios jungties ir vienetinės jungties*. Kadangi bendruoju atveju klasterį sudaro ne vienas elementas, o jų aibė, atstumą tarp klasterių galima apskaičiuoti įvairiais būdais. Vidutinės jungties metodas skaičiuoja visų galimų klasterių porų vidutinį atstumą

$(d(r,s) = \frac{1}{n_r n_s}\sum_{i=1}^{n_r}\sum_{j=1}^{n_s} dist(x_{r_i}, x_{s_j}), i \in (1,\text{K}\ ,n_r), j \in (1,\text{K}\ ,n_s))$, pilnosios jungties skaičiuoja atstumą tarp labiausiai nutolusių taškų $(d(r,s) = \max(dist(x_{r_i}, x_{s_j}))$, vienetinės jungties – tarp arčiausių taškų $(d(r,s) = \min(dist(x_{r_i}, x_{s_j}))$.

Nehierarchinius jungimo metodus sudaro du pagrindiniai metodai:

- *k vidurkių* algoritmas;
- *k medianų* algoritmas.

Dažniausiai taikomas *k vidurkių* algoritmas. Tai vienas iš paprasčiausių klasterizavimo algoritmų, priskiriamų vienas kitą išskiriančių algoritmų klasei. Taikant šį algoritmą yra minimizuojama tikslo funkcija *O*:

$$O = \sum_{i=1}^{k}\sum_{j=1}^{n}\left\|x_j^{(i)} - c_i\right\|^2, \quad (9.1)$$

čia $\left\|x_j^{(i)} - c_i\right\|^2$ yra atstumas tarp elementų $x_j^{(i)}$ ir klasterio centro c_j.

Taikant *k vidurkių* algoritmą klasterių skaičius turi būti žinomas iš anksto. Klasterių skaičiui nustatyti galima taikyti *scree* testą, kuris taikomas faktorinei analizei, nustatant veiksnių skaičių. Apskaičiuotos fak-

torių tikrinės vertės pateikiamos diagramos pavidalu ir padeda nustatyti optimalų veiksnių skaičių.

Šio metodo silpnoji savybė – pirmasis žingsnis arba pradinių klasterių nustatymas. Skirtingai sudarius pradinius klasterius, gaunami skirtingi rezultatai. Stengiamasi juos sudaryti taip, kad pradiniai klasteriai būtų kuo labiausiai nutolę vienas nuo kito.

Skaičiavimo algoritmą sudaro šie žingsniai:

1. Parenkamas klasterių skaičius ir jų centrai.
2. Kiekvienas elementas priskiriamas klasteriui, kurio centras yra arčiausiai.
3. Kai visi elementai priskirti, iš naujo apskaičiuojami klasterių centrai.
4. Antrasis ir trečiasis žingsnis kartojamas tol, kol klasterių centrai nebesikeičia.

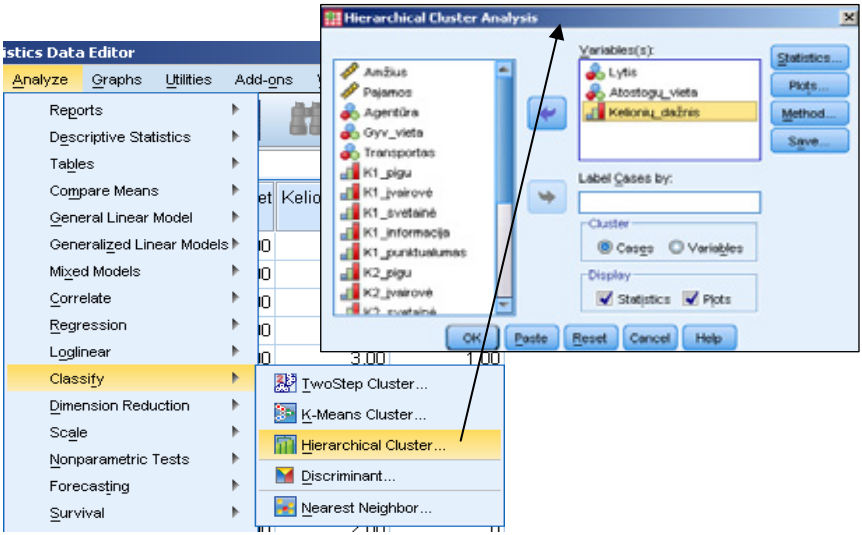
Nors aprašytas klasterizavimo procesas yra visada baigtinis, šis algoritmas nebūtinai randa globalų tikslo funkcijos minimumą. Algoritmas jautrus pradinių klasterių centrų parinkimui. Kad sumažintume šį efektą, algoritmą galima taikyti keletą kartų.

Faktorinės analizės ryšys su klasterine analize. Faktorinės analizės skirtumas nuo klasterinės analizės tas, kad faktorinė analizė yra taikoma koreliacinei matricai, o klasterinei analizei taikomi kiti panašumo matai. Skirtingai nei faktorinė analizė, šie matai negali įvertinti koreliacinei matricai būdingų savybių, tokių kaip priešingų koreliacijos koeficientų ženklų. Pavyzdžiui, pasikeitęs koreliacijos koeficiento ženklas visai pakeistų klasterinės analizės rezultatą, o faktorinė analizė atpažįsta ženklą ir pasikeičia tik veiksnių svorių ženklai, bet niekas daugiau. Kitas faktorinės analizės metodo pranašumas prieš klasterinę analizę tas, kad faktorinė analizė atpažįsta koreliacinių ryšių savybes.

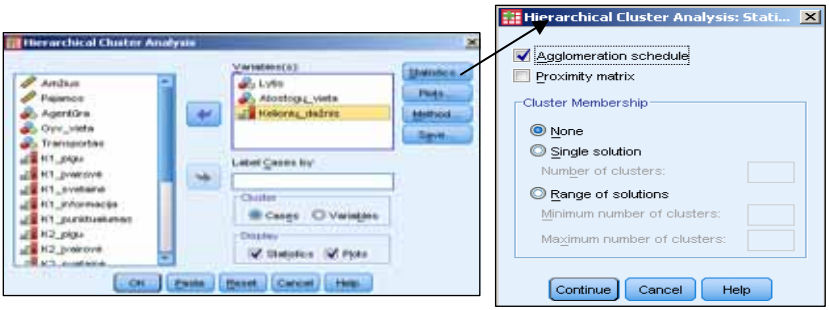
9.1. Klasterinė analizė SPSS terpėje

Hierarchinės klasterinės analizės dialogo langas *Hierarchical Cluster Analysis* atveriamas komanda *Analyze* → *Classify* → *Hierarchical Cluster...* (žr. 9.2 pav.). Kintamieji *Lytis*, *Atostogų_vieta* ir *Kelionių_dažnis* perkeliami į langelį *Variables.*

Mygtuko *Statistics* paspaudimu atveriamas dialogo langas *Hierarchical Cluster Analysis: Statistics.* Šiame lange pažymėtas nustatytasis langelis *Agglomeration schedule* (rodyti jungimosi į klasterius lentelę). Zonoje *Cluster Membership* galima nurodyti klasterių skaičių: *None* – klasterių skaičius parenkamas automatiškai, *Single solution* – nurodomas nustatytasis klasterių skaičius, *Range of solution* – nurodomas minimalusis (*minimum number of clusters*) ir maksimalusis (*maximum number of clusters*) klasterių skaičius. Paspaudę mygtuką *Continue* grįžtame į pradinį langą (žr. 9.3 pav.).



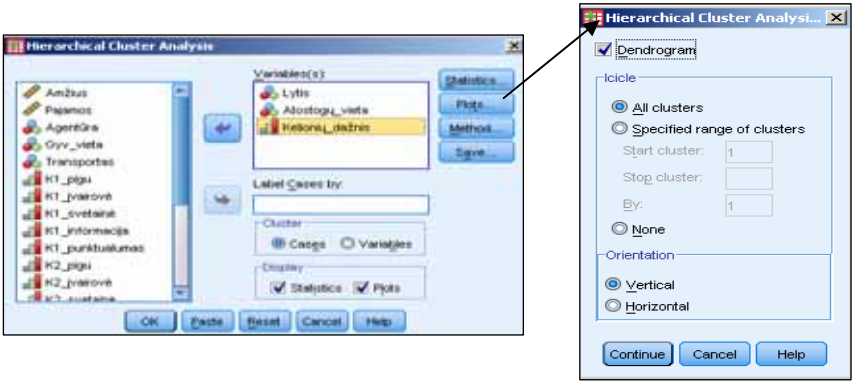
9.2 pav. **Dialogo langas *Hierarchical Cluster Analysis***



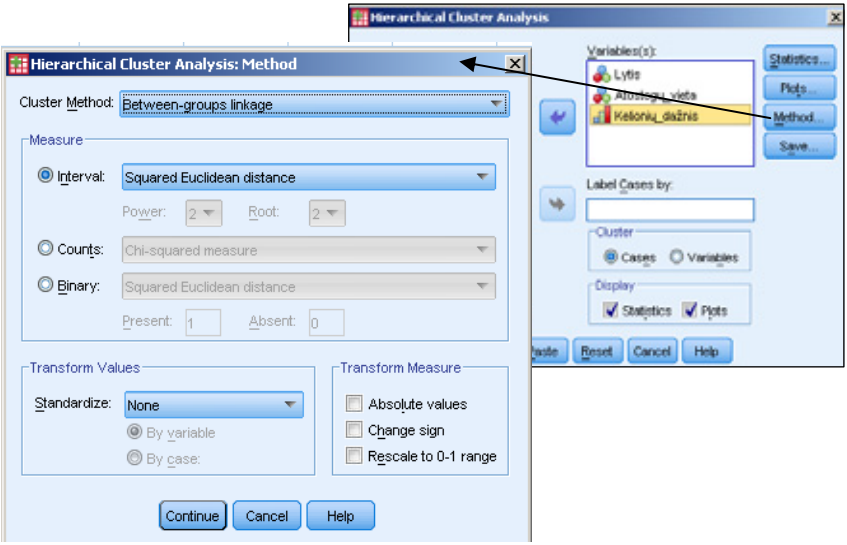
9.3 pav. **Dialogo langas *Hierarchical Cluster Analysis: Statistics***

Mygtuko *Plots…* paspaudimu atveriamas dialogo langas *Hierarchical Cluster Analysis: Plots* (žr. 9.4 pav.). Šiame lange galima pažymėti langelį *Dendrogram* (pateikti klasterių dendrogramą). Paspaudę mygtuką *Continue* grįžtame į pradinį langą.

Mygtuko *Method…* paspaudimu atveriamas dialogo langas *Hierarchical Cluster Analysis: Method* (žr. 9.5 pav.). Šiame lange pasirenkamas klasterizavimo metodas (*Cluster Method*) ir atstumo skaičiavimo būdas (*Interval*). Pagal nustatytuosius parametrus taikomas ryšio tarp grupių metodas (*Between-groups linkage*) ir skaičiuojamas Euklido atstumo kvadratas $\|X-Y\| = \sum_{i=1}^{n} (x_i - y_i)^2$ (*Squared Euclidean distance*). Paspaudę mygtuką *Continue* grįžtame į pradinį langą.



9.4 pav. **Dialogo langas *Hierarchical Cluster Analysis: Plots***



9.5 pav. **Dialogo langas *Hierarchical Cluster Analysis: Method***

Procedūra baigiama mygtuko *OK* paspaudimu. Lange *Output-SPSS Viewer* rodomi analizės rezultatai. Lentelėje *Agglomeration Schedule* rodomas klasterių jungimo procesas (žr. 9.6 pav.). Šiame pavyzdyje buvo panaudota 19 elementų ir jie sujungti per 18 žingsnių. Pirmasis stulpelis (*Stage*) – klasterizavimo žingsniai. Antrasis ir trečiasis stulpeliai (bendru pavadinimu *Cluster Combined*) rodo, kurie elementai buvo jungiami. Lentelės pirmoje eilutėje matome, kad buvo jungiami 8 ir 19 elementai, antruoju žingsniu sujungti 7 ir 14 elementai. Stulpelyje *Coefficients* rodomi atstumai tarp jungiamų elementų. Atstumo reikšmė yra labai svarbi, nes pagal ją nustatoma, kada reikia nutraukti klasterių jungimo procedūrą. Jungimo procesą rekomenduojama nutraukti tada, kai atstumas šuoliškai padidėja, nes priešingu atveju į klasterius bus sujungti vienas nuo kito labai nutolę elementai. Pavaizduotu atveju tai turėtų būti 8 žingsnis. Trys dešinieji stulpeliai rodo, ar elementai jau buvo prieš tai jungti į klasterius (*Stage Cluster First Appears*) ir kada bus kitas jungimas (*Next Stage*).

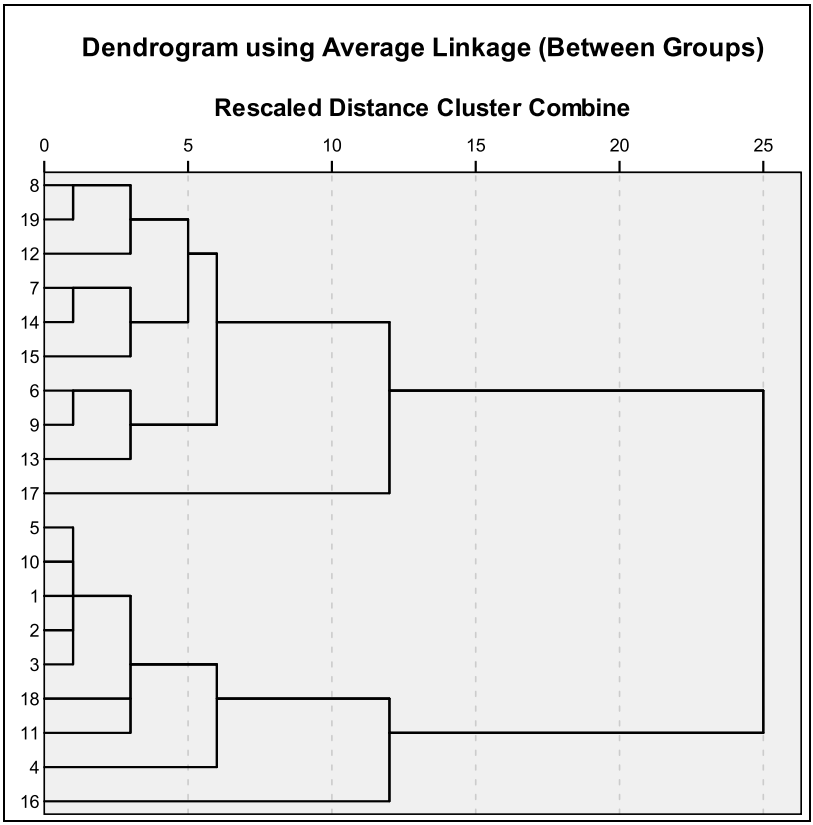
Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	8	19	,000	0	0	11
2	7	14	,000	0	0	9
3	5	10	,000	0	0	5
4	6	9	,000	0	0	10
5	1	5	,000	0	3	7
6	2	3	,000	0	0	7
7	1	2	,000	5	6	8
8	1	18	1,000	7	0	12
9	7	15	1,000	2	0	13
10	6	13	1,000	4	0	14
11	8	12	1,000	1	0	13
12	1	11	1,167	8	0	15
13	7	8	2,000	9	11	14
14	6	7	2,444	10	13	16
15	1	4	2,714	12	0	17
16	6	17	5,222	14	0	18
17	1	16	5,375	15	0	18
18	1	6	11,367	17	16	0

9.6 pav. **Klasterių jungimo rezultatai**

Jungimo procesas grafiškai pateikiamas dendrograma (žr. 9.7 pav.). Klasterinės analizės metodai leidžia greitai išskirti tikslines vartotojų grupes. Pavyzdžiui, turime kintamuosius *Lytis*, *Atostogų vieta* ir *Kelionių dažnis*. Atlikta hierarchinė klasterinė analizė. Klasterių jungimo dendrograma (žr. 9.7 pav.) leidžia išskirti du klasterius. Analizuojant respondentų, priskirtų kiekvienam klasteriui, savybes, galima pasakyti, kad pagrindinis kintamasis, kuris nustato klasterių skirtumą, tai atostogų vietos pasirinkimas. Pirmajam klasteriui priklauso vyrai ir moterys, kurie pasirenka pažintines keliones, antrajam – vyrai ir moterys, kurie pasirenka poilsio keliones. Kelionių dažnis reikšmės neturi.

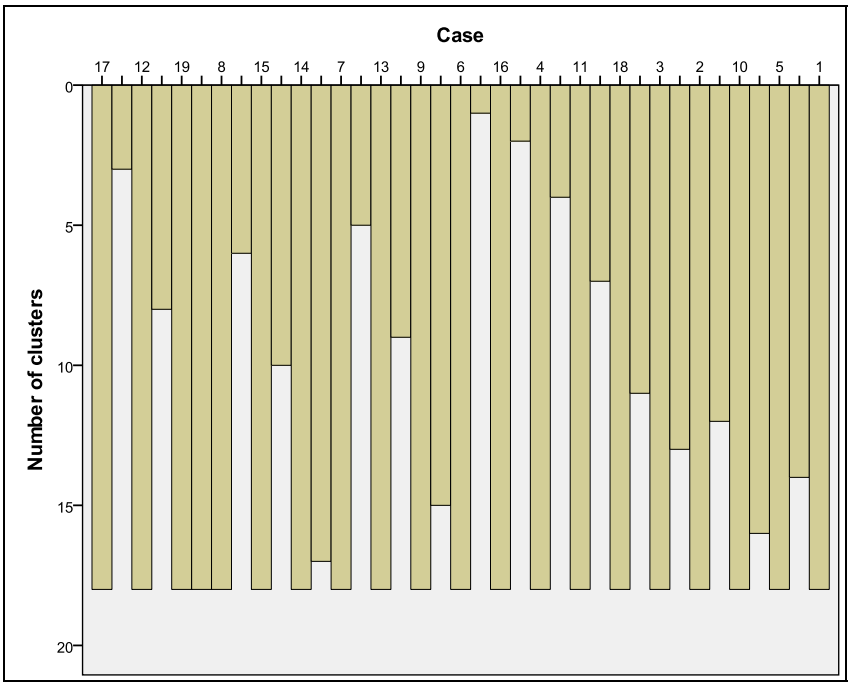
Elementų skaidymo į klasterius procesą rodo varveklinė diagrama (žr. 9.8 pav.). Pradžioje visi respondentai sudaro vieną klasterį, paskui išskiriamos respondentų grupės, kol kiekvienas respondentas sudarys atskirą klasterį. Horizontalioje ašyje pateikiami respondentų numeriai, vertikalioje – klasterų skaičius.



9.7 pav. **Klasterių jungimo dendrograma**

Atliktos procedūros sintaksė:
/SCATTERPLOT(BIVAR)=Pajamos WITH Pirkinių_suma
/MISSING=LISTWISE.

Jei dialogo lange *Hierarchical Cluster Analysis: Statistics* nurodyti klasterių skaičiai (pavyzdžiui, nuo 2 iki 4), tai lentelėje *Cluster Membership* parodoma, kaip elementai išsidėstė klasteriuose (žr. 9.9 pav.). Pavyzdžiui, pasirenkant du klasterius, į pirmąjį klasterį sujungti 1, 2, 3, 4, 5, 10, 11, 16 ir 18 elementai, o į antrąjį klasterį – 6, 7, 8, 9, 12, 13, 14, 15, 17 ir 19 elementai. Pasirenkant tris klasterius, į atskirą trečiąjį klasterį išskirtas 16 elementas.



9.8 pav. **Varveklinė diagrama**

Cluster Membership

Case	4 Clusters	3 Clusters	2 Clusters
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	1
6	2	2	2
7	2	2	2
8	2	2	2
9	2	2	2
10	1	1	1
11	1	1	1
12	2	2	2
13	2	2	2
14	2	2	2
15	2	2	2
16	3	3	1
17	4	2	2
18	1	1	1
19	2	2	2

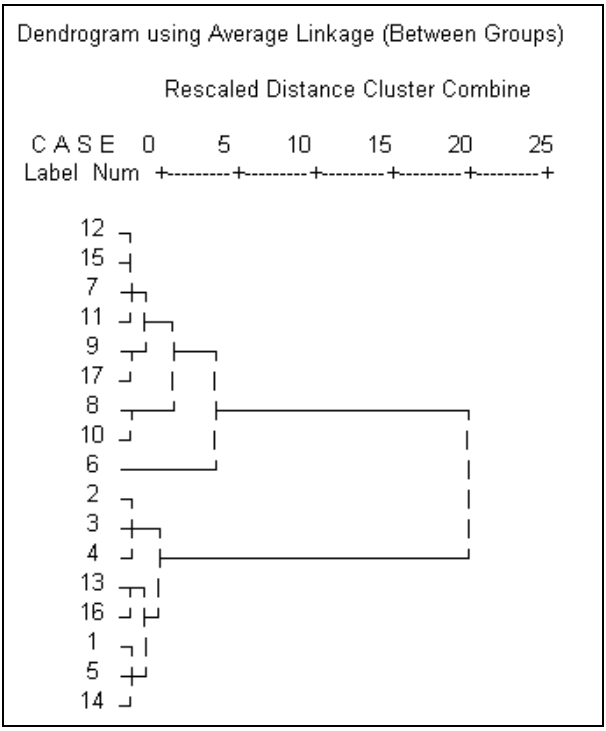
9.9 pav. **Elementų pasiskirstymas klasteriuose**

KONTROLINIAI KLAUSIMAI

9.1. Kokiais atvejais taikoma klasterinė analizė?
9.2. Kokie uždaviniai sprendžiami klasterinės analizės metodais?
9.3. Koks klasterinės analizės tikslas?
9.4. Kokie kintamieji tinka klasterinei analizei?
9.5. Koks hierarchinių ir nehierarchinių klasterizavimo metodų skirtumas?
9.6. Koks faktorinės ir klasterinės analizės ryšys?
9.7. Kaip nustatomas klasterių skaičius?
9.8. Ką rodo dendrograma?

UŽDUOTYS

9.1 užduotis
Atliekant klasterinę analizę gautas skaičiavimo rezultatas (žr. 9.10 pav.). Pasirinkite tinkamą klasterių skaičių.



9.10 pav. **Klasterių jungimo dendrograma**

LITERATŪRA

1. Blunch, N. J. *Introduction to structural equation modeling using SPSS and AMOS*. Los Angeles (Calif.): Sage Publications, 2008, p. 270.
2. Brace, N.; Kemp, R.; Snelgar, R. *SPSS for psychologists: a guide to data analysis using SPSS for Windows:* (*version 12 and 13*). 3rd ed. Basingstoke: Palgrave Macmillan, 2006, p. 450.
3. Buračas, A. *Statistikos terminai: anglų–lietuvių–prancūzų–rusų–vokiečių kalbos žodynas*. Vilnius: Mykolo Romerio universiteto Leidybos centras, 2008, p. 878.
4. Burns, R. A.; Burns, R. B. *Business research method and statistics using SPSS*. Los Angeles (Calif.); London: Sage Publications, 2008, p. 544.
5. Connolly, P. *Quantitative data analysis in education: a critical introduction using SPSS*. London; New York (N.Y.): Routledge: Taylor and Francis Group, 2006, p. 268.
6. Curwin, J.; Slater, R. *Quantitative methods for business decisions*. 6th ed. London: Thomson Learning, 2008, p. 790.
7. Čekanavičius, V.; Murauskas, G. *Statistika ir jos taikymai*. T 1. Vilnius: TEV, 2003, p. 237.
8. Čekanavičius, V.; Murauskas, G. *Statistika ir jos taikymai*. T 2. Vilnius: TEV, 2004, p. 272.
9. Dancey, C. P.; O'Riada, S. *Statistics without maths for psychology: using SPSS for Windows*. 3rd ed. Harlow: Pearson: Prentice Hall, 2004, p. 612.
10. Doane, D.; Seward, L. W. *Applied statistics in business and economics*. 3rd ed. New York (N.Y.): McGraw-Hill: Irwin, 2011, p. 831.
11. Faherty, V. E. *Compassionate statistics: applied quantitative analysis for social services: with exercises and instructions in SPSS*. Los Angeles (Calif.): Sage Publications, 2008, p. 319.
12. Greasley, P. *Quantative data analysis using SPSS: an introduction for health and social science*. Maidenhead: Open University Press: McGrawHill, 2008, p. 138.
13. Howitt, D.; Cramer, D. *Introduction to SPSS in psychology: with supplements for releases 10, 11, 12 and 13*. 3rd ed. Harlow: Pearson: Prentice Hall, 2005, p. 237.
14. Howitt, D.; Cramer, D. *Introduction to SPSS in psychology: for version 16 and earlier*. 4th ed. Harlow: Pearson: Prentice Hall, 2008, p. 377.

15. Krzanowski, W. J. *Statistical principles and techniques in scientific and social investigations*. Oxford; New York (N.Y.): Oxford University Press, 2007, p. 241.
16. Leonavičienė, T. *SPSS programų paketo taikymas statistiniuose tyrimuose*. Vilnius: Vilniaus pedagoginio universiteto leidykla, 2006, p. 125.
17. McClave, J. T.; Benson, P. G.; Sincich, T. *Statistics for business and economics*. 10th ed. Upper Saddle River (N.J.): Pearson: Prentice Hall, 2008, p. 904.
18. Meier, K. J.; Brudney, J. L.; Bohte, J. *Applied statistics for public and nonprofit administration*. 7th ed. Belmont (Calif.): Wadsworth/Thomson Learning, 2009, p. 572.
19. Palalant, J. *SPSS survival manual: a step by step guide to data analysis using SPSS for Windows*. Maidenhead: Open University Press, 2007, p. 335.
20. Rudzkienė, V. *Socialinė statistika*. Vilnius, 2005, p. 260.
21. Rudzkienė, V.; Burinskienė, M. *Plėtros krypčių vertinimo ir valdymo informaciniai modeliai*. Vilnius: Technika, 2007, p. 406.
22. Sarantakos, S. *A toolkit for quantitative data analysis: using SPSS*. Basingstoke: Palgrave Macmillan, 2007, p. 118.
23. Treiman, D. J. *Quantitative data analysis: doing social research to test ideas*. San Francisco (Calif.): Jossey-Bass: A Wiley Imprint, 2009, p. 443.
24. Vaitkevičius, R.; Saudargienė, A. *Statistika su SPSS psichologiniuose tyrimuose*. Kaunas: Vytauto Didžiojo universiteto leidykla, 2006, p. 223.
25. Дуброва, Т. А. *Прогнозирование социально-экономических процессов*. 2-е изд., испр. и доп. Москва: МФПА, 2010, с. 188.
26. *Многомерный статистический анализ в экономических задачах: компьютерное моделирование в SPSS*. Москва, 2009, с. 308.
27. Моосмюллер, Г.; Ребик, Н. *Маркетинговые исследования с SPSS*. Москва: ИНФРА-М, 2009, с. 158.
28. Соколов, Г. А.; Сагитов, Р. В. *Введение в регрессионный анализ и планирование регрессионных экспериментов в экономике*. Москва: ИНФРА-М, 2010, с. 200.
29. *Социальная статистика*. Под редакцией И. И. Елисеевой. Москва: Финансы и статистика, 2003, с. 480.
30. Таганов, Д. Н. *SPSS статистический анализ в маркетинговых исследованиях*. Санкт-Петербург: Питер, 2005, с. 190.

PRIEDAI

1 lentelė. **χ^2 skirstinio α lygmens kritinės reikšmės**

n/α	α=0,10	α=0,05	α=0,025	α=0,01
1	2,71	3,84	5,02	6,63
2	4,61	5,99	7,38	9,21
3	6,25	7,81	9,35	11,34
4	7,78	9,49	11,14	13,28
5	9,24	11,07	12,83	15,09
6	10,64	12,59	14,45	16,81
7	12,02	14,07	16,01	18,48
8	13,36	15,51	17,53	20,09
9	14,68	16,92	19,02	21,67
10	15,99	18,31	20,48	23,21
12	18,55	21,03	23,34	26,22
13	19,81	22,36	24,74	27,69
14	21,06	23,68	26,12	29,14
15	22,31	25,00	27,49	30,58
16	23,54	26,30	28,85	32,00
18	25,99	28,87	31,53	34,81
19	27,20	30,14	32,85	36,19
20	28,41	31,41	34,17	37,57
21	29,62	32,67	35,48	38,93
22	30,81	33,92	36,78	40,29
23	32,01	35,17	38,08	41,64
24	33,20	36,42	39,36	42,98
25	34,38	37,65	40,65	44,31
26	35,56	38,89	41,92	45,64
27	36,74	40,11	43,19	46,96
28	37,92	41,34	44,46	48,28
29	39,09	42,56	45,72	49,59
30	40,26	43,77	46,98	50,89
40	51,80	55,76	59,34	63,69
50	63,17	67,50	71,42	76,15
60	74,40	79,08	83,30	88,38
70	85,53	90,53	95,02	100,42
80	96,58	101,88	106,63	112,33
90	107,56	113,14	118,14	124,12
100	118,50	124,34	129,56	135,81

n – laisvės laipsnių skaičius; α – reikšmingumo lygmuo.

2 lentelė. **Stjudento skirstinio α lygmens kritinės reikšmės $t_\alpha(n)$**

n/α	α=0,40	α=0,10	α=0,05	α=0,025	α=0,01	α=0,005
1	0,325	3,078	6,314	12,706	31,821	63,657
2	0,289	1,886	2,920	4,303	6,965	9,925
3	0,277	1,638	2,353	3,182	4,541	5,841
4	0,271	1,533	2,132	2,776	3,747	4,604
5	0,267	1,476	2,015	2,571	3,365	4,032
6	0,265	1,440	1,943	2,447	3,143	3,707
7	0,263	1,415	1,895	2,365	2,998	3,499
8	0,262	1,397	1,860	2,306	2,896	3,355
9	0,261	1,383	1,833	2,262	2,821	3,250
10	0,260	1,372	1,812	2,228	2,764	3,169
11	0,260	1,363	1,796	2,201	2,718	3,106
12	0,259	1,356	1,782	2,179	2,681	3,055
13	0,259	1,350	1,771	2,160	2,650	3,012
14	0,258	1,345	1,761	2,145	2,624	2,977
15	0,258	1,341	1,753	2,131	2,602	2,947
16	0,258	1,337	1,746	2,120	2,583	2,921
17	0,257	1,333	1,740	2,110	2,567	2,898
18	0,257	1,330	1,734	2,101	2,552	2,878
19	0,257	1,328	1,729	2,093	2,539	2,861
20	0,257	1,325	1,725	2,086	2,528	2,845
21	0,257	1,323	1,721	2,080	2,518	2,831
22	0,256	1,321	1,717	2,074	2,508	2,819
23	0,256	1,319	1,714	2,069	2,500	2,807
24	0,256	1,318	1,711	2,064	2,492	2,797
25	0,256	1,316	1,708	2,060	2,485	2,787
30	0,256	1,310	1,697	2,042	2,457	2,750
40	0,255	1,303	1,684	2,021	2,423	2,704
60	0,254	1,296	1,671	2,000	2,390	2,660
120	0,254	1,289	1,658	1,980	2,358	2,617

n – laisvės laipsnių skaičius;
α – reikšmingumo lygmuo.

ATSAKYMAI

3.1 užduotis

Vidurkis ir mediana padidės 100 vienetų, kitkas nesikeis.

3.2 užduotis

Prieš nustatant medianą, diametro reikšmės nebuvo išdėstytos didėjimo (mažėjimo) tvarka.

4.1 užduotis

Tikrinant respondentų amžiaus skirstinio normalumą gauti Kolmogorovo–Smirnovo (*Kolmogorov-Smirnov*(*a*)) ir Šapiro–Vilko (*Shapiro-Wilk*) kriterijų skaičiavimo rezultatai. Kriterijų reikšmingumo lygmuo *p-level* (*Sig.*) $< \alpha = 0,05$, kintamojo skirstinio skirtumas nuo normaliojo statistiškai reikšmingas. Šiam kintamajam parametrinių kriterijų taikyti negalima.

5.1 užduotis

Hipotezės:

H_0: pirkimo dažnis nepriklauso nuo lyties,

H_1: pirkimo dažnis priklauso nuo lyties.

Lentelėje *Chi-Square Tests* pateikti analizės rezultatai (žr. 5.26 pav.). Užrašas po lentele informuoja, kad kintamieji tinka analizei. *P-level* $= 0,123 > \alpha = 0,05$, hipotezės H_0, kad pirkimo dažnis nepriklauso nuo lyties, atmesti negalima.

Apskaičiavus χ^2 kriterijų nustatoma, ar tarp požymių yra statistinis ryšys. Lentelės dydis 2 x 3. Ryšio stiprumui nustatyti dėl *ranginių* kintamųjų taikomas *Cramer's V* testas. Lentelėje *Symmetric Measures* pateikta šio testo reikšmė (*Cramer's V = 0,264*) ir jo stebimas reikšmingumo lygmuo *p-level* (stulpelis *Approx. Sig.). P-level* $= 0,123 > \alpha = 0,05$, *Cramer's V* koeficientas nereikšmingas.

5.2 užduotis

Hipotezės:

H_0: priimtina jogurto kaina nepriklauso nuo amžiaus,

H_1: priimtina jogurto kaina priklauso nuo amžiaus.

Lentelėje *Chi-Square Tests* pateikti analizės rezultatai (žr. 5.27 pav.). Užrašas po lentele informuoja, kad kintamieji tinka analizei. *P-level* $= 0,000 < \alpha = 0,05$, pasitvirtina hipotezė H_1, kad priimtina jogurto kaina priklauso nuo amžiaus. Eilutė *Linear-by-Linear Association* leidžia daryti išvadą dėl nenominaliųjų kintamųjų tiesinio ryšio, *p-level* $= 0,000 < \alpha = 0,05$, kintamųjų tiesinis ryšys reikšmingas.

Apskaičiavus χ^2 kriterijų nustatoma, ar tarp požymių yra statistinis ryšys. Abu kintamieji yra kiekybiniai. *Kiekybinių* kintamųjų ryšio stiprumui nustatyti taikomas Pirsono koreliacijos koeficientas (*Pearson's R = 0,370*, ryšys silpnas, *p-level = 0,000 < α = 0,05* – koreliacija reikšminga), tačiau kintamųjų skirstinys nežinomas, todėl geriau taikyti Spirmeno koreliacijos koeficientą (*Spearman Correlation = 0,356*, ryšys silpnas, *p-level = 0,000 < α = 0,05* – koreliacija reikšminga).

5.3 užduotis

Lentelės dydis 2 x 5, todėl norint taikyti Pirsono χ^2 suderinamumo kriterijų, tik 25 procentais narvelių dažnis gali būti mažesnis kaip 5. Nagrinėjamu atveju tokių narvelių trys (30 procentų), Pirsono χ^2 suderinamumo kriterijų taikymas duos nepatikimą rezultatą.

5.4 užduotis

Galima apskaičiuoti:

a) amžiaus ir vidutinio pinigų kiekio, išleidžiamo per mėnesį, prekei A Pirsono koreliacijos koeficientą (kintamieji kiekybiniai, duomenų skirstinys yra normalusis);
b) respondentų vidutinių pajamų ir priimtinos prekės A kainos Spirmeno koreliacijos koeficientą (vidutinės pajamos – ranginis kintamasis, priimtina prekės A kaina – kiekybinis kintamasis, tačiau jo skirstinys neatitinka normalumo reikalavimų);
c) respondentų amžiaus ir priimtinos prekės A kainos Spirmeno koreliacijos koeficientą (amžius – kiekybinis kintamasis, priimtina prekės A kaina – kiekybinis kintamasis, tačiau jo skirstinys neatitinka normalumo reikalavimų).

5.5 užduotis

Statistinės hipotezės:

H_0: pajamų ir pirkimo dažnio koreliacijos koeficientas lygus nuliui,

H_1: pajamų ir pirkimo dažnio koreliacijos koeficientas nelygus nuliui.

Spirmeno koreliacijos koeficientas $r_s = -0,050$ rodo nereikšmingą pajamų ir pirkimo dažnio koreliaciją. Stebimasis reikšmingumo lygmuo *p-level = 0,704*, tai daugiau už pasirinktą reikšmingumo lygmenį $\alpha = 0,05$. Hipotezės H_0, kad pajamų ir pirkimo dažnio koreliacijos koeficientas lygus nuliui, atmesti negalima.

5.6 užduotis

Statistinės hipotezės:

H_0: amžiaus ir pinigų, išleistų pirkiniams, koreliacijos koeficientas lygus nuliui,

H_1: amžiaus ir pinigų, išleistų pirkiniams, koreliacijos koeficientas nelygus nuliui.

Pirsono koreliacijos koeficientas $r = 0,604$ rodo vidutinę amžiaus ir pinigų, išleistų pirkiniams, koreliaciją. Stebimasis reikšmingumo lygmuo $p\text{-}level = 0,003$, tai mažiau už pasirinktą reikšmingumo lygmenį $\alpha = 0,05$. Patvirtinta hipotezė H_1: amžiaus ir pinigų, išleistų pirkiniams, koreliacijos koeficientas nelygus nuliui.

6.1 užduotis ir 6.2 užduotis

Determinacijos koeficiento R^2 reikšmės rodo, kad išleidžiamų kavinėje pinigų priklausomybę nuo amžiaus galima aprašyti logaritminės, kvadratinės ir kubinės regresijos modeliu ($R^2 > 0,25$). Geriausiai tinka kubinė regresija:

$y = a + b1x + b2x^2 + b3x^3$ arba $y = -17,866 + 2,711x - 0,054x^2 + 0,0x^3$,

čia y – išleidžiami pinigai kavinėje, x – amžius. Galima pasakyti, kad koeficientas *b3* iš esmės lygus nuliui, tačiau visi kiti koeficientai skiriasi nuo kvadratinio regresijos modelio koeficientų.

Tiesinės regresijos modelio tinkamumo tikrinimas

Suformuluojame hipotezes:

H_0: $\beta_1 = 0$,

H_1: $\beta_1 \neq 0$.

Stebimasis reikšmingumo lygmuo $p\text{-}level = 0,000$, tai mažiau už pasirinktą patikimumo lygmenį $\alpha = 0,05$. Patvirtinta hipotezė H_1, kad $\beta_1 \neq 0$.

6.3 užduotis

Lentelėje *Classification Table* pateikiamos priklausomojo kintamojo stebimos (*Observed*) ir apskaičiuotos, remiantis regresijos modeliu, tikėtinos (*Predicted*) reikšmės. Iš 27 kintamojo *Kelione_lektuvu* reikšmių 0 (*ne*) testu buvo patvirtintos 21, o kitos 6 priskirtos prie 1 (*taip*). Iš 11 kintamojo *Kelione_lektuvu* reikšmių 1 (*taip*) buvo patvirtintos 8, o kitos 3 priskirtos prie 0 (*ne*). Bendras teisingo atpažinimo procentas yra 76,3 procentai. Nėra griežtų taisyklių, nusakančių, koks teisingo klasifikavimo procentas laikytinas patenkinamu. Paprastai reikalaujama, kad kiekvienai kategorijai jis būtų ne mažesnis kaip 50 procentų, o logistinė

regresija apskritai taikoma tik tuo atveju, kai $y_i = 0$ (kintamojo *y* reikšmės) sudaro ne mažiau kaip 20 procentų ir ne daugiau kaip 80 procentų visų stebėjimų.

Lentelėje *Omnibus Tests of Model Coefficients* pateiktas Pirsono χ^2 modelio suderinamumo kriterijus, kuris yra regresijos *ANOVA* analogas. Formuluojamos hipotezės:

H_0: $b = 0$,

H_1: $b \neq 0$.

Hipotezė H_0 atmetama, jeigu *p-level* $< \alpha$. Hipotezė H_0 neatmetama, jeigu *p-level* $\geq \alpha$, čia α – reikšmingumo lygmuo. Šio pavyzdžio *p-level* (*Sig.*) = *0,001 < 0,05*, todėl nulinę hipotezę, kad $b = 0$ atmetame.

Logistinė regresija turi net keletą tiesinės regresijos determinacijos koeficientų analogų. Visiems jiems galioja ta pati taisyklė: kuo didesnis koeficientas, tuo geriau logistinė regresija suderinta su duomenimis. *Cox & Snell R Square* pseudodeterminacijos koeficientas = *0,261* ir normuotas šio koeficiento variantas – *Nagelkerke R Square = 0,357* pateikiami lentelėje *Model Summary*. Koeficientai rodo tą dispersijos dalį, kurią paaiškina logistinės regresijos modelis.

Lentelėje *Variables in the Equation* pateikiamos apskaičiuotų koeficientų *a* ir *b* reikšmės (*B* stulpelyje) ir jų nelygybės nuliui reikšmingumas. Voldo (*Wald*) kriterijus yra Stjudento kriterijaus tiesinėje regresijoje analogas, t. y. jis atsako į klausimą, ar konkretus koeficientas lygus nuliui. Formuluojamos hipotezės:

H_0: $b = 0$,

H_1: $b \neq 0$.

Hipotezė H_0 atmetama, jeigu *p-level* $< \alpha$. Hipotezė H_0 neatmetama, jeigu *p-level* $\geq \alpha$, čia α – reikšmingumo lygmuo. Visi šio pavyzdžio regresijos lygties koeficientai reikšmingai nelygūs nuliui, nes *p-level* (*Sig.*) *< 0,05*.

6.4 užduotis

Lentelės *Variables in the Equation Exp*(*B*) stulpelyje pateikiama galimybių santykio reikšmė. Galimybių santykis rodo, kaip kinta Y galimybė įgyti reikšmę 1. Šiuo pavyzdžiu galime teigti, kad kiekvienas papildomas planuojamų išlaidų litas padidina tikimybę keliauti lėktuvu 1,002 karto.

Galima užrašyti regresijos lygtį:

$$\hat{p}' = -3{,}486 + 0{,}002 \cdot planuojamos_išlaidos\,.$$

Pavyzdžiui, kai planuojamos kelionės išlaidos yra 500 litų, tikimybė keliauti lėktuvu apskaičiuojama taip:

$\hat{p}' = -3{,}486 + 0{,}002 \cdot 500 = -2{,}486,$

$\hat{p}\,(30) = exp(-2{,}489)/(1+exp(-2{,}486)) = (0{,}083/1{,}083) = 0{,}07.$

7.1 užduotis

Pirmoje lentelėje *One-Sample Statistics* pateikiamas apskaičiuotas kintamojo vidurkis (*Mean*) ir standartinis nuokrypis (*Std. Deviation*). Lentelė *One-Sample Test* leidžia įvertinti statistiškai reikšmingus skirtumus. Išvada daroma vertinant stebimojo reikšmingumo lygmens *p-level* (*Sig.* (*2 tailed*)) reikšmę. Jei *p-level*>α, tai kintamųjų vidurkių skirtumas statistiškai nereikšmingas. Nagrinėjamu atveju *p-level = 0,686* > α *= 0,05*, skirtumas 2 Lt (stulpelis *Mean Difference*) yra statistiškai reikšmingas. Realiai per vieną kartą išleidžiamų pinigų vidurkis reikšmingai skiriasi nuo 30 Lt sumos, už kurią duoda akcijos lipduką.

7.2 užduotis

Pirmoje lentelėje *Group Statistics* pateikti kintamojo *Pinigai* vidurkiai (*Mean*) respondentų grupėms >= *35 m. ir < 35 m*. Lentelė *Independent Samples Test* leidžia nustatyti statistinius grupių skirtumus. Livyno kriterijus (*Levene's Test*) taikomas hipotezei dėl analizuojamų kintamųjų dispersijų lygybei tikrinti. Jeigu stebimojo reikšmingumo lygmens *p-level* reikšmė *Sig*. stulpelyje *Levene's Test for Equality of Variances* rodo, kad rezultatas statistiškai nereikšmingas (*p-level* > α, kur α – pasirenkamas reikšmingumo lygmuo, pavyzdžiui, α *= 0,05*) (nagrinėjamu atveju *p-level = 0,344*), tai kintamųjų dispersijos statistiškai lygios ir vidurkių skirtumas nustatomas iš eilutės *Equal variances assumed* stulpelio *Sig.* (*2-tailed*). Nagrinėjamu atveju Livyno kriterijus rodo statistinį reikšmingumą. Stulpelio *Sig.* (*2-tailed*) reikšmė *0,602* (*p-level*>α) rodo, kad skirtingo amžiaus respondentų per vieną kartą pirkiniams išleidžiamų pinigų skirtumas statistiškai nereikšmingas.

7.3 užduotis

Lentelėje *Levene's Test of Equality of Error Variances p-level* (stulpelis *Sig.*) rodo *F* kriterijaus reikšmingumą. Jei *p-level = 0,066* > α *= 0,05* (*F* kriterijus nereikšmingas), dispersijos lygios, skirtumų analizei toliau taikomas *Scheffe* kriterijus. Nepriklausomųjų kintamųjų grupių skirtumų reikšmingumą rodo nepriklausomojo kintamojo *Amžius F* kriterijaus *p-level* (žr. stulpelį *Sig.*).

Lentelė *Post Hoc Tests*, *Multiple Comparisons* leidžia patikrinti, kaip viena grupė skiriasi nuo kitos ir į kokią pusę. Reikia žiūrėti rezultatus, kurie atitinka *Livyno* testu nustatytą kriterijų (nagrinėjamu atveju – *Scheffe* kriterijus). *Mean Difference* stulpelio reikšmės rodo grupių vidurkių skirtumą, o *p-level* (*Sig.* stulpelis) reikšmė – šių skirtumų reikšmingumą. Nagrinėjamu atveju skirtumai nereikšmingi, nes *p-level* reikšmės yra didesnės už pasirinktą patikimumo lygmenį $\alpha = 0,05$.

8.1 užduotis

Ar duomenys (kintamieji) tinka faktorinei analizei, tikrinama Kaizerio–Mejerio–Olkino (KMO) matu ir Bartleto sferiškumo kriterijumi. Lentelėje *KMO and Bartlett's Test KMO = 0,383* reikšmė rodo, kad duomenys faktorinei analizei netinka. Bartleto sferiškumo kriterijus (*Bartlett's test of signification*) rodo, ar tarp analizuojamų kintamųjų yra statistiškai reikšmingai koreliuojančių, t. y. tikrina tokią hipotezę:

H_0: visi kintamieji yra nekoreliuoti,

H_1: kintamieji koreliuoti.

Pasirenkame reikšmingumo lygmenį $\alpha = 0,05$, *p-level* $= 0,000 < \alpha$. Patvirtinta hipotezė H_1.

8.2 užduotis

Galima pasirinkti tris veiksnius, kurie paaiškina 71,88 procentus bendros dispersijos.

8.3 užduotis

Ar duomenys (kintamieji) tinka faktorinei analizei, tikrinama Kaizerio–Mejerio–Olkino (KMO) matu ir Bartleto sferiškumo kriterijumi. Lentelėje *KMO and Bartlett's Test KMO = 0,535* reikšmė rodo, kad duomenys faktorinei analizei tinka. Bartleto sferiškumo kriterijus (*Bartlett's test of signification*) rodo, ar tarp analizuojamų kintamųjų yra statistiškai reikšmingai koreliuojančių, t. y. tikrina tokią hipotezę:

H_0: visi kintamieji yra nekoreliuoti,

H_1: kintamieji koreliuoti.

Pasirenkame reikšmingumo lygmenį $\alpha = 0,05$, *p-level* $= 0,041 < \alpha$. Patvirtinta hipotezė H_1.

Galima pasirinkti du veiksnius, kurie paaiškina 57,32 procentus bendros dispersijos.

Pasuktųjų veiksnių svorių matrica rodo, kad veiksnys F_1 turi stiprų ryšį su kintamaisiais *mėnesinė draudimo įmoka* ir *apdraustų šeimos narių skaičius*, o veiksnys F_2 – su kintamaisiais *draudimo sutarties laikotarpis* ir *gyvybės draudimo suma.*

Bilevičienė Tatjana, Jonušauskas Steponas
Bi238 Statistinių metodų taikymas rinkos tyrimuose: vadovėlis. – Vilnius: Mykolo Romerio universiteto Leidybos centras, 2011. – 308 p., iliustr.
Bibliogr.: p. 298, 299.
ISBN 978-9955-19-242-8

Vadovėlyje nagrinėjamas kiekybinių metodų taikymas rinkos tyrimų procesams: duomenų rinkimo, analizės ir interpretavimo metodologija. Glaustai, be sudėtingos matematikos nurodoma, kaip atlikti rinkos tyrimus, kokią statistiką taikyti pateikiant ir interpretuojant gautus rezultatus. Vadovėlyje pateikiami tipinių uždavinių sprendimo būdai, statistinių išvadų pagrįstumo analizės ir rezultatų interpretavimo pavyzdžiai, naudojami studentų tyrimo darbų pavyzdžiai, kuriems spręsti pasirinktas programų paketas SPSS (17 versija). Pats paketas aprašomas tiek, kiek reikia skaitytojams praktikams.

Vadovėlis skirtas Mykolo Romerio universiteto studentams ir dėstytojams, atliekantiems statistinius tyrimus. Pateikta medžiaga gali būti naudojama ir dėstant kitus dalykus, susijusius su statistiniais tyrimais, taip pat atliekant bakalauro bei magistro darbo tyrimus, rašant mokslinius straipsnius ir disertacijas. Dėstoma medžiaga bus naudinga ir kitų aukštųjų mokyklų studentams, ypač būsimiems vadybos, ekonomikos ir valdymo specialistams.

UDK 311:339.13(075.8)

Tatjana Bilevičienė, Steponas Jonušauskas
STATISTINIŲ METODŲ TAIKYMAS RINKOS TYRIMUOSE
Vadovėlis

Redaktorė *Dovilė Vaitkevičienė*
Maketuotoja *Regina Silkovienė*
Viršelio autorė *Stanislava Narkevičiūtė*

SL 585. 2011 01 20. 14,31 leidyb. apsk. l.
Tiražas 100 egz. Užsakymas .
Išleido Mykolo Romerio universiteto Leidybos centras, Ateities g. 20, LT-08303 Vilnius
Tinklalapis internete www.mruni.eu
Elektroninis paštas leidyba@mruni.eu
Spausdino UAB „Baltijos kopija“, Kareivių g. 13 b, LT-09109 Vilnius
Tinklalapis internete www.kopija.lt
El. paštas info@kopija.lt