Kazimieras Pukėnas

KOKYBINIŲ DUOMENŲ ANALIZĖ *SPSS* PROGRAMA

LIETUVOS KŪNO KULTŪROS AKADEMIJA

Kazimieras Pukėnas

KOKYBINIŲ DUOMENŲ ANALIZĖ *SPSS* PROGRAMA

MOKOMOJI KNYGA

KAUNAS
2009

Mokomąją knygą recenzavo

2009-xx-xx Nr.X
Lietuvos kūno kultūros akademijos Senato rekomenduota

© Lietuvos kūno kultūros akademija, 2009

PUKĖNAS, Kazimieras
Kokybinių duomenų analizė SPSS programa: mokomoji knyga / Kazimieras Pukėnas; Lietuvos kūno kultūros akademija. – Kaunas: LKKA, 2009. – 93 p.

ISBN 9955 – 622 – 18 – 0

TURINYS

PRATARMĖ

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

SPSS programinis paketas (angl. – *Statistical Package for the Social Sciences*) – vienas labiausiai paplitusių statistinės informacijos apdorojimo programinių paketų, tinkamų ir pradedančiajam, ir patyrusiam vartotojui. Pagrindinis SPSS programinio paketo privalumas – didelė šiuolaikinių statistinių analizės metodų pasirinktis bei duomenų analizės rezultatų vizualizavimo priemonių (duomenų pateikimo lentelių, diagramų, skirstinių kreivių) įvairovė, lengvai įvaldoma dialoginė sąsaja. SPSS programinis paketas taikomas sociologijoje, psichologijoje, biologijoje, medicinoje, rinkodaroje, kokybės valdymo procese. Anglų kalba SPSS programinio paketo taikymas plačiai aprašytas (Norušis, 2005), (Norušis, 2006), (Bryman, Cramer), tačiau lietuvių kalba tokio pobūdžio literatūros trūksta.

Šios knygos tikslas – supažindinti studentus su populiariausiais kategorinių (vardinės ir rangų skalės) duomenų statistinės analizės metodais, įdiegtais programinio paketo SPSS 17.0 versijoje, gautų rezultatų interpretacija. Knygoje aprašomi tie vienmačiai ir daugiamačiai statistikos metodai ir modeliai, kurie plačiausiai taikomi socialiniuose tyrimuose: požymių dažnių analizė, klausimynų patikimumo analizė, faktorinė analizė, binarinė logistinė regresija, sprendimų medžių sudarymas. Daugiausia dėmesio skiriama šių metodų panaudojimo ypatumams, jų galimybių analizei, gautų rezultatų interpretacijai. Visų metodų taikymas iliustruojamas konkrečiais pavyzdžiais. Knyga orientuota į skaitytoją, išklausiusį matematinės statistikos kursą ir susipažinusi su SPSS pradmeninis – duomenų įvedimu ir redagavimu, duomenų transformavimu, perskaičiavimu ir pan. Skaitytojui glaustai primenamos tik tos matematinės statistikos sąvokos ir teiginiai, kurių žinojimas yra būtinas, norint suvokti taikomų metodų esmę ir sėkmingai dirbti su SPSS programiniu paketu. Išsamesnių žinių apie statistinius tyrimo metodus skaitytojas gali taip pat pasisemti iš puikių statistikos vadovėlių (Čekanavičius, Murauskas, 2000), (Čekanavičius, Murauskas, 2002), (Gonestas, Strielčiūnas, 2003), (Sakalauskas, 2003), skirtų skaitytojams, neturintiems specialaus matematinio parengtumo.

Ši mokomoji knyga pagrindinai skiriama Lietuvos kūno kultūros akademijos ir kitų Lietuvos aukštųjų mokyklų socialinių mokslų srities bakalauro ir magistro studijų studentams, doktorantams, vykdantiems įvairius tyrimus anketinių apklausų metodu.

Autorius nuoširdžiai dėkoja Vilniaus universiteto profesoriui **X. Yyyyyyy** ir Kauno technologijos universiteto profesoriui **Y. Xxxxxxx** už nuodugnų leidinio rankraščio recenzavimą, kritines pastabas ir naudingus patarimus, leidusius reikšmingai pagerinti pateiktą medžiagą. Autorius taip pat dėkoja LKKA redaktorei **Z. Zzzzzzz** už leidinio kalbinį redagavimą.

1. BENDROSIOS ŽINIOS

1.1. KINTAMIEJI

Duomenų analizės metodo parinkimas labai priklauso nuo jų prigimties. Populiacijos, kartu ir imties, elementus vienija tiriamasis požymis (Čekanavičius, Murauskas, 2000), (Gonestas, Strielčiūnas, 2003). Požymio reikšmės, kurios kinta kartu su imties ar populiacijos nariais, vadinamos kintamuoju (*variable*). Kintamieji priklausomai nuo požymio gali būti kokybiniai arba kiekybiniai. Kiekybinio kintamojo reikšmė parodo, kiek tiriamo požymio turi populiacijos elementas, kai tuo tarpu kokybiniai kintamieji nusako dydžius, kurių negalima įvertinti skaičiais. Pavyzdžiui, sporto rezultatai, išreiškiami sekundėmis, metrais, kilogramais ir pan., ekonominiai rodikliai, išreiškiami piniginiais vienetais, ir t. t. yra kiekybiniai kintamieji, o lytis, sporto šaka, geografinis regionas ir pan. – kokybiniai kintamieji. Paprastai, SPSS kokybinio kintamojo reikšmės koduojamos skaitmenimis. Pavyzdžiui, kintamojo "apskritis" reikšmes galima koduoti taip: "Vilniaus" = 1, "Kauno" = 2, "Klaipėdos" = 3 ir pan. Kiekybinius kintamuosius dar galima skirstyti į tolydžiuosius ir diskrečiuosius. Kiekybinis kintamasis yra vadinamas tolydžiuoju, jei jo reikšmių skirtumas gali būti kiek norima mažas. Kiekybinis kintamasis, kurio reikšmių skirtumas yra ne mažesnis už tam tikrą minimalų pokytį, vadinamas diskrečiuoju kintamuoju. Tolydžiųjų kintamųjų pavyzdžiai - ūgis, svoris, laikas; diskrečiųjų kintamųjų – įverčio balas *k*-balėje sistemoje (vertinant sveikais skaičiais), pataikytų baudų skaičius, prisitraukimų prie skersinio skaičius. Kintamasis, kuris gali įgyti tik dvi reikšmes (paprastai 0 arba 1) vadinamas dvireikšmiu (angl. – *dichotomous variable*) arba binariniu. Kintamasis, kurio reikšmės nepriklauso nuo nuo kitų kintamųjų vadinamas nepriklausomuoju (*independent variable*), kintamasis, kurio reikšmės priklauso nuo kito kintamojo – priklausomuoju (*dependent variable*). Pvz., priklausomybėje $y = f(x)$ y priklauso nuo x .

Ši studijų knyga pagrindinai skirta kokybinių duomenų analizei. Taikydami binarinę logistinę regresiją, sudarydami sprendimų medžius, šiai kategorijai priskirsime priklausomus kintamuosius.

1.2. SPSS DUOMENŲ MATAVIMŲ SKALĖS

Atsižvelgiant į matavimo duomenų pobūdį, SPSS naudojamos šios statistinių duomenų matavimo skalės (*Measures*):

- **Vardinė** arba **pavadinimų** (*Nominal*) skalė. Šioje matavimo skalėje kintamojo reikšmes galima klasifikuoti tik kokybiškai – negali būti vykdomi net paprasčiausi palyginamieji vertinimai: "lygu" arba "daugiau – mažiau" Būdingiausi vardinės skalės pavyzdžiai yra lytis, sportininkų komandų sąrašas, sporto šakų sąrašas ir t. t. Skaičiai, kuriais koduojami atskiri objektai ar jų savybės, neturi jokios empirinės reikšmės, tik rodo, kokia čia ypatybė ar objektas. Kintamųjų, priklausančių vardinei skalei (nominaliųjų kintamųjų) apdorojimo galimybės gana ribotos – galima tiktai įvertinti, kurių objektų (savybių) yra daugiau ar mažiau, jų proporcijas, koks bendras visų sąraše esančių objektų kiekis. Pagal vardinius kintamuosius dažniausiai vykdoma kokybinė duomenų klasifikacija arba grupavimas – imtis suskaldoma pagal šių kintamųjų kategorijas. Gautoms dalinėms imtims taikomi vienodi statistiniai testai, jų rezultatai palyginami tarpusavyje.
- **Rangų** (*Ordinal*) skalė. Joje nustatoma objekto (reiškinio) vieta pagal pasirinktą kiekybinį arba kokybinį požymį vienos rūšies objektų (reiškinių) grupėje. Pavyzdžiui, sportininko užimta vieta varžybose, studentų sportinis aktyvumas (pvz.: 1=nesportuoja, 2=sportuoja retkarčiais, 3=sportuoja intensyviai), automobilio klasė

(pvz.: 1=aukščiausios klasės, 2=vidutinės klasės, 3=žemiausios klasės) ir t. t. Šiai skalei priklauso taip pat kintamieji, gauti grupuojant duomenis, pvz. – pagal pinigines įplaukas. Su matuojamais pagal rangų skalę kintamaisiais (ranginiais kintamaisiais) galima atlikti gerokai daugiau statistinių operacijų negu su vardiniais kintamaisiais. Be dažnių įvertinimo, galima apskaičiuoti medianą, rangų koreliacijos koeficientą, palyginti atskiras imtis naudojant neparametrinius testus.

- **Intervalų** (*Scale*) skalė. Šioje skalėje nurodomi kiekybiniai kintamųjų reikšmių skirtumai, išreikšti matavimo vienetais (metrais, sekundėmis, laipsniais, taškais ir pan.). 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. Rangų skalės kintamieji, turintys ne mažiau 15 galimų reikšmių, gali būti laikomi intervaliniais (Yaffee, 2003). Duomenis intervalų skalėje galima apdoroti visais be apribojimų statistikos metodais.
- **Santykių** skalė. Ši skalė skiriasi nuo intervalų skalės tik tuo, kad joje nulinis taškas yra griežtai apibrėžtas ir visiškai atitinka dydžio nebuvimą. SPSS taikymo praktikoje skirtumas tarp kintamųjų, matuojamų pagal intervalų arba santykių skalę, yra neesminis.

Ranginiai ir nominalieji kintamieji vadinami kategoriniais. Sąvokos “kokybinis” ir “kategorinis” vartojamos kaip sinonimai.

1.3. PRIKLAUSOMOS IR NEPRIKLAUSOMOS IMTYS

Dvi imtys priklauso viena nuo kitos, jeigu kiekvienai vienos imties reikšmei galima dėsningai ir vienareikšmiškai nurodyti vieną atitinkamą kitos imties reikšmę. Panašiai apibrėžiama ir kelių imčių priklausomybė.

Dažniausiai pasitaiko priklausomos imtys, kai matavimai atliekami tam tikrais laiko momentais. Tada priklausomos imtys sudaro tiriamo vyksmo parametrų reikšmes skirtingais laiko momentais.

SPSS programoje priklausomos imtys pateikiamos kaip skirtingi kintamieji, kurių analizei taikomas atitinkamas kriterijus toje pačioje stebėjimų visumoje. Nepriklausomos imtys, t. y. imtys, kurioms negalima nustatyti dėsningo ir vienareikšmio atitikimo, gali turėti skirtingus stebėjimus, kuriuos paprastai skiria kategorinis vardinės skalės kintamasis.

1.4. IMTIES SUDARYMO BŪDAI IR IMTIES DYDIS

Imties sudarymą pagrindinai sąlygoja du veiksniai – imtis turi būti reprezentatyvi, t. y., kuo pilniau atspindėti populiaciją, iš kurios ji sudaryta ir užtikrinti mažą imties paklaidą. Imties paklaida suprantama kaip skirtumas tarp populiacijos tam tikros charakteristikos skaitinės reikšmės (parametro) ir šios charakteristikos įverčio, gauto iš imties (statistikos) (Sakalauskas, 2003). Šiems tikslams pasiekti paprastai naudojama tikimybinė atranka, kurios išdavoje kiekvienas populiacijos elementas turi tokias pat galimybes patekti į imtį. Yra didelė tikimybinės atrankos metodų įvairovė. Dažniausiai naudojamos paprasta atsitiktinė, sisteminė atsitiktinė, sluoksninė bei klasterinė atranka (Dattalo, 2008).

Pagal labiausiai žinomą paprastą atsitiktinę atranką populiacijos nariai sunumeruojami, o į imtį atrenkami pagal atsitiktinių skaičių generatorių.

Sisteminė atsitiktinė atranka paremta atsitiktiniu populiacijos narių sąrašu. Imtis sudaroma atsitiktinai pasirinkus pradinį (startinį) sąrašo numerį ir atrenkant narius tam tikru intervalu.

Sluoksninė atsitiktinė atranka naudojama tada, kai norima užtikrinti imtyje kiekvienos grupės (vadinamojo sluoksnio) elementų kiekį, proporcingą sluoksnio dydžiui populiacijos atžvilgiu. Populiacija išskaidoma į nepersidengiančius sluoksnius (vadinąmasias stratas) ir pagal kiekvieną grupę atliekama atsitiktinė atranka. Šis metodas leidžia išvengti situacijos, kai vienos iš tiriamų grupių elementai gali nepatekti į imtį arba patekti neproporcingai, t. y. imtis bus nereprezentatyvi.

Klasterinė atsitiktinė atranka naudojama atsitiktinei imčiai iš labai didelės arba geografiškai įvairiapusės populiacijos sudaryti.

Dėl laiko, kaštų sumetimų taip pat naudojami netikimybinės atrankos metodai. Dažniausiai naudojami tinkamumo (*availability*), sniego gniūžtės (*snowball*) ir kvotos metodai (Dattalo, 2008). Tačiau vykdant netikimybinę atranką, populiacijos elementų patekimo į imtį galimybės yra nežinomos, didelė sisteminės paklaidos galimybė.

Pagal tinkamumo metodą atrenkami tie elementai, kurie yra pasiekiami tyrėjui. Esminis metodo trūkumas – respondentai parenkami apklausos vykdytojo nuožiūra arba tik sutinkantys dalyvauti apklausoje, kas žymiai padidina sisteminių paklaidų atsiradimo tikimybę.

Kvotos metodas yra sluoksninės atrankos netikimybinė versija, kada atrenkami tik elementai, tenkinantys užsiduotų požymių visumą.

Sniego gniūžtės metodas yra naudojamas išplėsti apklausos dalyvių ratą per žinomus tiriamos populiacijos atstovus, t. y. pagal pirminės respondentų grupės atsakymus nustatomi kiti (pirmųjų pažįstami, draugai, partneriai) respondentai, kurie vėliau irgi apklausiami. Procedūra gali būti kartojama.

Imties dydis yra svarbus veiksnys, apsprendžiantis statistinį tikslumą, kuriuo vertinami populiacijos požymiai. Imties dydį esminiai sąlygoja du veiksniai – populiacijos dydis ir imties paklaida. Įvairių duomenų imties dydžiui nustatyti skirta daug darbų – plati apžvalga pateikiama (Dattalo, 2008). Praktiniam apklausų planavimui galima vadovautis 1.1 lentelėje pateikiamu imties dydžiu priklausomai nuo populiacijos dydžio ir imties paklaidos, kuri šiuo atveju reiškia populiacijos proporcijų nustatymo paklaidą (pagal elektroninį išteklių www.library.nhs.uk/nlhdocs/FOLIO13_choosing_a_sample.doc).

1.1 lentelė. Rekomenduojamas imties dydis

Populiacijos dydis	Imties paklaida	Imties paklaida	Imties paklaida
	± 3%	± 5%	± 10%
100	92	80	49
250	203	152	70
500	341	217	81
750	441	254	85
1000	516	278	88
2500	748	333	93
5000	880	357	94
10000	964	370	95
25000	1023	378	96
50000	1045	381	96
100000	1056	383	96
1000000	1066	384	96
100000000	1067	384	96

1.5. STATISTINĖS HIPOTEZĖS IR JŲ TIKRINIMAS

Stastistine hipoteze vadinama bet kokia prielaida apie populiacijos požymio skirstinį ar jo parametrus arba apie kelių populiacijų nepriklausomumą. Iškeltoji hipotezė patvirtinama arba atmetama remiantis imties duomenimis. Tikrinant statistines hipotezes apibrėžiama nulinė hipotezė bei jai alternatyvioji hipotezė. Hipotezė, kurioje teigiama, kad nėra esminių skirtumų tarp lyginamųjų visumų, tarp lyginamųjų rodiklių, vadinama nuline hipoteze H_0. Priešingas nulinei hipotezei teiginys vadinamas alternatyviąja hipoteze H_1.

Hipotezėms tikrinti yra naudojami įvairūs kriterijai. Statistinis kriterijus – tai taisyklė, pagal kurią, remiantis imties duomenimis, hipotezė H_0 priimama arba atmetama. Nesigilindami (plačiau galima rasti: Čekanavičius, Murauskas, 2000, Gonestas, Strielčiūnas, 2003 ir Sakalauskas, 2003), pateiksime statistinio kriterijaus reikšmingumo lygmens, pagal kurį yra atmetama nulinė hipotezė H_0 ir priimama alternatyvioji hipotezė H_1, sąvoką. Tardami, kad hipotezė H_0 yra teisinga, iš imties duomenų sudarome atsitiktinę funkciją (vadinamą statistiką) $S = S(X_1, X_2, \cdots, X_n)$, čia $X_1, X_2, \cdots, X_n$ – suprantama kaip seka iš *n* atsitiktinių dydžių, ir jos reikšmių aibę R_α tokią, kad tikimybė $P(S \in R_\alpha \mid H_0 - teisinga) = \alpha$ būtų maža. Aibė R_α vadinama kritine sritimi, o skaičius α — statistinio kriterijaus reikšmingumo lygmeniu (*significance level*). Jeigu pagal imties duomenis statistikos *S* reikšmė patenka į kritinę sritį, tai hipotezė H_0 yra atmetama su reikšmingumo lygmeniu α ir priimama alternatyvioji hipotezė H_1. Priešingu atveju, teigiama, kad imties duomenys neprieštarauja hipotezei H_0. Arba dar paprasčiau, kriterijaus reikšmingumo lygmenį galima suprasti kaip klaidos atmetus hipotezę H_0, nors iš tikrųjų ji teisinga, tikimybę $\alpha = P(H_0$ *atmetame* $\mid H_0$ *teisinga)*. Dar ši tikimybė vadinama pirmosios rūšies klaida. Tradiciškai reikšmingumo lygmenys yra $\alpha = 0{,}05$; $\alpha = 0{,}01$ ir $\alpha = 0{,}001$. Tikrinant hipotezę H_0 galima taip pat pat priimti hipotezę, nors ji iš tikrųjų yra klaidinga. Tai vadinama antrosios rūšies klaida. Hipotezių sprendinių klasifikacija pateikiama 1.2 lentelėje.

1.2 lentelė. Hipotezių sprendinių vertinimo lentelė

	H_0 teisinga	H_0 klaidinga
atmetame H_0	I rūšies klaida	teisingas sprendimas
neatmetame H_0	teisingas sprendimas	II rūšies klaida

Antrosios rūšies klaida dažnai žymima β, o $1-\beta$ vadinamas kriterijaus galia. Jei taikomi keli kriterijai, labiau patikimi rezultatai gaunami skaičiuojant pagal galingesnius kriterijus. SPSS (kaip beje ir kitose statistinėse programose) hipotezės tikrinamos pagal *p*-reikšmės metodą. Tikimybė, kad kriterijaus statistika *S* (tuo atveju, kai H_0 teisinga) ne mažesnė už konkrečią statistikos *S* realizaciją s^*, vadinama *p*-reikšme $p = P(S \geq s^*)$, kai H_0 teisinga.

Bendra taisyklė, tinkanti visų rušių (su vienpusia ir dvipusia alternatyva) nulinėms hipotezėms H_0 ir alternatyvoms H_1 formuluojama taip:

Tegul α yra reikšmingumo lygmuo, o *p* yra *p*-reikšmė.

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

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

2. SPSS POŽYMIŲ DAŽNIŲ LENTELĖS

Priklausomybės tarp vardinių ir rangų skalės kintamųjų analizei SPSS naudojamos taip vadinamos požymių dažnių lentelės (*Contingency table*, *Crosstabs*) ir didelė testų įvairovė priklausomybės laipsniui tarp nagrinėjamų kintamųjų įvertinti. Pagrindinį dėmesį skirsime Chi-kvadratu (χ^2) kriterijui, kuris yra vienas populiariausių ir plačiausiai taikomų neparametrinių kriterijų. χ^2 kriterijus naudojamas hipotezėms apie kintamojo skirstinį populiacijoje tikrinti (t. y., ar empirinio ir teorinio skirstinių skirtumas yra reikšmingas), dviejų kintamųjų nepriklausomumui (vienoje populiacijoje stebima kintamųjų pora) ir vieno kintamojo homogeniškumui (keliose populiacijose stebimas vienas ir tas pats kintamasis) tikrinti (Sakalauskas, 2003). χ^2 kriterijus yra pagrindinis anketinių apklausų duomenų analizės įrankis.

SPSS χ^2 kriterijus skaičiuojamas trejopai (Garson, 2009): pagal Pirsono (*Pearson*) formulę, pagal tikėtinumo santykio (*Likelihood Ratio*) formulę bei pagal Mantelio-Haenzelio (*Linear-by-Linear*) formulę. Kai duomenys aprašomi keturlauke (2x2) dažnių lentele ir kai nors vienas tikėtinas stebėjimų skaičius mažiau penkių, papildomai skaičiuojamas tikslus Fišerio (*Fisher's*) kriterijus. Plačiausiai naudojama Pirsono formulė χ^2 kriterijaus reikšmei apskaičiuoti:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{2.1}$$

čia O_i – nustatyti dažniai, E_i – tikėtini dažniai, k – bendras kintamųjų kategorijų ir grupių skaičius, lygus *Crosstabs* lentelės eilučių ir stulpelių sandaugai.

Asimptotinio χ^2 kriterijaus rezultatų patikimumas yra sąlygojamas šiais reikalavimais: imties tūris turi būti ne mažesnis kaip 30, keturlaukėse (2x2) dažnių lentelėse tikėtini dažniai turi būti ne mažesni kaip 5, didesnėse lentelėse bent 80% dažnių lentelės ląstelių tikėtini dažniai turi būti ne mažesni kaip 5 ir neturi būti ląstelių su nuliniais tikėtinais dažniais (Garson, 2009).

Tikėtinumo santykio formulė χ^2 kriterijaus reikšmei apskaičiuoti yra ši:

$$\chi^2 = -2 \cdot \sum_{i=1}^{k} O_i \cdot ln\frac{E_i}{O_i}, \tag{2.2}$$

Pagal Pirsono ir tikėtinumo santykio formules gaunami artimi rezultatai, o esant didelėms imtims jie praktiškai sutampa. Mantelio-Haenzelio formulė χ^2 kriterijaus reikšmei apskaičiuoti yra ši:

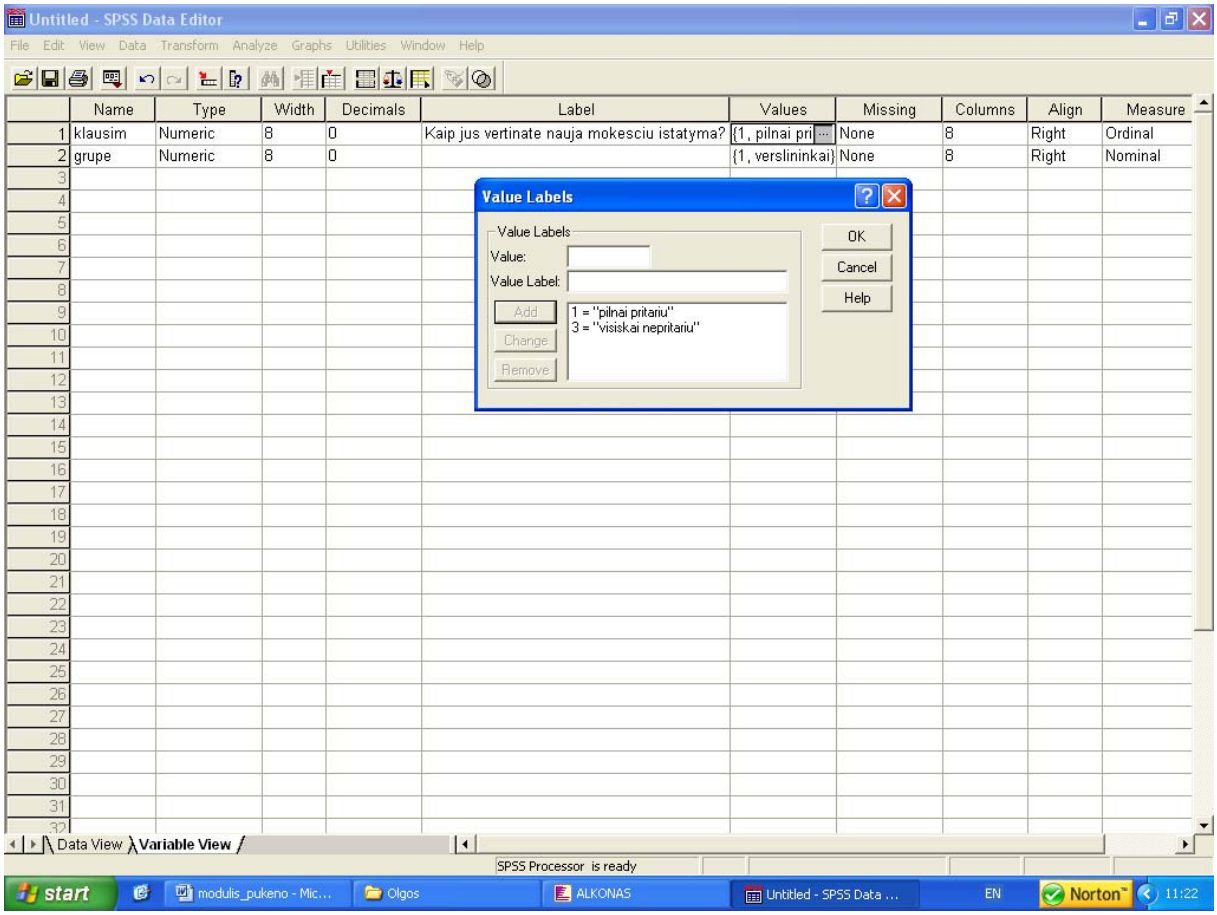
$$\chi^2 = r^2 \cdot (n-1), \tag{2.3}$$

čia r — Pirsono koreliacijos koeficientas, n — imties dydis.

Išvesties lentelėje šis kriterijus nurodomas ***Linear-by-Linear*** pavadinimu. Taikomas hipotezėms apie tiesinį ranginių kintamųjų nepriklausomumą patikrinti, t.y. kai $p < \alpha$ H_0 atmetama ir daroma išvada, kad kintamuosius sieja tiesinė priklausomybė. Netaikomas vardiniams kintamiesiems.

2.1. POŽYMIŲ DAŽNIŲ LENTELIŲ SUDARYMAS

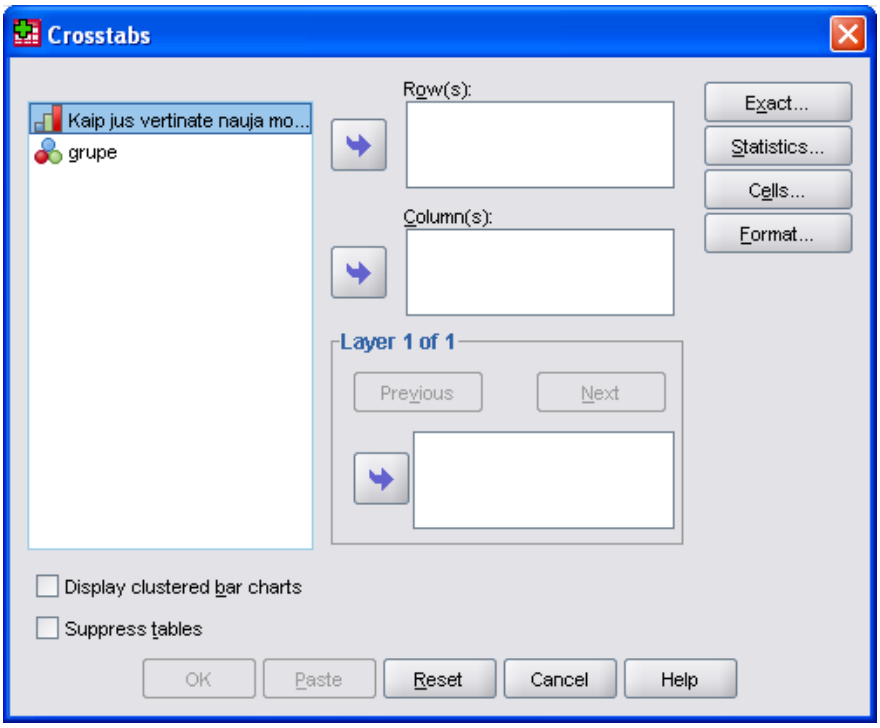
Panagrinėkime pavyzdį vieno kintamojo homogeniškumui (keliose populiacijose stebimas vienas ir tas pats kintamasis) patikrinti. Tiriama, kaip verslininkai ir pramonininkai vertina naujai priimtą mokesčių įstatymą. Parinksime paprastą trijų lygių atsakymų į klausimą "Kaip jūs vertinate naują mokesčių įstatymą?" vertinimo skalę – nuo "pilnai pritariu" iki "visiškai nepritariu". Kintamasis *klausim* reiškia atsakymus į klausimą, o kintamasis *grupe* nurodo respondentų priklausomybę konkrečiai populiacijai – šiuo atveju verslininkų ir pramonininkų. SPSS duomenų rinkmenos langas ***Variable View*** turi atrodyti taip, kaip parodyta 2.1 pav. Apklausus 40 atsitiktinai parinktų verslininkų ir 30 atsitiktinai parinktų pramonininkų, duomenys suvedami duomenų redaktoriaus lange ***Data View***. Statistinė hipotezė H_0 teigia, kad verslininkai ir pramonininkai naują mokesčių įstatymą vertina vienodai, hipotezė H_1 – verslininkai ir pramonininkai naują mokesčių įstatymą vertina nevienodai. Sprendimo priėmimo taisyklė – jeigu χ^2 kriterijaus *p*-reikšmė mažesnė už reikšmingumo lygmenį α $(\alpha = 0{,}05)$, nulinę hipotezę atmetame, t. y. verslininkų ir pramonininkų vertinimai skiriasi statistiškai reikšmingai, jeigu χ^2 kriterijaus *p*-reikšmė didesnė už reikšmingumo lygmenį, nulinės hipotezės atmesti negalima, t. y. verslininkų ir pramonininkų vertinimai nesiskiria.



2.1 pav. Kintamųjų langas ***Variable View***

Turėdami analizuojamus duomenis požymių dažnių lentelę sudarykite taip:

- Nurodykite komandas ***Analyze*** → ***Descriptive Statistics*** →***Crosstabs...*** Atsidarys dialogo langelis ***Crosstabs*** (2.2 pav.).
- Įkelkite į laukelį ***Colum(s)*** atsakymų į klausimą duomenis, o į laukelį ***Row(s)*** – kintamąjį, nurodantį respondentų priklausomybę konkrečiai populiacijai. Galima ir atvirkščiai, bet nurodytas variantas yra įprastas. Į laukelius ***Colum(s)*** ir ***Row(s)*** galima iškart įkelti po kelis kintamuosius, t. y., grupuojant juos pagal populiacijas ir kintamųjų stebėjimus populiacijose. Kiekvienam dviejų kintamųjų deriniui bus sukurta atskira dažnių lentelė. Pavyzdžiui, jeigu ***Row(s)*** sąraše yra du kintamieji, o ***Colum(s)*** sąraše – trys kintamieji, tai gausime 3*2=6 dažnių lenteles.
- Spragtelėkite dialogo langelio ***Crosstabs*** mygtuką ***Statistics...***
- Atsidariusiame naujame dialogo langelyje ***Crosstabs: Statistics*** (2.3 pav.) pasirinkite ***Chi-square*** .
- Spragtelėkite dialogo langelio ***Crosstabs: Statistics*** mygtuką ***Continue***.
- Spragtelėkite dialogo langelio ***Crosstabs*** mygtuką ***Cells...***
- Dialogo langelio ***Crosstabs: Cell Display*** (2.4 pav.) komandų grupėje ***Percentages*** pasirinkite vieną ar kelis šių variantų:
  - ***Row*** (pagal eilutes): procentinės reikšmės skaičiuojamos pagal eilutes, t. y., kiekvienos ląstelės reikšmė atžvilgiu eilutės sumos.
  - ***Column*** (pagal stulpelius): procentinės reikšmės skaičiuojamos pagal stulpelius, t. y., kievienos ląstelės reikšmė atžvilgiu stulpelio sumos.

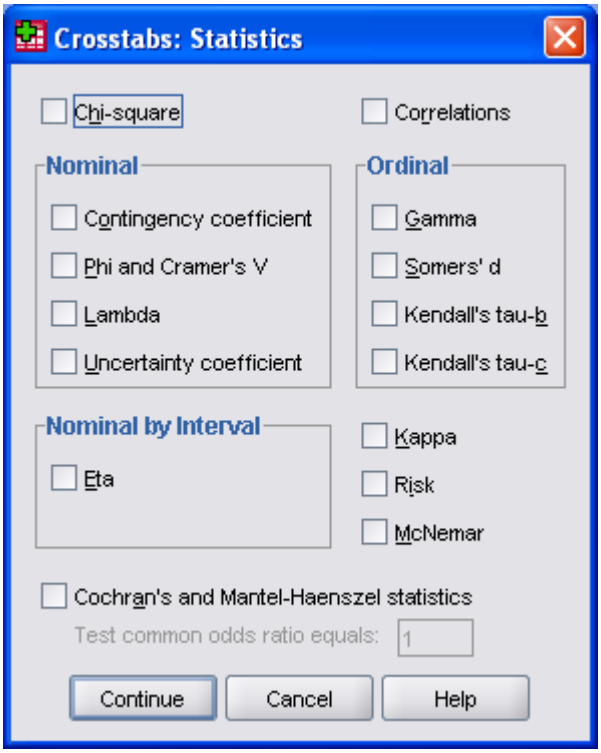


2.2 pav. Dialogo langelis ***Crosstabs***

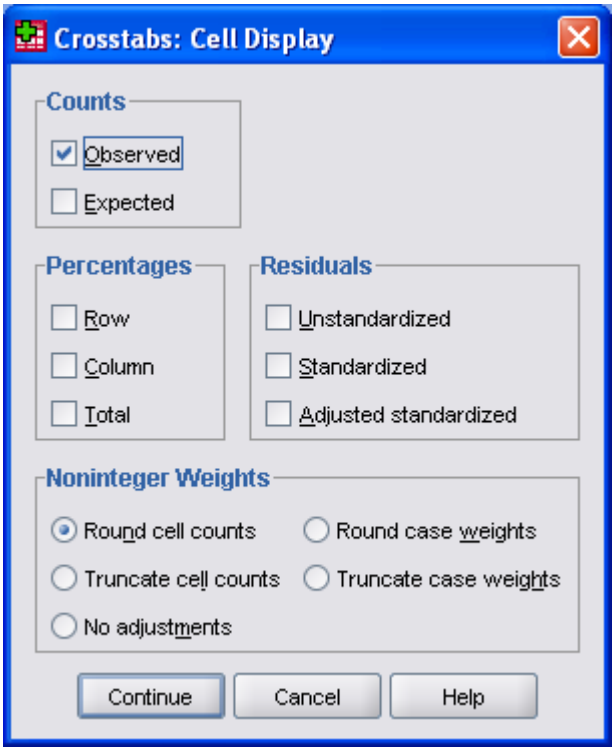
  - ***Total*** (viso): kiekvienos ląstelės reikšmė atžvilgiu bendro stebėjimų skaičiaus.

Įkėlus į laukelį ***Row(s)*** kintamąjį, nurodantį respondentų priklausomybę konkrečiai populiacijai, paprastai užtenka pažymėti ***Column*** laukelį – turėsime kiekvieno atsakymo procentinę dalį atskirai kiekvienai populiacijai.

- Spragtelėkite dialogo langelio ***Crosstabs: Cell Display*** mygtuką ***Continue***, po to – dialogo langelio ***Crosstabs*** mygtuką ***OK***.



2.3 pav. Dialogo langelis ***Crosstabs: Statistics***



2.4 pav. Dialogo langelis ***Crosstabs: Cell Display***

Mūsų atveju gausime 2.5 pav. parodytas ***Crosstabs*** išvesties lenteles. Lentelėje *Case Processing Summary*, kaip paprastai, pateikiama analizuojamų duomenų suvestinė. Dažnių lentelėje *Crosstabulation* sutinkamai su mūsų nustatymu atsakymai į klausimą įrašyti į atskirus stulpelius, o respondentų populiacijos – į atskiras eilutes.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
grupe * Kaip jus vertinate nauja mokesciu istatyma?	70	100,0%	0	,0%	70	100,0%

grupe * Kaip jus vertinate nauja mokesciu istatyma? Crosstabulation

			Kaip jus vertinate nauja mokesciu istatyma?			Total
			pilnai pritariu	2	visiskai nepritariu	
grupe	verslininkai	Count	17	15	8	40
		% within Kaip jus vertinate nauja mokesciu istatyma?	56,7%	65,2%	47,1%	57,1%
	pramonininkai	Count	13	8	9	30
		% within Kaip jus vertinate nauja mokesciu istatyma?	43,3%	34,8%	52,9%	42,9%
Total		Count	30	23	17	70
		% within Kaip jus vertinate nauja mokesciu istatyma?	100,0%	100,0%	100,0%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1,321[a]	2	,517
Likelihood Ratio	1,325	2	,516
Linear-by-Linear Association	,223	1	,637
N of Valid Cases	70		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 7,29.

2.5 pav. Požymių dažnių analizės lentelės

Kiekvienoje lentelės ląstelėje yra įrašytas stebėjimų skaičius (pasirodymo dažnis) – *Count*, o *Total* eilutėje ir stulpelyje – atitinkamai eilučių ir stulpelių reikšmių sumos.

Kadangi pažymėjome dialogo langelio ***Crosstabs: Cell Display*** laukelį ***Column*** kiekvienoje lentelės ląstelėje yra įrašyta taip pat atsakymų skaičiaus procentinė dalis atskirai kiekvienai populiacijai. Pagaliau, lentelėje *Chi-Square Tests* yra pateiktos χ^2 kriterijaus reikšmės (stulpelyje *Value*) ir atitinkamos *p*-reikšmės (stulpelyje *Asymp. Sig*.). Kadangi mūsų atveju Pirsono χ^2 kriterijaus $p > 0{,}05$ $\left(p = 0{,}517\right)$, nulinės hipotezės atmesti negalima – verslininkų ir pramonininkų požiūris į naują mokesčių įstatymą skiriasi statistiškai nereikšmingai.

Pažymėję dialogo langelio ***Crosstabs: Cell Display*** grupės ***Counts*** laukelį ***Expected*** (laukelis ***Observed*** pažymėtas pagal nustatymą) šalia stebimų dažnių gausime ir tikėtinus (prognozuojamus) dažnius, kurie skaičiuojami kaip atitinkamos eilutės ir atitinkamo stulpelio

sumų sandauga, padalinta iš bendros stebėjimų sumos ir kurių galėtume tikėtis jei rezultatai tarp populiacijų nesiskirtų. Lyginant tikėtiną stebėjimų skaičių su eksperimentiniu būdu nustatytu ir atsižvelgiant į nukrypimo pobūdį (daugiau ar mažiau), galima padaryti tam tikras išvadas apie kintamojo pasiskirstymą.

Požymių nepriklausomumo uždavinį turime kai vienoje populiacijoje stebima kintamųjų pora. Pavyzdžiui, reikia nustatyti:

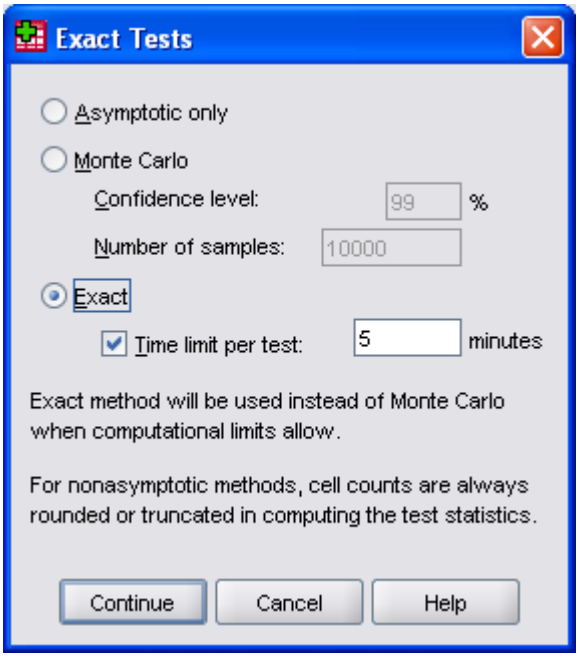
- Ar yra priklausomybė tarp sergamumo tam tikromis ligomis ir įpročių;
- Ar vykdoma ekonominė politika priklauso nuo to, kokia partija yra valdžioje ;

Požymių nepriklausomumo tikrinimo procedūra analogiška aukščiau aprašytajai. Sprendimo priėmimo taisyklė – jeigu χ^2 kriterijaus *p*-reikšmė mažesnė už reikšmingumo lygmenį α, požymiai statistiškai priklausomi, jeigu χ^2 kriterijaus *p*-reikšmė didesnė už reikšmingumo lygmenį α, požymiai statistiškai nepriklausomi.

Požymių dažnių lenteles kintamiesiems, nurodytiems ***Row(s)*** ir ***Column(s)*** laukeliuose galima sudaryti pagal kiekvieną kintamojo, nurodyto ***Crosstabs*** dialogo langelio ***Layer*** (sluoksnis) laukelyje, kategoriją. Norint sudaryti antrą sluoksnį, reikia spragtelėti mygtuką ***Next***. Lentelės sudaromos kiekvienai pirmojo ir antrojo sluoksnio kintamųjų kategorijų kombinacijai.

Norėdami gauti rezultatus pagal tikslųjį Fišerio (*Fisher's*) kriterijų (kai keturlaukėje 2x2 dažnių lentelėje nors vienas tikėtinas stebėjimų skaičius mažiau už 5):

- Spragtelėkite dialogo langelio ***Crosstabs*** mygtuką ***Exact...***
- Dialogo langelyje ***Exact Test*** (2.6 pav.) pasirinkite ***Exact***.

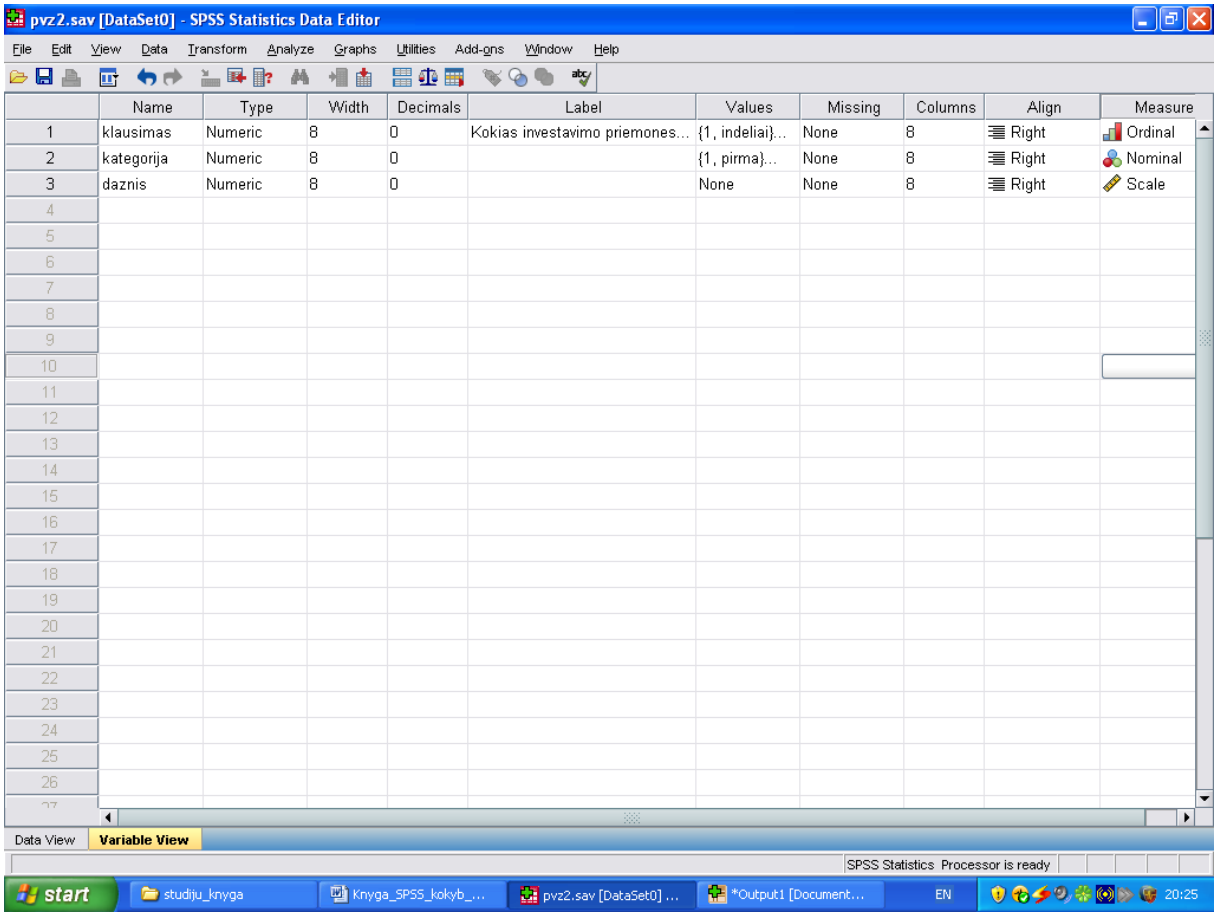


2.6 pav. Dialogo langelis ***Exact Tests***

2.2. KONCENTRUOTŲ DUOMENŲ POŽYMIŲ DAŽNIŲ LENTELĖS

Aukščiau nagrinėtame pavyzdyje respondentai galėjo pasirinkti vieną galimą atsakymo į pateiktą klausimą variantą. Tačiau daugelyje anketinių apklausų pateikiami ir tokie klausimai, atsakydamas į kuriuos respondentas gali pasirinkti kelis (arba visus) galimus

atsakymų variantus. SPSS programų pakete statistinės hipotezės klausimams su daugelio atsakymų galimybe yra tikrinamos t. v. koncentruotų duomenų metodu. Pateiksime hipotetinį pavyzdį. Tiriama, kokios investavimo priemonės yra populiariausios: indėliai, obligacijos, akcijos ar akcijų fondai. Apklausiamos trys respondentų kategorijos, kurias sąlyginai pavadinsime *pirma*, *antra* ir *trečia*. SPSS duomenų rinkmenos langas ***Variable View*** turi atrodyti taip, kaip parodyta 2.7 pav. Atkreipkite dėmesį, kad be klausimo (*klausimas*) ir kategorijos (*kategorija*) kintamųjų atsirado intervalinis kintamasis *daznis*.



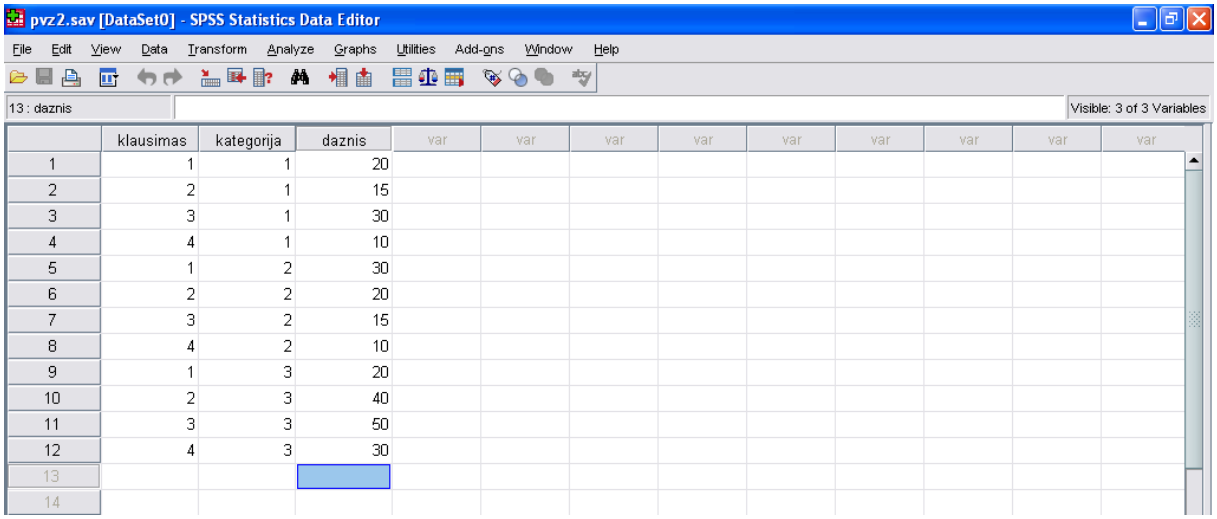
2.7 pav. Kintamųjų langas ***Variable View***

Duomenys lange ***Data View*** surašomi taip, kaip parodyta 2.8 pav., t. y. pateikiama turimų duomenų suvestinė, nurodant kiek kartų atitinkamas atsakymo variantas buvo pasirinktas kiekvienos kategorijos respondentų.

Tokiems koncentruotiems duomenims pirmiausia suteikiamas svorio formatas (*Weight Cases*):

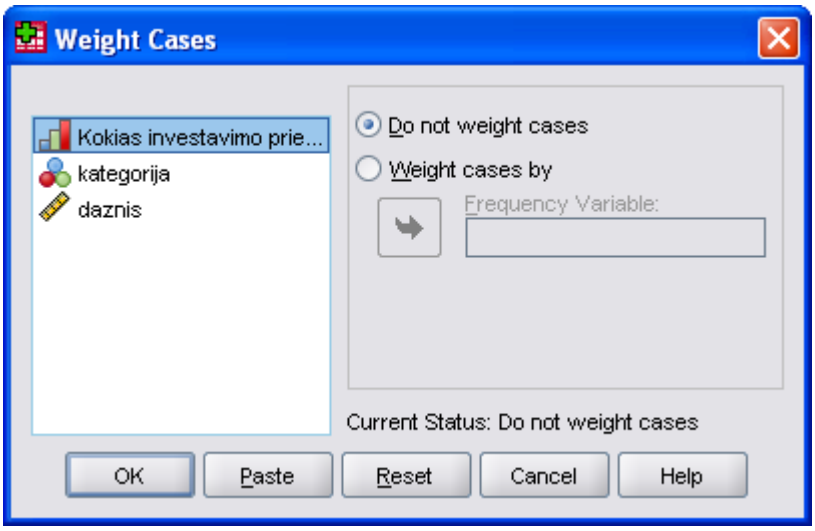
- Nurodome komandas ***Data → Weight Cases...***
- Dialogo langelyje ***Weight Cases*** (2.9 pav.) pažymime variantą ***Weight cases by*** ir perkeliame kintamąjį *daznis* į laukelį ***Frequency Variable*** (dažnių kintamasis).
- Spragtelime mygtuką ***OK***.

Tolimesnė tyrimo, kurio tikslas nustatyti ar vienodai populiarios investavimo priemonės tarp respondentų kategorijų, eiga yra visiškai analogiška išdėstytai 2.1 skyriuje, skirtame statistinių hipotezių klausimams su vieno atsakymo galimybe tikrinimui.



	klausimas	kategorija	daznis	var	var	var	var	var	var	var	var	var
1	1	1	20									
2	2	1	15									
3	3	1	30									
4	4	1	10									
5	1	2	30									
6	2	2	20									
7	3	2	15									
8	4	2	10									
9	1	3	20									
10	2	3	40									
11	3	3	50									
12	4	3	30									
13												
14												

2.8 pav. Duomenų lango ***Data View*** *fragmentas*



2.9 pav. Dialogo langelis ***Weight Cases***

2.3. KATEGORINIŲ DUOMENŲ RYŠIO MATAI

Koreliacijos koeficientas yra tiesinės priklausomybės tarp kintamųjų kiekybinio įvertinimo kriterijus arba ryšio stiprumo matas. Matuojamiems pagal intervalų skalę kintamiesiems yra skaičiuojamas Pirsono (*Pearson*) koreliacijos koeficientas. Kai stebimi kategoriniai kintamieji matuojami pagal rangų arba vardinę skalę naudojami kiti ryšio stiprumo matai (Čekanavičius, Murauskas, 2000), (Yaffee, 2003), (Garson, 2009).

2.3.1. Ranginių kintamųjų ryšio matai

Taikomi matuojamų pagal rangų skalę kintamųjų ryšio stiprumui įvertinti. Paprastai tai neparametriniai ryšio stiprumo matai, nepriklausantys nuo kintamųjų skirstinio pobūdžio. Šie koeficientai matuoja tiesinę kintamųjų priklausomybę: teigiamas koreliacijos koeficientas rodo tiesioginę kintamųjų priklausomybę (didesnes vieno kintamojo reikšmes atitinka

didesnės kito kintamojo reikšmės), neigiamas – atvirkštinę (didesnes vieno kintamojo reikšmes atitinka mažesnės kito kintamojo reikšmės).

Be dažniausiai taikomo Spearman‘o ranginės koreliacijos koeficiento dar naudojami Kendall'o τ ir Gamma ranginės koreliacijos koeficientai. Tačiau Spearman‘o ir Kendall'o τ bei Gamma koeficientai interpretuojami skirtingai – Spearman‘o koeficientas analogiškas Pirsono (*Pearson*), tik skaičiuojamas ranginiams duomenims (o jei duomenys yra intervaliniai – jie paverčiami ranginiais), gi Kendall'o τ ir Gamma koeficientai turi tikimybės dimensiją. Gamma ranginės koreliacijos koeficientas koeficientas apskaičiuojamas pagal formulę

$$\gamma = \frac{P - Q}{P + Q}, \qquad (2.4)$$

čia P – suderintų porų (*concordances*) skaičius, Q – nesuderintų porų (*disconcordances*) skaičius. Dvi duomenų poros (x_i, y_i) ir (x_j, y_j), $(i \neq j)$ yra suderintos, jei $(x_i > x_j \ ir \ y_i > y_j)$ arba $(x_i < x_j \ ir \ y_i < y_j)$. Poros nesuderintos, jei $(x_i > x_j \ ir \ y_i < y_j)$ arba $(x_i < x_j \ ir \ y_i > y_j)$. Iš viso suderintų ir nesuderintų porų yra $n(n-1)/2$, čia n – imties dydis.

Kendall'o τ ranginės koreliacijos koeficientas koeficientas apskaičiuojamas pagal formulę (Garson, 2009)

$$\tau = \frac{P - Q}{\frac{n \cdot (n-1)}{2}}. \qquad (2.5)$$

SPSS yra pateikiami du Kendall'o ranginės koreliacijos koeficiento skaičiavimo variantai, kuriuos Jūs galite pasirinkti dialogo langelyje ***Crosstabs: Statistics***, ***Ordinal*** grupėje – ***Kendall’s tau-b*** ir ***Kendall’s tau-c***.

Kendall'o τ_b koeficientas, kuris dažniausiai naudojamas keturlaukių kontingencijos lentelių atveju, įvertina sutampančias kintamųjų reikšmes (*tied values*), t. y. vienodus rangus ir skaičiuojamas pagal formulę (Garson, 2009)

$$\tau_b = \frac{P - Q}{\sqrt{(P + Q + X_0)(P + Q + Y_0)}}, \qquad (2.6)$$

čia X_0 – nesutampančių porų skaičius atžvilgiu kintamojo *x* ir Y_0 – nesutampančių porų skaičius atžvilgiu kintamojo *y*.

Kendall'o τ_c koeficientas, kuris naudojamas didesnių negu 2x2 dimensijų lentelių atveju, skaičiuojamas pagal formulę (Garson, 2009)

$$\tau_c = \frac{(P - Q) \cdot 2m}{n^2 (m - 1)}, \qquad (2.7)$$

Čia m – kontingencijos lentelės eilučių arba stulpelių skaičius (pasirenkamas mažesnis).

Asymetrinis Somers’o *d* ranginės koreliacijos koeficientas parodo priklausomybės pobūdį tarp pasirinkto nepriklausomo kintamojo ir pasirinkto priklausomo kintamojo. Kai nepriklausomu laikomas *x* kintamasis, Somers’o *d* koeficientas skaičiuojamas pagal formulę (Garson, 2009)

$$d_{yx} = \frac{P - Q}{P + Q + Y_0}, \qquad (2.8)$$

Kai nepriklausomu laikomas *y* kintamasis, Somers'o *d* koeficientas skaičiuojamas pagal formulę (Garson, 2009)

$$d_{xy} = \frac{P - Q}{P + Q + X_0}. \qquad (2.9)$$

Visus šiuos ranginės koreliacijos koeficientus (***Gamma***, ***Sommers'd***, ***Kendall's tau-b***, ***Kendall's tau-c***) Jūs galite pasirinkti dialogo langelyje ***Crosstabs: Statistics***, grupėje ***Ordinal***.

Pavyzdžiui, tiriama koks ryšys tarp atsakymų į aukščiau nagrinėtą klausimą "Kaip jūs vertinate naują mokesčių įstatymą?" ir respondento atstovaujamos įmonės dydžio. Įmonę apibūdinsime kaip "maža", "vidutinė" ir "didelė".

Norėdami apskaičiuoti ranginės koreliacijos koeficientą:

- Nurodykite komandas ***Analyze → Descriptive Statistics →Crosstabs...***
- Atsakymų į klausimą "Kaip jūs vertinate naują mokesčių įstatymą?"duomenis įkelkite į laukelį ***Colum(s)***, o atsakymų į klausimą "Kaip jūs apibūdintumėte savo įmonę" duomenis – į laukelį ***Row(s)***.
- Spragtelėkite dialogo langelio ***Crosstabs*** mygtuką ***Statistics...***
- Atsidariusiame naujame dialogo langelyje ***Crosstabs: Statistics*** (2.3 pav.) pažymėkite laukelį ***Correlations***, o ***Ordinal*** grupėje pasirinkite visus aukščiau išvardintus koreliacijos koeficientus.
- Spragtelėkite dialogo langelio ***Crosstabs: Statistics*** mygtuką ***Continue***, po to – dialogo langelio ***Crosstabs*** mygtuką ***OK***.

Rezultatai bus pateikti išvesties lentelėse *Directional Measures* ir *Symmetric Measures* (2.10 pav.).

Directional Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Somers' d	Symmetric	,128	,107	1,193	,233
		imone Dependent	,126	,105	1,193	,233
		Kaip jus vertinate nauja mokesciu istatyma? Dependent	,131	,109	1,193	,233

a. Not assuming the null hypothesis.

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

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Kendall's tau-b	,128	,107	1,193	,233
	Kendall's tau-c	,122	,103	1,193	,233
	Gamma	,200	,165	1,193	,233
	Spearman Correlation	,142	,119	1,186	,240[c]
Interval by Interval	Pearson's R	,147	,118	1,227	,224[c]
N of Valid Cases		70			

a. Not assuming the null hypothesis.

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

c. Based on normal approximation.

2.10 pav. Ranginės koreliacijos koeficiento skaičiavimo razultatų pavyzdys

Koreliacijos koeficiento reikšmė nurodoma stulpelyje *Value*, o stulpelyje *Approx. Sig*. – kriterijaus *p*-reikšmė, kuria remiantis sprendžiama ar koreliacija statistiškai reikšminga, t. y. ar koreliacijos koeficientas statistiškai reikšmingai skiriasi nuo nulio. Koreliacija statistiškai reikšminga, kai $p < \alpha$, statistiškai nereikšminga, kai $p >= \alpha$, čia α – nustatytas reikšmingumo lygmuo. Šiuo atveju, koreliacijos ryšys yra silpnas – Spearman'o koreliacijos koeficiento (*Spearman Correlation*) reikšmė lygi 0,240, *Kendall's tau-b* koreliacijos koeficiento reikšmė lygi 0,233 ir t. t. – ir statistiškai nereikšmingas. Kadangi kintamieji nepriklauso intervalų skalei, Pirsono koreliacijos koeficiento (*Pearson's R*) reišmės nenagrinėjame.

2.3.2. Vardinių kintamųjų ryšio matai

Koreliacijos koeficiento negalima taikyti tarpusavio ryšiui tarp vardinių kintamųjų, turinčių daugiau dviejų kategorijų, apibūdinti, nes jie yra tarpusavyje nepalyginami, jų negalima surikiuoti. Todėl, priklausomybei tarp tokių kintamųjų nustatyti paprastai naudojamas χ^2 kriterijus. Tačiau SPSS taip pat yra pateikiami kriterijai kiekybiniam ryšio tarp dviejų vardinių kintamųjų įvertinimui. Šie kriterijai parodo dviejų priklausančių vardinei skalei kintamųjų priklausomumo ar nepriklausomumo laipsnį. Kriterijaus reikšmė, lygi nuliui, atitinka visišką kintamųjų nepriklausomumą, o reikšmė, lygi vienetui – didžiausią priklausomumą. Šie tarpusavio ryšio matai negali turėti neigiamų reikšmių, kadangi nesant galimybės išrikiuoti kintamuosius, negalima nustatyti ir priklausomybės krypties.

Trumpai apibūdinsime šiuos kriterijus, kuriuos Jūs galite pasirinkti ***Crosstabs: Statistics*** dialogo langelio kriterijų grupėje ***Nominal***.

Phi – ϕ koeficientas skaičiuojamas χ^2 pagrindu eliminuojant imties dydžio įtaką (Garson, 2009)

$$\phi = \sqrt{\frac{\chi^2}{n}}, \qquad (2.10)$$

Naudojamas tada, kai duomenys aprašomi keturlaukėmis (2x2) kontingencijos lentelėmis, t. y. taikomas binariniams kintamiesiems. Didesnių lentelių atveju didžiausia ϕ reikšmė priklauso nuo lentelės dydžio ir gali viršyti 1.

Contingency Coefficient – kontingencijos koeficientas yra ϕ modifikacija, pritaikyta didesnėms kontingencijos lentelėms. Šis koeficientas skaičiuojamas pagal formulę

$$c = \sqrt{\frac{\chi^2}{\chi^2 + n}}, \qquad (2.11)$$

Kadangi *n* visada daugiau už nulį, koeficiento reikšmė mažesnė už vienetą ir priklauso nuo imties dydžio. Todėl pagal šį koeficientą negalima lyginti kelių atvejų su skirtingais *n*. Kai kurie tyrėjai rekomenduoja šį koeficientą taikyti 5x5 ir didesnėms lentelėms.

Cramer's V – Kramerio *V* koeficientas yra dažniausiai naudojamas vardinių kintamųjų ryšio matas, skaičiuojamas χ^2 pagrindu. Jis nepriklauso nuo lentelės dydžio, kai eilučių skaičius lygus stulpelių skaičiui. Keturlaukėms lentelėms Kramerio *V* koeficientas sutampa su ϕ koeficientu. Skaičiuojamas pagal formulę

$$V = \sqrt{\frac{\chi^2}{n \cdot (m-1)}}, \qquad (2.12)$$

čia *m* – mažiausias iš kontingencijos lentelės eilučių ir stulpelių skaičiaus.

Visi aukščiau išvardinti kriterijai yra skaičiuojami χ^2 kriterijaus pagrindu. Kiti matuojamų pagal vardinę skalę kintamųjų tarpusavio ryšio matai: ***Lambda*** (liamda, λ), ***Goodman and Kruscal's tau*** (liamda modifikacija – Gudmano-Kruskalio tau, τ) ir ***Uncertainly coefficient*** (neapibrėžtumo koeficientas) koeficientai skaičiuojami taip vadinamos proporcingo klaidos mažinimo koncepcijos pagrindu. Jie įvertina priklausomo kintamojo nuspėjamumo santykinį klaidos sumažėjimą, kai žinomas nepriklausomas kintamasis.

Pateiksime pavyzdį. Buvo daroma trijų fakultetų studentų apklausa apie jų domėjimąsi sportu. Anketoje buvo tik vienas klausimas – "Ar Jūs aktyviai domitės sportu?", į kurį reikėjo atsakyti "taip" arba "ne". Taigi, turime du vardinės skalės kintamuosius: *sportas*, kurio reikšmės 1 (*taip*) ir 2 (*ne*) ir *fak_tas*, kurio reikšmėms 1, 2, 3 suteiksime sąlygines žymenas – atitinkamai *A*, *B*, *C*. Dialogo langelyje ***Crosstabs: Statistics*** pasirinksime visus ***Nominal*** variantus. Rezultatai parodyti 2.11 pav.

Directional Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Lambda	Symmetric	,110	,090	1,142	,254
		fak_tas Dependent	,000	,000	.[c]	.[c]
		Ar jus aktyviai domites sportu? Dependent	,235	,182	1,142	,254
	Goodman and Kruskal tau	fak_tas Dependent	,047	,032		,039[d]
		Ar jus aktyviai domites sportu? Dependent	,108	,070		,024[d]
	Uncertainty Coefficient	Symmetric	,064	,043	1,464	,020[e]
		fak_tas Dependent	,052	,036	1,464	,020[e]
		Ar jus aktyviai domites sportu? Dependent	,081	,055	1,464	,020[e]

a. Not assuming the null hypothesis.

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

c. Cannot be computed because the asymptotic standard error equals zero.

d. Based on chi-square approximation

e. Likelihood ratio chi-square probability.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,329	,023
	Cramer's V	,329	,023
	Contingency Coefficient	,312	,023
N of Valid Cases		70	

2.11 pav. Ryšio matų tarp dviejų vardinės skalės kintamųjų skaičiavimo pavyzdys

Lentelėse *Directional Measures* ir *Symmetric Measures* pateiktos aukščiau minėtų koeficientų reikšmės. Pagal kriterijus, skaičiuojamus χ^2 kriterijaus pagrindu, ryšį tarp kintamųjų (t. y. tarp studijų tam tikrame fakultete ir aktyviu domėjimųsi sportu) galime vertinti kaip silpną, bet statistiškai reikšmingą – visų šios kategorijos kriterijų *p*-reikšmė (*Approx. Sig*.) mažesnė už nustatytą reikšmingumo lygmenį (0,05). Pagal kriterijus proporcingo klaidos mažinimo pagrindu ryšys tarp kintamųjų yra gerokai silpnesnis, o pagal liamda kriterijų ir statistiškai nereikšmingas. Tuo atveju, kai neaišku kuris kintamasis gali būti priskirtas priklausomam, skaičiuojami abu variantai bei pateikiamas vidurkis (*Symmetric* – lentelėje *Directional Measures*).

2.3.3. Kiti tarpusavio ryšio matai

Eta (***Nominal by Interval***) koeficientas yra taikomas, kai nepriklausomas kintamasis yra vardinės arba rangininės skalės (kategorinis kintamasis), o priklausomas kintamasis – intervalinis. Kategorinis kintamasis turi turėti skaitmeninį kodavimą (*Numeric* tipo).

Kappa koeficientas taikomas dviejų ekspertų, vertinančių tą patį objektą ar reiškinį, išvadų suderinamumui nustatyti. Koeficiento reikšmė 1 rodo visišką ekspertų vertinimų sutapimą. ***Kappa*** koeficientas taikomas tik tada, kai abudu kintamieji turi tas pačias kategorijų reikšmes ir vienodą kategorijų skaičių.

Risk (rizikos laipsnis) yra skaičiuojamas keturlaukėms 2x2 dažnių lentelėms, sudarytoms laikantis tam tikrų žemiau pateiktų reikalavimų. Skaičiuojant rizikos matą, analizuojamas taip vadinamas rizikos kintamasis, kuris turi dvi kategorijas ir nurodo, įvyko tam tikras įvykis ar ne. Rizikos kintamasis analizuojamas atžvilgiu priežastinio (nepriklausomo) kintamojo, kuris irgi turi būti binarinis (turintis dvi kategorijas). Šį teiginį pailiustruosime paprastu pavyzdžiu. Buvo atlikta 90–ies respondentų apklausa polinkio į depresiją atžvilgiu. Apklausos rezultatai pateikti lentelėje.

Depresija	Taip	Ne
Moterys	a=12	b=44
Vyrai	c=3	d=31

Depresija yra rizikos kintamasis, o *lytis* – nepriklausomas (priežastinis) kintamasis. SPSS rizikos matas yra apskaičiuojamas pagal šias formules:

$$R_0 = \frac{a \cdot d}{b \cdot c}, \text{ galimybių santykis (}\textit{Odds Ratio}\text{)}, \qquad (2.13)$$

$$R_1 = \frac{a \cdot (c + d)}{c \cdot (a + b)}, \text{ santykinės rizikos koeficientas}, \qquad (2.14)$$

$$R_2 = \frac{b \cdot (c + d)}{d \cdot (a + b)}, \text{ santykinės rizikos koeficientas}, \qquad (2.15)$$

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

Kad teisingai apskaičiuoti santykinės rizikos koeficientus SPSS paketu, reikia laikytis šių taisyklių:

- Nepriklausomą (priežastinį) kintamąjį patalpinkite į dialogo langelio ***Crosstabs*** lentelės eilučių (***Row(s)***) sąrašą, o rizikos kintamąjį – į stulpelių (***Column(s)***) sąrašą.
- Pirmoje lentelės eilutėje turi būti didžiausios rizikos grupė, t. y. nepriklausomas kintamasis koduojamas pradedant nuo didžiausios rizikos kategorijos. Mūsų atveju, kintamojo *lytis* reikšmės yra 1 – *moterys*, 2 – *vyrai*.
- Pirmame lentelės stulpelyje turi būti duomenys, jeigu įvykis įvyks, t. y. rizikos kintamasis koduojamas pradedant nuo to, kad įvykis įvyks. Mūsų atveju, kintamojo *depresija* reikšmės yra 1 – *taip*, 2 – *ne*.

Norėdami apskaičiuoti rizikos laipsnį:

- Patalpinkite kintamuosius į dialogo langelio ***Crosstabs*** laukelius ***Row(s)*** ir ***Column(s)*** aukščiau aprašyta tvarka.
- Spragtelėkite dialogo langelio ***Crosstabs*** mygtuką ***Statistics...***

- Atsidariusiame naujame dialogo langelyje ***Crosstabs: Statistics*** pasirinkite variantą ***Risk*** ir pažymėkite laukelį ***Cochran's and Mantel-Haenszel statistics***.
- Laukelyje ***Test common odds ratio equals*** (bendras galimybių santykis) palikite nustatytąją reikšmę 1.
- Spragtelėkite mygtuką ***Continue***, po to – ***OK*** dialogo langelyje ***Crosstabs***.

Pagrindiniai išvesties rezultatai parodyti 2.12 pav.

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for lytis (moterys / vyrai)	2,818	,733	10,828
For cohort depresija = taip	2,429	,738	7,993
For cohort depresija = ne	,862	,725	1,024
N of Valid Cases	90		

Tests of Conditional Independence

	Chi-Squared	df	Asymp. Sig. (2-sided)
Cochran's	2,420	1	,120
Mantel-Haenszel	1,580	1	,209

Under the conditional independence assumption, Cochran's statistic is asymptotically distributed as a 1 df chi-squared distribution, only if the number of strata is fixed, while the Mantel-Haenszel statistic is always asymptotically distributed as a 1 df chi-squared distribution. Note that the continuity correction is removed from the Mantel-Haenszel statistic when the sum of the differences between the observed and the expected is 0.

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			2,818
ln(Estimate)			1,036
Std. Error of ln(Estimate)			,687
Asymp. Sig. (2-sided)			,131
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	,733
		Upper Bound	10,828
	ln(Common Odds Ratio)	Lower Bound	-,310
		Upper Bound	2,382

The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1,000 assumption. So is the natural log of the estimate.

2.12 pav. Rizikos laipsnio įvertinimo pavyzdys

Santykinės rizikos koeficientai apskaičiuoti tiek galimybei susirgti – *For cohort depresija=taip*, tiek galimybei nesusirgti – *For cohort depresija=ne*. Taip, rizika susirgti depresija moterims yra 2,429 karto didesnė negu vyrams, o galimybė nesusirgti depresija moterims sudaro 0,862 galimybės nesusirgti vyrams. Galimybių santykis (*Odds Ratio*) įvertina santykinę riziką pagal (2.13) formulę. Nulinė hipotezė, teigianti, kad galimybių santykis lygus 1, t. y. kintamieji yra nepriklausomi, yra tikrinama Mantelio-Haenzelio

(*Mantel-Haenszel*) kriterijumi. Lentelėje *Mantel-Haenszel Common Odds Ratio Estimate* (Mantelio-Haenzelio bendro galimybių santykio įvertis) pateikta kriterijaus *p*-reikšmė (*Asymp. Sig. (2-sided)*) rodo, kad nulinės hipotezės atmesti negalima. Analogiškas rezultatas pateikiamas lentelėje *Test of Conditional Independence.* Norėdami patikrinti galimybių santykių lygybę pagal kiekvieną kintamojo, nurodyto dialogo langelio ***Crosstabs*** laukelyje ***Layer*** (sluoksnis), kategoriją:

- Patalpinkite kintamuosius į dialogo langelio ***Crosstabs*** laukelius ***Row(s)*** ir ***Column(s)*** aukščiau aprašyta tvarka.
- Patalpinkite kintamąjį, pagal kurio kategorijas bus skaičiuojamas galimybių santykis, į dialogo langelio ***Crosstabs*** laukelį ***Layer***.
- Spragtelėkite dialogo langelio ***Crosstabs*** mygtuką ***Statistics...***
- Atsidariusiame naujame dialogo langelyje ***Crosstabs: Statistics*** pasirinkite ***Cochran's and Mantel-Haenszel statistics***.
- Laukelyje ***Test common odds ratio equals*** (bendras galimybių santykis) palikite nustatytąją reikšmę 1.
- Spragtelėkite mygtuką ***Continue***, po to – ***OK*** dialogo langelyje ***Crosstabs***.

Pavyzdžiui, sugrupavę aukščiau pateikto pavyzdžio apie polinkį į depresiją duomenis pagal dvi amžiaus kategorijas (iki 20 m. ir daugiau 20 m.) ir įkėlę kintamąjį *amzius* į dialogo langelio ***Crosstabs*** laukelį ***Layer*** gautume 2.13 pav. parodytą papildomą lentelę galimybių santykio homogeniškumui pagal į laukelį ***Layer*** įkelto kintamojo kategorijas įvertinti.

Tests of Homogeneity of the Odds Ratio

	Chi-Squared	df	Asymp. Sig. (2-sided)
Breslow-Day	,154	1	,695
Tarone's	,153	1	,696

2.13 pav. Galimybių santykio homogeniškumo pagal atskiras kategorijas skaičiavimo pavyzdys

Nulinė hipotezė, teigianti, kad galimybių santykis yra homogeniškas pagal atskiras sluoksnio kintamojo kategorijas, yra tikrinama pagal *Breslow-Day* ir *Tarone's* kriterijus. Mūsų atveju, šių kriterijų *p*-reikšmė (*Asymp. Sig. (2-sided)*) rodo, kad nulinė hipotezė neatmestina, t. y. galimybių santykis yra homogeniškas pagal amžiaus kategorijas.

McNemar'o kriterijus yra taikomas tada, kai tiriamas dvireikšmis (binarinis) kintamasis (nuostata, gebėjimai, sveikata ir t. t.) matuojamas du kartus – iki poveikio ir po jo.

3. KLAUSIMYNŲ PATIKIMUMO VERTINIMAS

Klausimyno patikimumas (angl. *reliability*) yra suprantamas kaip koreliacija tarp gautų testo rezultatų ir hipotetinių "tikrų" rezultatų (Norušis, 2005). Kadangi nėra galimybės gauti šiuos "tikrus" rezultatus, klausimyno patikimumui įvertinti yra naudojamos šios pagrindinės charakteristikos (Garson, 2009), (Norušis, 2005):

- klausimyno skalės vidinis nuoseklumas (angl. *scale internal consistency*), kuris remiasi atskirų klausimų, sudarančių klausimyną, koreliacija (tiksliau, atsakymų į atskirus klausimyną sudarančius klausimus, koreliacija);
- klausimyno patikimumas pakartotinių tyrimų atžvilgiu (angl. *test-retest reliability)*, kuris remiasi dviejų (ar daugiau) bandymų koreliacija;
- vertinimo patikimumas (angl. *inter-rater reliability*), kuris remiasi koreliacija tarp dviejų (ar daugiau) ekspertų vertinimų.

3.1. KLAUSIMYNO SKALĖS VIDINIS NUOSEKLUMAS

Klausimyno skalės vidiniam nuoseklumui (angl. *scale internal consistency*) įvertinti dažniausiai yra naudojamas Cronbacho alfa (***Cronbach's alpha***) koeficientas, kuris, kaip minėta aukščiau, remiasi atskirų klausimų, sudarančių klausimyną, koreliacija ir įvertina, ar visi skalės klausimai pakankamai atspindi tiriamąjį dydį bei įgalina patikslinti reikiamų klausimų skaičių skalėje. Jeigu atskirų klausimų dispersijų suma yra artima visos skalės dispersijai, reiškia atskiri klausimai tarpusavyje nekoreliuoja, t. y. jie neatspindi to paties dalyko. Šiuo atveju, klausimyno skalė yra sudaryta iš atsitiktinių klausimų ir Cronbacho alfa koeficientas yra artimas 0. Jeigu visos skalės dispersija yra ženkliai didesnė už atskirų klausimų dispersijų sumą, reiškia atskiri klausimai tarpusavyje koreliuoja, t. y. jie atspindi tą patį dalyką (Norušis, 2005). Šiuo atveju, Cronbacho alfa koeficientas yra artimas 1. Cronbacho alfa koeficientas didėja didinant klausimų, sudarančių klausimyną, skaičių. Tai reiškia, kad klausimynai su didesniu klausimų skaičiumi yra patikimesni. Taip pat, kad negalima lyginti pagal Cronbacho alfa koeficientą dviejų skalių su skirtingu elementų skaičiumi. Cronbacho alfa koeficientas apskaičiuojamas pagal formulę (Yaffee, 2003)

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{i=1}^{k} S_i^2}{S_p^2}\right), \qquad (3.1)$$

čia k – skalės elementų skaičius, S_i^2 yra i-tojo skalės elemento dispersija, S_p^2 – bendra skalės dispersija. Skalės dispersija apskaičiuojama pagal formulę

$$S_p^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n} P_j^2 - n\left(\sum_{i=1}^{k} \overline{T}_i\right)^2\right], \qquad (3.2)$$

čia:

$P_j = \sum_{i=1}^{k} X_{ji}$ – j-tojo respondento atsakymų į visus k klausimus įverčių suma;

$T_i = \sum_{j=1}^{n} X_{ji}$ – *i*-tojo skalės elemento (atsakymų į *i*-tąjį klausimą) įverčių suma per visus respondentus;

$\overline{T}_i$ – *i*-tojo skalės elemento įverčių vidurkis, , *n* – imties dydis;

$G = \sum_{i=1}^{k} \sum_{j=1}^{n} X_{ji}$ – bendra visų *n* respondentų atsakymų į visus *k* klausimus įverčių suma;

X_{ji} – *j*-tojo respondento atsakymo į *i*-tąjį klausimą įvertis;

$$\begin{array}{cccccccc} & 1 & 2 & \cdots & i & \cdots & k & \\ 1 & X_{11} & X_{12} & & & & X_{1k} & P_1 \\ \vdots & & & & & & & \vdots \\ j & & & & X_{ji} & & & P_j \\ \vdots & & & & & & & \vdots \\ n & X_{n1} & X_{n2} & & & & X_{nk} & P_n \\ & T_1 & T_2 & \cdots & T_i & \cdots & T_k & G \end{array} \qquad (3.3)$$

Cronbacho alfa koeficientas gali būti interpretuojamas dvejopai (Garson, 2009):

- Tai dispersijos dalis, kurią gali paaiškinti duotoji skalė, palyginus su hipotetine tikrąja skale, sudarytą iš visų įmanomų klausimų, skirtų mus dominančios charakteristikos nustatymui.
- Tai koreliacija tarp duotosios skalės ir visų kitų įmanomų skalių, skirtų mus dominančios charakteristikos nustatymui ir sudarytų iš to paties klausimų skaičiaus.

Standartizuotų duomenų alfa koeficientas skirtas įvertinti atsakymų į atskirus klausimus dispersijų skirtingumą. Jis dar vadinamas Spearman-Brown‘o padidinto patikimumo koeficientu (*Spearman-Brown stepped-up reliability coefficient*) ir skaičiuojamas pagal formulę (Garson, 2009)

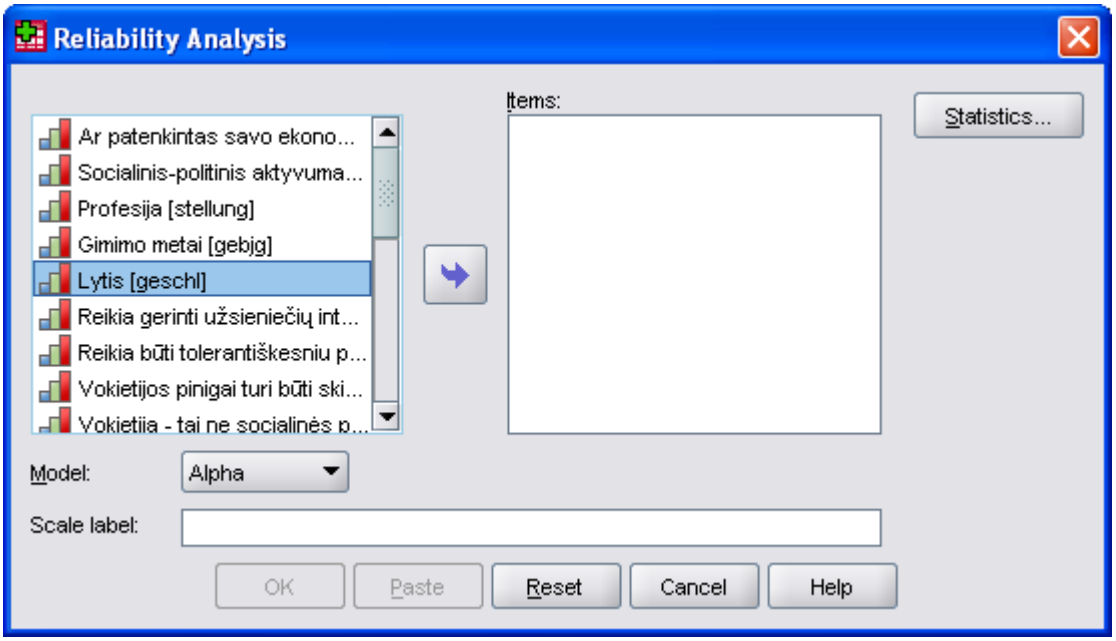
$$\alpha = \frac{k \cdot \bar{r}}{1 + (k-1)\bar{r}}, \qquad (3.4)$$

čia $\bar{r}$ – koreliacijos koeficiento tarp visų įmanomų atsakymų į klausimus porų vidurkis (*the average of inter-item correlations*).

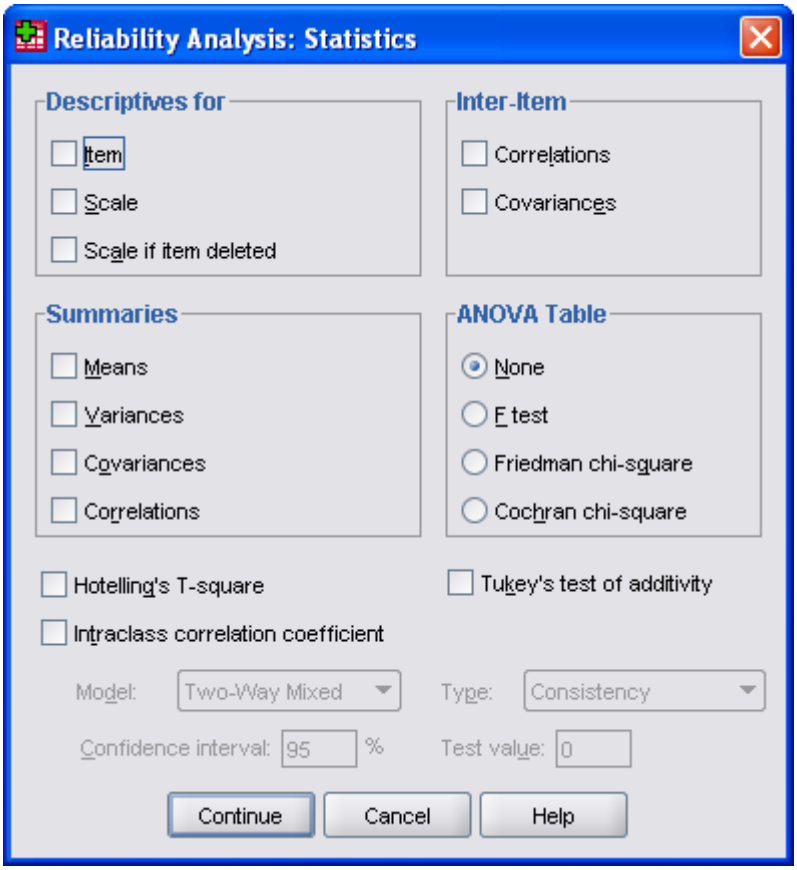
Norėdami apskaičiuoti Cronbacho alfa koeficientą SPSS paketu:

- Atidarykite bylą su analizuojamais duomenimis.
- Nurodykite komandas ***Analyze → Scale →Reliability Analysis...*** Atsidarys dialogo langelis ***Reliability Analysis*** (3.1 pav.).
- Įkelkite kintamuosius iš sąrašo į laukelį ***Items***; laukelyje ***Model*** palikite nustatytą ***Alpha***.
- Spragtelėkite dialogo langelio ***Reliability Analysis*** mygtuką ***Statistics...***
- Atsidariusiame naujame dialogo langelyje ***Reliability Analysis: Statistics*** (3.2 pav.) pažymėkite laukelį ***Scale if item deleted***.
- Pažymėkite ***Summaries*** laukelį ***Correlations*** jeigu norite gauti Spearman-Brown‘o padidinto patikimumo koeficiento reikšmę.
- Pažymėkite ***Inter-Item*** laukelį ***Correlations*** jeigu norite gauti koreliacijos koeficientų reikšmes tarp atsakymų į visus klausimyną sudarančius klausimus.

- Spragtelėkite dialogo langelio ***Reliability Analysis: Statistics*** mygtuką ***Continue***, po to – dialogo langelio ***Reliability Analysis*** mygtuką ***OK***.



3.1 pav. Dialogo langelis ***Reliability Analysis*** *Cronbacho alfa koeficientui apskaičiuoti*



3.2 pav. Dialogo langelis ***Reliability Analysis: Statistics***

Išvesties lentelėje *Reliability Statistics* (3.3 pav.) pateikta Cronbacho alfa koeficiento reikšmė, kuri gerai sudarytam klausimynui turėtų būti didesnė už 0,7 (kai kurių autorių teigimu – už 0,6). Spearman-Brown'o padidinto patikimumo koeficientas yra įvardintas kaip *Cronbach's Alpha Based on Standardized Items*. Jo reikšmė artima Cronbacho alfa koeficiento reikšmei, kas reiškia, kad atsakymų į atskirus klausimus dispersijos yra panašios. Lentelėje *Summary Item Statistics* pateikiamas koreliacijos koeficiento tarp visų atsakymų į klausimus porų vidurkis, mažiausia reikšmė, didžiausia reikšmė ir kt.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,609	,636	11

Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Inter-Item Correlations	,137	-,193	,540	,733	-2,802	,019	11

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Siulomos aukstos kokybes sporto paslaugos	20,28	25,611	,326	,577
Siulomos sporto paslaugos turi specifini atlikimo stiliu	19,94	25,231	,329	,575
Siulomas gausus sporto paslaugu pasirinkimas	20,35	24,408	,412	,558
Apie paslaugas teikiama profesionali informacija	20,47	25,416	,367	,570
Suteikiama operatyvi informacija apie produktu ir paslaugu kainas, ju pokycius ir nuolaidas	20,29	25,170	,360	,570
Imone isikurusi patogioje vietoje	20,56	26,750	,175	,604
Patogus darbo laikas (lankymas laisvu grafiku)	20,75	27,393	,152	,607
Isores reklama gerai matosi	20,17	23,496	,341	,569
Paslaugos reklamuojamos pasitelkiant lauko reklama	20,29	24,506	,333	,572
Paslaugos teikiamos moderniuose patalpose	19,38	25,005	,175	,615
Prie imones patogi automobiliu stovejimo aikstele	19,70	25,289	,142	,626

3.3 pav. Klausimyno skalės vidinio nuoseklumo skaičiavimo pavyzdys

Lentelės *Item-Total Statistics* stulpelyje *Corrected Item-Total Correlation* pateiktos koreliacijos koeficiento tarp atsakymų į konkretų klausimą ir suminės klausimyno skalės reikšmės. Klausimai, turintys silpną koreliacinį ryšį su klausimyno skale (kurių koreliacijos su klausimyno skale koeficientas mažesnis už 0,1 – 0,2), gali būti iš klausimyno pašalinti, jeigu tai sąlygoja Cronbacho alfa koeficiento pastebimą padidėjimą ir jeigu jie teoriškai nėra būtini klausimyne. Kaip pasikeis Cronbacho alfa koeficiento reikšmė pašalinus klausimą iš klausimyno nurodoma lentelės *Item-Total Statistics* stulpelyje *Cronbach's Alpha if Item deleted.* Neigiamos koreliacijos koeficiento reikšmės (lentelės *Item-Total Statistics* stulpelyje *Corrected Item-Total Correlation*) reiškia, kad atsakymus į atitinkamą klausimą reikia perkoduoti atvirkštine tvarka. Pašalinus kurio nors klausimo atsakymus arba juos perkodavus, analizę reikia atlikti iš naujo.

Iš kitų, mažiau populiarių už Cronbacho alfa, klausimyno patikimumo vertinimo matų galima paminėti Gutmano koeficientus. Pasirinkus dialogo langelio ***Reliability Analysis*** (3.1 pav.) laukelyje ***Model*** metodą ***Guttman***, išvestyje bus pateiktos šešios Gutmano koeficientų reikšmės. Iš jų, kai kurių tyrėjų yra naudojamas *Lambda 2*, *Lambda 3* lygus Cronbacho alfa. *Lambda 5* yra rekomenduojamas, kai vieno klausimo atsakymai stipriai koreliuoja su atsakymais į kitus klausimus, kurie, savo ruožtu, tarpusavyje neturi stipraus koreliacinio ryšio. Tokia situacija galima, pavyzdžiui, kada kiekvienas klausimas turi ryšį su atskira žinių sritimi, o į vieną klausimą galima atsakyti turint žinių iš bet kurios šių sričių. *Lambda 6* yra rekomenduojamas, kai koreliacijos ryšys tarp atsakymų į klausimus yra silpnas, palyginus su koeficientu R^2, nusakančiu regresinę atsakymų į vieną klausimą priklausomybę nuo atsakymų į visus kitus klausimus (Garson, 2009), (Norušis, 2005).

3.2. KLAUSIMYNO PATIKIMUMAS PAKARTOTINIŲ TYRIMŲ ATŽVILGIU

Klausimyno stabilumas pakartotiniams tyrimams (angl. *Test-retest reliability*) vertinamas atliekant pakartotinę tų pačių respondentų apklausą po tam tikro laiko. Laiko intervalas yra nustatomas atsižvelgiant į atliekamą tyrimą, bet paprastai tai būna kelios savaitės. Klausimyno patikimumas pakartotinių vykdymų atžvilgiu matuojamas ***Spearman-Brown*** koeficientu, kuris apskaičiuojamas pagal koreliacijos tarp pirmo ir antro bandymų duomenų koeficientą (Yaffee, 2003), (Garson, 2009). Nors kai kurie tyrinėtojai skeptiškai vertina šios charakteristikos patikimumą, klausimyno stabilumas pakartotiniams vykdymams vis dar yra gana populiarus. Norėdami įvertinti klausimyno patikimumą pakartotinių tyrimų atžvilgiu:

- Atidarykite bylą su analizuojamais duomenimis. Joje turi būti abiejų bandymų duomenys.
- Nurodykite komandas ***Analyze → Scale →Reliability Analysis...*** Atsidarys dialogo langelis ***Reliability Analysis*** (3.1 pav.).
- Įkelkite iš sąrašo į laukelį ***Items*** kintamuosius: pradžioje pirmo bandymo kintamuosius, po jų – antro bandymo kintamuosius tuo pačiu eiliškumu. Programa automatiškai padalina visus kintamuosius į dvi lygias grupes. Jeigu kintamųjų skaičius pasirodytų nelyginis, pirmai kintamųjų grupei bus priskirta vienu kintamuoju daugiau.
- Laukelyje ***Model*** nustatykite ***Split-half***.
- Spragtelėkite dialogo langelio ***Reliability Analysis*** mygtuką ***OK***.

Lentelėje *Reliability Statistics* (3.4 pav.) pateikta Cronbacho alfa koeficiento reikšmė kiekvienam bandymui, o taip pat Spearman-Brown‘o koeficiento, kuris yra patikimumo pakartotinių tyrimų atžvilgiu matas, reikšmė.

Reliability Statistics

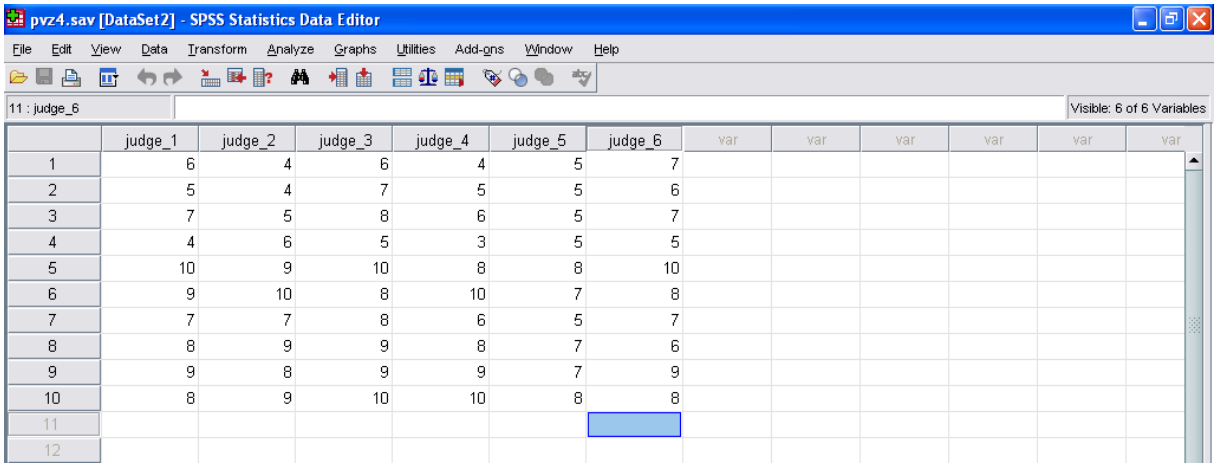
Cronbach's Alpha	Part 1	Value	,609
		N of Items	11[a]
	Part 2	Value	,607
		N of Items	11[b]
		Total N of Items	22
		Correlation Between Forms	,995
Spearman-Brown Coefficient		Equal Length	,998
		Unequal Length	,998
		Guttman Split-Half Coefficient	,998

3.4 pav. Klausimyno stabilumo pakartotiniams vykdymams vertinimo pavyzdys (išvesties lentelės fragmentas)

3.3. VERTINIMO PATIKIMUMAS

Vertinimo patikimumas (*angl. Inter-rater reliability*) nustato vertinimo homogeniškumą ir taikomas nustatyti vertintojų konsenso laipsnį, kai du ar daugiau vertintojų taiko tuos pačius vertinimo kriterijus tiems patiems žmonėms ar reiškiniams vertinti. Kategoriniams duomenims, konsensas yra apskaičiuojamas sutarimų skaičių padalinant iš bendro stebėjimų skaičiaus. Intervaliniams duomenims konsensas yra matuojamas intraklasiniu koreliacijos koeficientu (angl. *Intraclass Correlation Coefficient – ICC*), kuris

taip pat naudojamas ir pakartotinių tyrimų stabilumui nustatyti. Intraklasinis koreliacijos koeficientas savo esme yra tarpgrupinės dispersijos santykis su bendra dispersija. Jis įvertina ne tik priklausomybės tarp dviejų kintamųjų laipsnį, bet ir šių kintamųjų suderinamumą jų vidurkių atžvilgiu (dviejų ekspertų vertinimai gali koreliuoti tarpusavyje, bet labai skirtis savo dydžiais). SPSS duomenų rinkmenoje atskiri kintamieji reiškia atskirų ekspertų vertinimus intervalinėje arba ją atitinkančioje (pvz. Likerto) skalėje (3.5 pav.) Intraklasinio koreliacijos koeficiento reikšmė bus tuo arčiau 1, kuo labiau sutaps ekspertų vertinimai, vertinant kiekvieną atvejį. Intraklasinio koreliacijos koeficiento skaičiavimo modeliai priklauso nuo vertintojų (ekspertų) statuso – pasirinkti norimi ar sudaro atsitiktinę imtį, nuo tiriamųjų vertinimo – vertinami visi duotos kategorijos tiriamieji ar tik sudarantys atsitiktinę imtį bei nuo to, ar patikimumas paremtas atskirų ekspertų vertinimu ar vertinimų vidurkiu (Yaffee, 2003).



	judge_1	judge_2	judge_3	judge_4	judge_5	judge_6	var	var	var	var	var	var
1	6	4	6	4	5	7						
2	5	4	7	5	5	6						
3	7	5	8	6	5	7						
4	4	6	5	3	5	5						
5	10	9	10	8	8	10						
6	9	10	8	10	7	8						
7	7	7	8	6	5	7						
8	8	9	9	8	7	6						
9	9	8	9	9	7	9						
10	8	9	10	10	8	8						
11												
12												

3.5 pav. Duomenų rinkmena intraklasiniam koreliacijos koeficientui nustatyti

SPSS yra numatyti šie intraklasinio koreliacijos koeficiento skaičiavimo modeliai ir tipai:

- Vieno faktoriaus atsitiktinių veiksnių modelis (angl. *One-way random effects model*), kada kiekvieną subjektą vertina skirtinga atsitiktinai parinktų ekspertų grupė. Tai retai pasitaikantis atvejis.
- Dviejų faktorių atsitiktinių veiksnių modelis (angl. *Two-way random effects model*), kada kiekvienas ekspertas iš atsitiktinai parinktų ekspertų grupės vertina kiekvieną atsitiktinai parinktą subjektą.
- Dviejų faktorių mišrus modelis (angl. *Two-way mixed model*), kada kiekvienas parinktas ekspertas vertina kiekvieną atsitiktinai parinktą subjektą. Modelis suprantamas kaip mišrus, nes ekspertai yra pastovus faktorius, o ne atsitiktinė imtis.
- Visiško sutarimo (angl. *Absolute agreement*) tipas. Pasirenkamas tada, kai ekspertų vertinimai yra identiški savo absoliutine išraiška ir kai svarbu įvertinti sisteminę ekspertų paklaidą.
- Nuoseklus (angl. *Consistency*) tipas. Taikomas tada, kai ekspertų vertinimai yra stipriai koreliuoti arba nėra identiški savo absoliutine išraiška.

Kiekvienas intraklasinio koreliacijos koeficiento modelis turi dvi versijas:

- Pavienių vertinimų patikimumas – analizuojami atskirų ekspertų vertinimo duomenys.
- Vertinimų vidurkio patikimumas – analizuojami ekspertų vertinimo vidurkiai.

Norėdami apskaičiuoti intraklasinio koreliacijos koeficiento reikšmę:

- Atidarykite bylą su analizuojamais duomenimis.

- Nurodykite komandas ***Analyze → Scale →Reliability Analysis...*** Atsidarys dialogo langelis ***Reliability Analysis*** (2.1 pav.).
- Įkelkite kintamuosius iš sąrašo į laukelį ***Items***; laukelyje ***Model*** palikite nustatytą ***Alpha***.
- Spragtelėkite dialogo langelio ***Reliability Analysis*** mygtuką ***Statistics...***
- Atsidariusiame naujame dialogo langelyje ***Reliability Analysis: Statistics*** (2.2 pav.) pažymėkite laukelį ***Item*** bei laukelį ***Intraclass Correlation Coefficient***.
- Pasirinkite reikiamą ICC skaičiavimo modelį bei tipą.
- Spragtelėkite dialogo langelio ***Reliability Analysis: Statistics*** mygtuką ***Continue***, po to – dialogo langelio ***Reliability Analysis*** mygtuką ***OK***.

Lentelėje *Intraclass Correlation Coefficient* (3.6 pav.) pateikiamos dvi intraklasinio koreliacijos koeficiento reikšmės: pavienio vertinimo *Single measures* ir vertinimų vidurkio *Average measures*. Pavienio vertinimo atvejis taikytinas tada, kai tyrėjas ketina remtis vieno eksperto vertinimais. Vertinimų vidurkio atvejis taikytinas tada, kai tyrimas remiasi kiekvieno tiriamojo suvidurkintu kelių ekspertų vertinimu. Intraklasinis koreliacijos koeficientas statistiškai reikšmingai skiriasi nuo 0, jeigu F kriterijaus *p*-reikšmė (*Sig*) mažesnė už reikšmingumo lygmenį (0,05).

Reikia pažymėti, kad pasirinkus ***Two-way random effects model*** ar ***Two-way mixed model*** vidurkio intraklasinis koreliacijos koeficientas yra lygus Cronbacho alfa koeficientui.

Intraclass Correlation Coefficient

	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	,745[b]	,527	,914	18,557	9	45	,000
Average Measures	,946[c]	,870	,985	18,557	9	45	,000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

b. The estimator is the same, whether the interaction effect is present or not.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

3.6 pav. Intraklasinio koreliacijos koeficiento skaičiavimo išvesties pavyzdys

Ranginiams duomenims yra skaičiuojamas Friedman‘o χ^2 kriterijus bei Kendall‘o konkordacijos koeficientas, o dvireikšmiams (binariniams) duomenims – Cochran‘o χ^2 kriterijus. Friedmano χ^2 kriterijauss statistika nusakoma formule

$$\chi^2 = \frac{\left(12/nk(k+1)\right)\sum_j^k C_j^2 - 3n(k+1)}{1-\sum T/nk\left(k^2-1\right)}, \qquad (3.5)$$

čia $\sum T = \sum_{i=1}^{n}\sum_{j=1}^{k}\left(t^3-t\right)$, C_j – j kintamojo ($j=1,2,\cdots,k$) rangų suma, k – kintamųjų skaičius, n – imties dydis, t – kintamųjų su sutampančiais rangais per visus tiriamuosius skaičius.

Kendall‘o konkordacijos koeficientas yra apskaičiuojamas pagal formulę

$$W = \left(\frac{F}{n(k-1)}\right)\left(\frac{n^2k(k^2-1)/12}{n^2k(k^2-1)/12 - n\sum T/12}\right), \qquad (3.6)$$

čia F – Friedman‘o χ^2 statistika.

Norėdami gauti vertinimo patikimumo rodiklius ranginiams duomenims:

- Atidarykite bylą su analizuojamais duomenimis.

Pastaba. Norėdami gauti Kendall‘o konkordacijos koeficiento reikšmę duomenis SPSS duomenų rinkmenoje suveskite taip, kad eilutės reikštų ekspertus, o stulpeliai – tiriamuosius! Tada Friedman‘o kriterijaus *p*-reikšmė rodys skirtumo tarp tiriamųjų reikšmingumą, o Kendall‘o konkordacijos koeficientas – ekspertų sutarimo laipsnį.

- Nurodykite komandas ***Analyze*** ➞ ***Scale*** ➞***Reliability Analysis...*** Atsidarys dialogo langelis ***Reliability Analysis*** (2.1 pav.).
- Įkelkite kintamuosius iš sąrašo į laukelį ***Items***.
- Spragtelėkite dialogo langelio ***Reliability Analysis*** mygtuką ***Statistics...***
- Atsidariusiame naujame dialogo langelyje ***Reliability Analysis: Statistics*** (2.2 pav.) ***ANOVA Table*** grupėje pažymėkite laukelį ***Friedman chi–square*** (bus pateiktas ir Kendall‘o konkordacijos koeficientas) arba laukelį ***Cochran chi–square***, jeigu duomenys binariniai.
- Spragtelėkite dialogo langelio ***Reliability Analysis: Statistics*** mygtuką ***Continue***, po to – dialogo langelio ***Reliability Analysis*** mygtuką ***OK***.

Lentelėje *ANOVA with Friedman‘s Test* (3.7 pav.) Friedman‘o kriterijaus *p*-reikšmė (*Sig*) rodo, kad tiriamųjų rangai skiriasi statistiškai reikšmingai. Tuo tarpu Kendall‘o konkordacijos koeficiento reikšmė rodo, kad ekspertų nuomonės didžia dalimi sutampa (koreliuoja).

ANOVA with Friedman's Test

		Sum of Squares	df	Mean Square	Friedman's Chi-Square	Sig
Between People		16,133	5	3,227		
Within People	Between Items	155,733[a]	9	17,304	42,994	,000
	Residual	39,867	45	,886		
	Total	195,600	54	3,622		
Total		211,733	59	3,589		

Grand Mean = 7,07

a. Kendall's coefficient of concordance W = ,736.

3.7 pav. Friedman‘o χ^2 kriterijaus bei Kendall‘o konkordacijos koeficiento skaičiavimo išvesties pavyzdys

4. FAKTORINĖ ANALIZĖ

Faktorinės analizės (Čekanavičius, Murauskas, 2002), (Garson, 2009) užduotis – atsižvelgiant į tarpusavio koreliaciją, suskirstyti stebimus kintamuosius į grupes, kurias vienija koks nors tiesiogiai nestebimas faktorius. Pereidami nuo didelio skaičiaus kintamųjų prie faktorių mes koncentruojame informaciją, padarome ją labiau aprėpiamą. Patys faktoriai dažnai neturi kiekybinio mato, pvz. kūrybiškumas, agresija, altruizmas negali būti išmatuoti betarpiškai, bet šias sąvokas galime įsivaizduoti kaip atitinkamas požymių grupes vienijančias kategorijas. Šiame skyriuje nagrinėsime tiriančiosios faktorinės analizės metodus, kai faktorių skaičius bei juos sudarantys kintamieji iš anksto nėra žinomi. Faktorinės analizės idėją pailiustruosime šiuo paprastu pavyzdžiu, pateiktu elektroniniame ištekliuje. Tarkime, tiriama, kodėl dalis studentų neigiamai žiūri į kompiuterinės statistikos dalyką, t. y. kokie faktoriai sąlygoja tą nenorą. Respondentams pateikiami klausimai apima įvairius to nenoro aspektus. Atsakymai vertinami penkių balų sistema nuo "pilnai sutinku" iki "griežtai nesutinku". Faktorinėje analizėje pagal respondentų vertinimų koreliacijas studentai yra suskirstomi į kelias grupes. Tada sprendžiama, koks faktorius galėtų vienyti konkrečios grupės studentus. Pavadinimą faktoriui suteikia pats tyrėjas, išanalizavęs grupės sudėtį. Šiuo atveju, tai gali būti nepasitikėjimas savo darbo kompiuteriu įgūdžiais, silpnos matematikos žinios ir t. t. Detali pavyzdžio analizė pateikta žemiau.

Faktorine analize siekiama:

- Sumažinti didelį kintamųjų skaičių pereinant prie mažesnio bendrųjų faktorių skaičiaus. Tai gali būti savarankiškas tikslas arba latentinių faktorių reikšmių įverčiai gali būti naudojami kaip pradinių duomenų pakaitalas klasterinėje, regresinėje ar kt. analizėje;
- Patvirtinti naudojamą skalę, parodant, kad skalės sudedamosios dalys patenka į tą patį faktorių bei tuo pačiu pašalinti tas sudedamąsias dalis, kurios patenka į kelis faktorius;
- Sudaryti ortogonalius (tarpusavyje nekoreliuotus) faktorius, kuriuos galima naudoti regresinėje analizėje, išvengiant kintamųjų multikolinearumo problemos;

Faktorinė analizė priklauso bendrojo tiesinio modelio (G*eneral Linear Model – GLM*) kategorijai ir remiasi daugialypei tiesinei regresijai analogiškomis prielaidomis (Garson, 2009), pagrindinės kurių – tiesinė kintamųjų priklausomybė, intervaliniai arba jiems artimi duomenys, tinkamas kitamųjų parinkimas, kintamųjų multikolinearumo nebuvimas. Daugialypio normalumo sąlyga keliama, kai faktorių išskyrimui naudojamas didžiausiojo tikėtinumo metodas. Šiap, kintamųjų pasiskirstymo pagal normalųjį dėsnį sąlyga nėra kritinė faktorinei analizei. Faktorinė analizė taikoma ir ranginiams kintamiesiems, turintiems suderintą, artimą intervalinei matavimų skalę. Tačiau binarinių duomenų atveju išskiriama per daug faktorių, kuriems priskiriami daugelis kintamųjų. Ranginės Likerto skalės duomenims yra plačiai taikomi intervalų skalės analizės metodai (tame tarpe, faktorinė analizė), kai skalė sudaroma iš ne mažiau kaip 5-ių , dar geriau – 7-ių reikšmių.

Faktorinės analizės etapai:

- Patikrinama, ar duomenys faktorinei analizei tinka;
- Faktorių išskyrimas – faktorių skaičiaus nustatymas;
- Faktorių sukimas ir interpretavimas;
- Faktorių reikšmių įverčių skaičiavimas;

4.1. MATEMATINIS FAKTORINĖS ANALIZĖS MODELIS

Bendriausias faktorinės analizės modelis, siejantis *k* kintamųjų $X_1, X_2, \cdots, X_k$ su *m* bendrųjų latentinių (nepastebėtų, nepastebimų) faktorių $F_1, F_2, \cdots, F_m$ ir specifiniu (charakteringuoju) latentiniu faktoriumi e_i aprašomas lygčių sistema (Čekanavičius, Murauskas, 2002)

$$X_i = \sum_{j=1}^{m} \lambda_{ij} Fj + e_i \,, \tag{4.1}$$

čia $i = 1, \cdots, k$, $m < k$, t. y. bendrųjų faktorių yra mažiau nei kintamųjų. Daugikliai λ_{ij} vadinami faktorių svoriais. Esant prielaidoms, kad:

- stebimi kintamieji pasiskirstę pagal normalųjį dėsnį, t. y. $X_i \sim N\left(\mu_i, \sigma_i^2\right)$,
- bendrieji faktoriai F_j yra nekoreliuoti ir jų dispersija $DF_j = 1$,
- charakteringieji faktoriai e_i nekoreliuoti ir jų dispersija $De_j = \tau_i$,
- faktoriai F_j ir e_i nekoreliuoti, čia $i = 1, \cdots, k$, $j = 1, \cdots, m$

stebimų kintamųjų dispersijas galima užrašyti taip (Čekanavičius, Murauskas, 2002):

$$DX_i = \sigma_i^2 = \lambda_{i1}^2 + \cdots + \lambda_{im}^2 + \tau_i = h_i^2 + \tau_i \,. \tag{4.2}$$

Dydis $h_i^2 = \sum_{j=1}^{m} \lambda_{ij}^2$ yra vadinamas kintamojo X_i bendrumu, o dydis τ_i – specifiškumu. Kuo didesnis h_i^2, palyginti su σ_i^2, tuo daugiau informacijos apie kintamąjį išsaugoma pereinant nuo pradinių kintamųjų prie bendrųjų faktorių.

Faktorinės analizės uždavinys – žinant X_i reikšmes, reikia padaryti išvadas apie bendruosius faktorius, sąlygojančius kintamųjų X_i elgseną, t. y. nustatyti faktorių svorių λ_{ij}, specifinių dispersijų τ_i (dispersijų, kurias lemia bendraisiais faktoriais nepaaaiškinama paties kintamojo variacija), o taip pat bendrųjų faktorių $F_1, F_2, \cdots, F_m$ reikšmių įverčius.

4.2. DUOMENŲ TINKAMUMAS FAKTORINEI ANALIZEI

Faktorinė analizė neturi prasmės nekoreliuotiems kintamiesiems. Todėl visų pirma reikia įsitikinti, ar stebimi kintamieji tarpusavyje koreliuoja. Tai padeda nustatyti Bartlett'o sferiškumo kriterijus, pagal kurį yra tikrinama hipotezė, kad kintamųjų koreliacijų matrica yra vienetinė, t. y. visi stebimi kintamieji yra nekoreliuoti. Jeigu taikant Bartlett'o sferiškumo kriterijų *p*-reikšmė yra didesnė už pasirinktąjį reikšmingumo lygmenį α, t. y. minėta hipotezė priimama, tai turimiems duomenims faktorinė analizė yra netaikytina. Tiesinis faktorinės analizės modelis yra taikomas tiesine priklausomybe susietų kintamųjų rinkiniui. Tai reiškia, kad dalinės koreliacijos koeficientai tarp dviejų kintamųjų turėtų būti maži palyginus su koreliacijos koeficientais tarp tų kintamųjų. Ar šiuo aspektu duomenys tinka faktorinei analizei, įvertina Kaizerio-Mejerio-Olkino (*KMO*) matas. *KMO* yra empirinių koreliacijos koeficientų reikšmių ir dalinių koreliacijos koeficientų rekšmių palyginamasis indeksas. Jis skaičiuojamas pagal formulę (Čekanavičius, Murauskas, 2002)

$$KMO = \frac{\sum\sum_{i \neq j} r_{ij}^2}{\sum\sum_{i \neq j} r_{ij}^2 + \sum\sum_{i \neq j} a_{ij}^2}, \qquad (4.3)$$

čia r_{ij} – kintamųjų X_i ir X_j koreliacijos koeficientas, a_{ij} – kintamųjų X_i ir X_j dalinės koreliacijos koeficientas.

Jei *KMO* mato reikšmė maža, tai nagrinėjamų kintamųjų faktorinė analizė nerezultatyvi. Maža *KMO* mato reikšmė rodo, kad kintamųjų porų koreliacija nėra paaiškinama kitais kintamaisiais. Laikoma, kad *KMO* turėtų būti ne mažesnis kaip 0,7, ribiniu atveju – ne mažesnis kaip 0,6 (Čekanavičius, Murauskas, 2002).

Kiekvieno kintamojo tinkamumo matą galima apskaičiuoti pagal formulę

$$MSA_i = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} a_{ij}^2}, \qquad (4.4)$$

t. y. vietoj visų įmanomų kintamųjų porų, įmamos tik atskiro kintamojo sudarytos poros. MSA_i yra vadinamas imties adekvatumo matu (*Measure of Sampling Adequacy – MSA*). Kintamuosius su mažomis MSA_i reikšmėmis tikslinga iš faktorinės analizės pašalinti.

4.3. FAKTORIŲ IŠSKYRIMAS

Panagrinėsime vieną iš dažniausiai nadojamų faktorių išskyrimo metodų, grindžiamą esminių komponenčių analize. Tarkime, turime k kintamųjų $X_1, X_2, \cdots, X_k$. Daugelio kintamųjų tarpusavio priklausomybė gali būti įvertinta jų koreliacijomis arba kovariacijomis, iš koreliacijos (kovariacijos) koeficientų suformuojant koreliacinę (kovariacinę) matricą. Taikant esminių komponenčių analizę, randamos tarpusavyje nekoreliuojančios kintamųjų $X_1, X_2, \cdots, X_k$ tiesinės kombinacijos $Y_1, Y_2, \cdots, Y_k$

$$Y_1 = \sum_{j=1}^{k} \alpha_{1j} X_j, \quad \cdots, \quad Y_k = \sum_{j=1}^{k} \alpha_{kj} X_j, \tag{4.5}$$

tenkinančios šias sąlygas:

1) $cov(Y_i, Y_j) = 0, \quad i, j = 1, \cdots, k, i \neq j$,
2) $DY_1 \geq DY_2 \geq \cdots \geq DY_k$,
3) $\sum_{i=1}^{k} DY_i = \sum_{i=1}^{k} DX_i$,

t. y., kintamieji $Y_1, Y_2, \cdots, Y_k$ turi būti nekoreliuoti ir išdėstyti dispersijų mažėjimo tvarka, o jų dispersijų suma turi būti lygi pradinių kintamųjų dispersijų sumai. Iš (4.5) seka, kad esminių komponenčių paieška – tai koeficientų $\alpha_{ij}, \quad i, j = 1, \cdots, k$ suradimas. Įrodyta, kad šie koeficientai yra pradinių kintamųjų kovariacijų matricos tikriniai (nuosavi) vektoriai. Nesigilindami į matricų algebros operacijas (išsamesnę analizę skaitytojas gali rasti Čekanavičius, Murauskas, 2002) pastebėsime, kad kovariacijos matricos C nuosavas vektorius (angl. *eigenvector*) $\vec{\alpha}_n$ ir jį atitinkanti nuosava reikšmė (angl. *eigenvalue*) λ_n yra lygties $C\vec{\alpha}_n = \lambda_n \vec{\alpha}_n$ sprendinys; λ_n reikšmė randama iš charakteringosios lygties $|C - \lambda_n I| = 0,$ čia I yra vienetinė matrica, kurios matmenys tokie pat, kaip ir matricos C. Gauta tiesinė priklausomybė $Y_1 = \alpha_{11} X_1 + \cdots + \alpha_{1k} X_k$ vadinama kintamųjų $X_1, X_2, \cdots, X_k$ pirmąja esmine komponente, kuri paaiškina $100 \cdot DY_1 / D$ procentų bendrosios dispersijos. Analogiškai tiesinė priklausomybė $Y_2 = \alpha_{21} X_1 + \cdots + \alpha_{2k} X_k$ vadinama antrąja esmine komponente, kuri paaiškina $100 \cdot DY_2 / D$ procentų bendrosios dispersijos, ir t. t. Kuo daugiau bendrosios kintamųjų dispersijos paaiškina pagrindinė komponentė, tuo jinai svarbesnė kaip akumuliuojanti informaciją apie kintamuosius. Pavyzdžiui, jei pagrindinė komponentė paaiškina 70% bendrosios dispersijos, galime teigti, kad palikdami vietoje pradinių kintamųjų $X_1, X_2, \cdots, X_k$ tik tą vieną komponentę, išlaikysime 70% informacijos apie pradinių kintamųjų reikšmių sklaidą. Visos pagrindinės komponentės (jų yra tiek, kiek ir pradinių kintamųjų) paaiškina visą bendrąją kintamųjų dispersiją, tačiau tik m pirmųjų komponenčių $Y_1, Y_2, \cdots, Y_m$, paaiškinančių didžiąją dalį bendrosios dispersijos, panaudojamos faktoriams nustatyti. Paprastai, m yra parenkamas lygus ne mažesnių už vienetą koreliacijos matricos tikrinių (nuosavų) reikšmių skaičiui. Taigi, turint k kintamųjų stebėjimus $(x_{1j}, x_{2j}, \cdots, x_{kj})$, $j = 1, \cdots, m$, iš pradžių apskaičiuojami k pagrindinių komponenčių įverčiai

$$\hat{Y}_i = \sum_{j=1}^{k} a_{ij} X_j, \quad i = 1, \cdots, k; \tag{4.6}$$

čia a_{ij} yra koeficientų α_{ij} empiriniai įverčiai. Latentiniais bendraisiais faktoriais laikomos m pirmųjų pagrindinių komponenčių, normuotų standartiniais nuokrypiais, t. y.

$$\hat{F}_j = \frac{\hat{Y}_j}{\sqrt{s^2\left(\hat{Y}_j\right)}}, \tag{4.7}$$

čia $s^2\left(\hat{Y}_i\right)$ yra j-osios pagrindinės komponentės dispersijos įvertis lygus j-tajai pagal dydį koreliacijų matricos tikrinei reikšmei, kurios atitinkamas tikrinis vektorius yra $\vec{a}_j = \left(a_{j1}, a_{j2}, \cdots, a_{jk}\right)', \quad j = 1, \cdots, k$. Faktorių svorių įverčiai išreiškiami lygybe

$$\hat{\lambda}_{ij} = a_{ij}\sqrt{s^2\left(\hat{Y}_j\right)}, \quad i = 1, \cdots, k; \quad j = 1, \cdots, m; \tag{4.8}$$

o specifinių faktorių įverčiai išreiškiami lygybe

$$\hat{e}_i = \sum_{j=m+1}^{k} a_{ij}\hat{Y}_j, \quad i = 1, \cdots, k. \tag{4.9}$$

Rezultate, kiekvienam kintamajam galima parašyti įvertį

$$\hat{X}_i = \sum_{j=1}^{m} \hat{\lambda}_{ij}\hat{F}_j + \hat{e}_i, \quad i = 1, \cdots, k. \tag{4.10}$$

Galioja paprasta taisyklė (Čekanavičius, Murauskas, 2002), (Garson, 2009) – faktorius F_j laikomas susijusiu su tais kintamaisiais $X_1, X_2, \cdots, X_k$, kuriems svorių įverčiai $\hat{\lambda}_{1j}, \cdots, \hat{\lambda}_{kj}$ absoliučiuoju dydžiu ne mažesni kaip 0,4. Teigiamas svoris rodo, kad kintamasis su faktoriumi koreliuoja teigiamai, o neigiamas svoris – neigiamai. Kintamieji yra vienodai svarbūs nepriklausomai nuo svorio ženklo.

4.4. FAKTORIŲ SUKIMAS IR INTERPRETAVIMAS

Gauta pradinė faktorių svorių matrica nenusako vienareikšmiškai sprendinio (tas pats kintamasis gali būti susijęs su keliais faktoriais ne mažesniais kaip 0,4 svoriais). Kad palengvinti faktorių diferencijaciją bei suteikti jiems lengviau interpretuojamą pavidalą, sudaromos tiesinės gautų faktorių kombinacijos, kurios tarpusavyje nekoreliuoja (yra ortogonalios). Šios naujų faktorių kombinacijų nustatymo procedūros, kuri vadinama ortogonaliuoju sukimu (rotacija), tikslas – suprastinti faktorių svorių matricos struktūrą, pasiekti, kad kiekvienas kintamasis turėtų tik kelis nenulinius faktorių svorius. Populiariausias iš ortogonaliųjų sukimų yra *Varimax* metodas. Transformuotos (pasuktos) matricos bendrieji faktoriai įvardijami remiantis kintamuosius, su kuriais koreliuoja atskiri faktoriai, vienijančiomis savybėmis. Interpretuojant faktorius negalima išvengti subjektyvumo, nes faktorių įvardijimas priklauso nuo tyrėjo kompetencijos, jo išsilavinimo.

4.5. FAKTORIŲ REIKŠMIŲ SKAIČIAVIMAS

Nustačius bendruosius faktorius, reikia apskaičiuoti šių faktorių reikšmes, kad būtų galima įvertinti jų atžvilgiu konkretų tiriamąjį. Faktorių įverčiai skaičiuojami mažiausių kvadratų metodu, regresinės analizės metodu. Pastarasis SPSS yra nustatytasis metodas. Laikant, kad faktoriai yra priklausomi kintamieji, o pradiniai kintamieji – nepriklausomi, galima užrašyti lygtį (Čekanavičius, Murauskas, 2002)

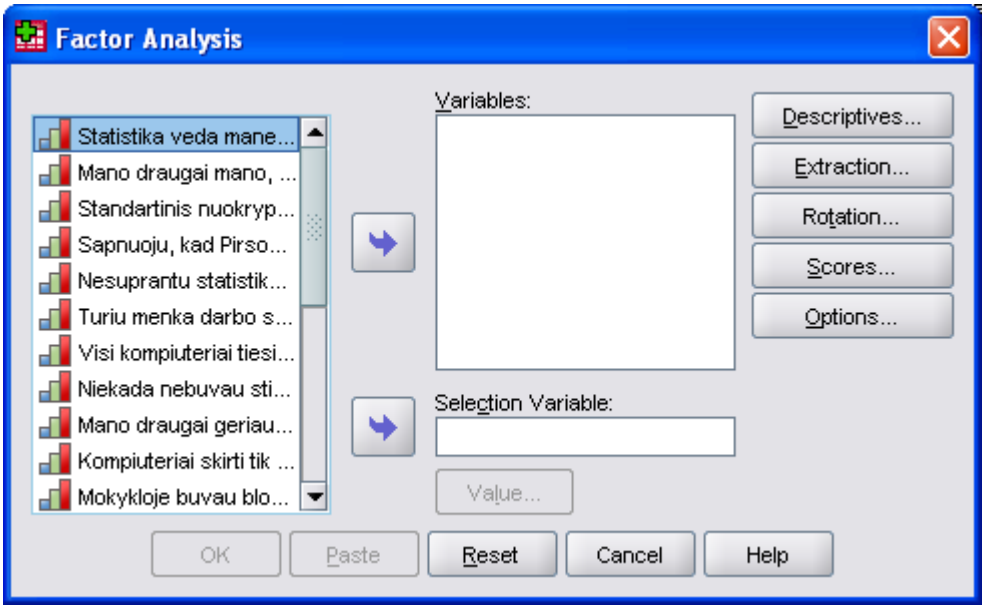
$$\hat{F}_j = \sum_{i=1}^{k} b_{ij} z_i, \quad j = 1, \cdots, m; \qquad (4.11),$$

čia $\hat{F}_j$ yra j-tojo faktoriaus reikšmė, z_i – i-tojo kintamojo standartizuota reikšmė, b_{ij} – regresijos koeficientų įverčiai.

4.6. FAKTORINĖS ANALIZĖS PAVYZDYS

Grįžkime prie skyriaus pradžioje minėto pavyzdžio. Norėdami atlikti faktorinę analizę:

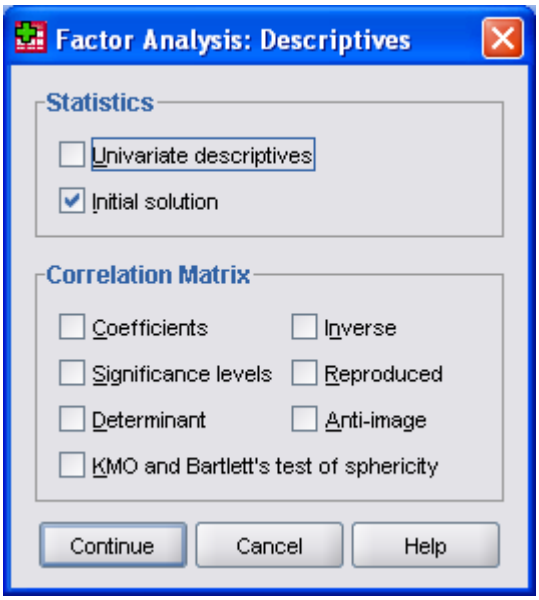
- Atidarykite bylą su analizuojamais duomenimis.
- Nurodykite komandas ***Analyze*** → ***Data Reduction*** →***Factor...*** Atsidarys dialogo langelis ***Factor Analysis*** (4.1 pav.).
- Įkelkite kintamuosius iš sąrašo į laukelį ***Variables***.
- Spragtelėkite dialogo langelio ***Factor Analysis*** mygtuką ***Descriptives...***



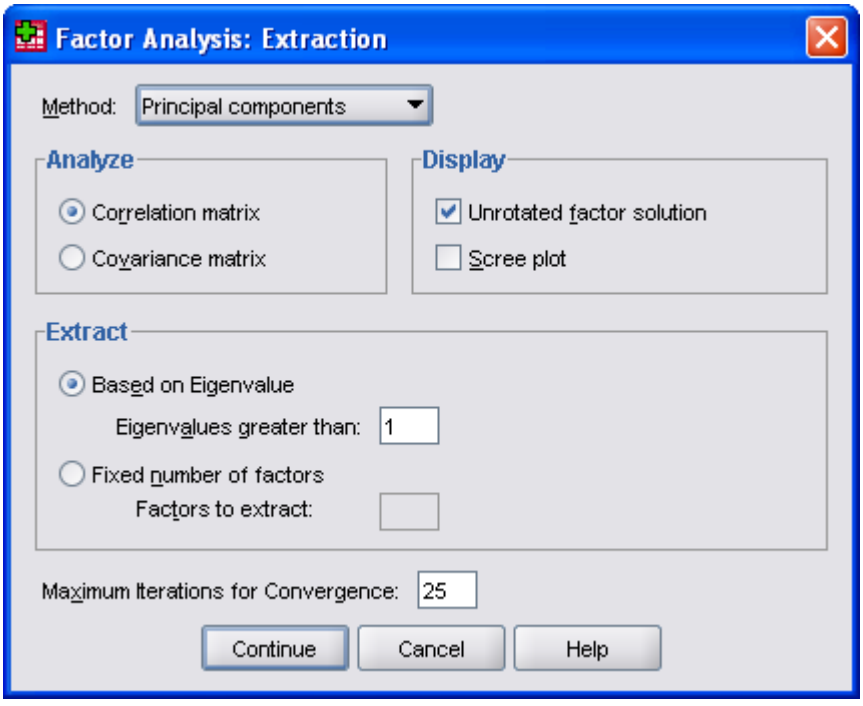
4.1 pav. Dialogo langelis ***Factor Analysis***

- Atsidariusiame naujame dialogo langelyje ***Factor Analysis: Descriptives*** (4.2 pav.) pažymėkite laukelį ***KMO and Bartlett's test of sphericity***, o taip pat laukelį ***Anti-image*** *MSA* matui kiekvienam kintamajam apskaičiuoti. Spragtelėkite dialogo langelio mygtuką ***Continue***.
- Spragtelėkite dialogo langelio ***Factor Analysis*** mygtuką ***Extraction...***

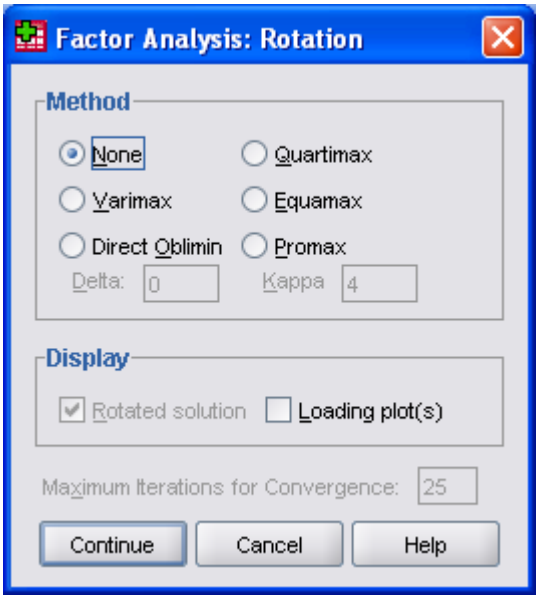
- Dialogo langelyje ***Factor Analysis: Extraction*** (4.3 pav.) palikite nustatytą esminių komponenčių (***Principal component***) metodą bei kitus nustatytus pažymėjimus.
- Spragtelėkite dialogo langelio ***Factor Analysis*** mygtuką ***Rotation...*** ir atsidariusiame dialogo langelyje ***Factor Analysis: Rotation*** (4.4 pav.) pažymėkite ***Varimax*** laukelį bei palikite pažymėtą ***Rotated solution*** laukelį.



4.2 pav. Dialogo langelis ***Factor Analysis: Descriptives***

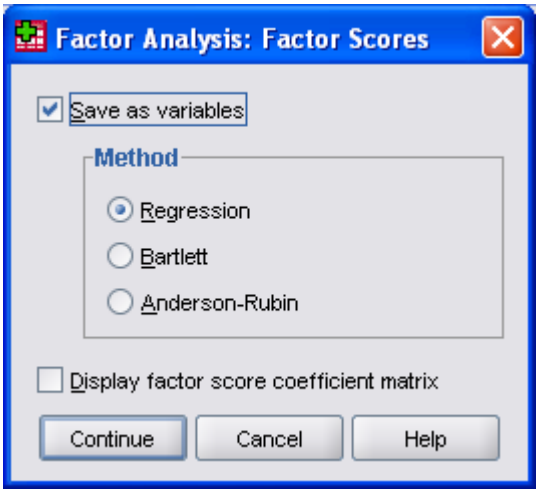


4.3 pav. Dialogo langelis ***Factor Analysis: Extraction***



4.4 pav. Dialogo langelis ***Factor Analysis: Rotation***

- Spragtelėkite dialogo langelio ***Factor Analysis*** mygtuką ***Scores...*** ir atsidariusiame dialogo langelyje ***Factor Analysis: Factor Scores*** (4.5 pav.) pažymėkite laukelį ***Save as variables*** faktorių reikšmių išsaugojimui duomenų rinkmenoje. Palikite nustatytą regresinį faktorių reikšmių skaičiavimo metodą.
- Spragtelėkite dialogo langelio ***Factor Analysis*** mygtuką ***Options...*** ir atsidariusiame dialogo langelyje ***Factor Analysis: Options*** (4.6 pav.) pažymėkite laukelį ***Supress small coefficients*** ir nustatykite ***Absolute value below*** reikšmę, lygią 0,4.

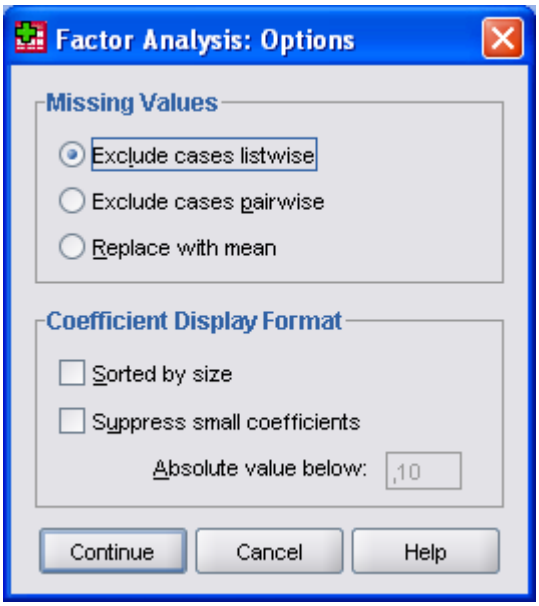


4.5 pav. Dialogo langelis ***Factor Analysis: Factor Scores***

- Spragtelėkite mygtuką ***OK*** pagrindiniame dialogo langelyje ***Factor Analysis***. Padrindinės faktorinės analizės išvesties lentelės parodytos 4.7.1–4.7.4 pav.

Lentelėje *KMO and Bartlett's Test* (4.7.1 pav.) pateikta *KMO* ktiterijaus reikšmė bei Bartlett'o sferiškumo kriterijaus *p*-reikšmė. Kadangi Bartlett'o sferiškumo kriterijaus *p*-reikšmė $p < 0{,}05$ rodo, kad kintamieji nėra nepriklausomi, o *KMO* = 0,930 – duomenys puikiai tinka faktorinei analizei. *MSA* reikšmės kiekvienam kintamajam yra nurodytos lentelės *Anti-image Matrices* (dėl didelės apimties čia nepateikiama) dalinių koreliacijos koeficientų

(su priešingu ženklu) matricos *Anti-image Correlation* įstrižajinėje ir pažymėtos išnaša *Measure of Sampling Adequacy*. Šiame pavyzdyje *MSA* reikšmės yra 0,875-0,967 ribose, kas reiškia, kad visi kintamieji tinkami faktorinei analizei.



4.6 pav. Dialogo langelis ***Factor Analysis: Options***

Lentelėje *Communalities* (4.7.2 pav.) pateikti pradinių kintamųjų bendrumai – pradinių kintamųjų variacijų dalys, kurios paaiškinamos bendraisiais faktoriais. Teigiama (Čekanavičius, Murauskas, 2002), kad atrinktosiose pagrindinėse komponentėse išliko pakankamai daug informacijos apie kintamąjį, jeigu jo bendrumas ne mažesnis kaip 0,20.

Lentelėje *Total Variance Explained* (4.7.3 pav.) nurodoma, kokią bendrosios kintamųjų dispersijos dalį paaiškina kiekviena pagrindinė komponentė (stulpelyje *Extraction Sums of Squared Loadings – % of Variance*), o taip pat kokią suminę bendrosios kintamųjų dispersijos dalį paaiškina pirmosios pagrindės komponentės (stulpelyje *Extraction Sums of Squared Loadings – Cumulative %*). Šiame pavyzdyje keturios pagrindinės komponentės, kurių nuosavos reikšmės (*Eigenvalues*) didesnės už 1, paaiškina 50,317 % bendrosios kintamųjų dispersijos. Stulpeliuose *Rotation Sums of Squared Loadings – % of Variance* ir *Cumulative %* nurodoma, kokią bendrosios kintamųjų dispersijos dalį paaiškina kiekviena pagrindinė komponentė, o taip pat kokią suminę bendrosios kintamųjų dispersijos dalį paaiškina pirmosios pagrindės komponentės galutiniame rezultate, po komponentų matricos sukimo procedūros.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,930
Bartlett's Test of Sphericity	Approx. Chi-Square	19334,492
	df	253
	Sig.	,000

4.7.1 pav. Pagrindiniai faktorinės analizės rezultatai

Communalities

	Initial	Extraction
Statistika veda mane prie asaru	1,000	,435
Mano draugai mano, kad as per kvailas(a) susidoroti su SPSS	1,000	,414
Standartinis nuokrypis mane dirgina	1,000	,530
Sapnuoju, kad Pirsonas atakuoja mane su koreliacijos koeficientu	1,000	,469
Nesuprantu statistikos	1,000	,343
Turiu menka darbo su kompiuteriais patirti	1,000	,654
Visi kompiuteriai tiesiog nekencia manes	1,000	,545
Niekada nebuvau stiprus(i) matematikoje	1,000	,739
Mano draugai geriau uz mane ismano statistika	1,000	,484
Kompiuteriai skirti tik zaidimams	1,000	,335
Mokykloje buvau blogas matematikas	1,000	,690
Zmones nori jums iteigti, kad SPSS daro statistika paprastesne, bet taip nera	1,000	,513
Bijau, kad del savo nekompetencijos sugadinsiu kompiuteri	1,000	,536
Kompiuteris yra nedraugiskas vartotojui	1,000	,488
Kompiuteriai man neiveikiami	1,000	,378
Aiskinimai apie centrine tendencija veda mane prie verksmo	1,000	,487
Lygtys veda mane prie komos	1,000	,683
SPSS visada luzta, kai tik as pradedu su juo dirbti	1,000	,597
Visi spokso i mane kai as dirbu su SPSS	1,000	,343
Per tuos vektorius as negaliu miegoti	1,000	,484
Sapnuoju normaluji pasiskirstyma	1,000	,550
Mano draugai geriau ismano SPSS negu as	1,000	,464
Jeigu gerai ismanysiu statistika, draugai laikys mane nuobodziu(ia)	1,000	,412

Extraction Method: Principal Component Analysis.

4.7.2 pav. Pagrindiniai faktorinės analizės rezultatai

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	7,290	31,696	31,696	7,290	31,696	31,696	3,730	16,219	16,219
2	1,739	7,560	39,256	1,739	7,560	39,256	3,340	14,523	30,742
3	1,317	5,725	44,981	1,317	5,725	44,981	2,553	11,099	41,841
4	1,227	5,336	50,317	1,227	5,336	50,317	1,950	8,476	50,317
5	,988	4,295	54,612						
6	,895	3,893	58,504						
7	,806	3,502	62,007						
8	,783	3,404	65,410						
9	,751	3,265	68,676						
10	,717	3,117	71,793						
11	,684	2,972	74,765						
12	,670	2,911	77,676						
13	,612	2,661	80,337						
14	,578	2,512	82,849						
15	,549	2,388	85,236						
16	,523	2,275	87,511						
17	,508	2,210	89,721						
18	,456	1,982	91,704						
19	,424	1,843	93,546						
20	,408	1,773	95,319						
21	,379	1,650	96,969						
22	,364	1,583	98,552						
23	,333	1,448	100,000						

Extraction Method: Principal Component Analysis.

4.7.3 pav. Pagrindiniai faktorinės analizės rezultatai

Lentelėje *Rotated Component Matrix* (4.7.4 pav.) yra pateikiami išskirtų keturių faktorių svoriai po sukimo procedūros. Matyti, kad pirmasis faktorius koreliuoja su kintamaisiais, kuriuos apibendrintai galima apibūdinti kaip tiriamųjų nepasitikėjimą savo darbo kompiuteriu įgūdžiais, antras faktorius koreliuoja su kintamaisiais, kurie reiškia tiriamųjų neigiamą nusistatymą statistikos dalyko atžvilgiu, trečią faktorių galėtume įvertinti kaip silpną matematikos dalyko žinojimą, o ketvirtą faktorių – baimė pasirodyti prieš draugus nekompetetingu. Taigi, įvardijant išskirtus faktorius svarbų vaidmenį vaidina ir subjektyvusis veiksnys, t. y. tyrėjo pozicija.

Be dažniausiai naudojamo faktorių išskyrimui esminių komponenčių analize pagrįsto metodo, dialogo langelyje ***Factor Analysis: Extraction*** (4.3 pav.) galima pasirinkti kitus, rečiau naudojamus faktorių išskyrimo metodus (Garson):

1. ***Image factoring***. Metodo esmė yra ta, kad esminių komponenčių analizė yra taikoma ne stebimų kintamųjų koreliacinei matricai, o numatytų kintamųjų koreliacinei matricai, kada kiekvienas numatytasis kintamasis yra išskaičiuojamas iš kitų pagal daugialypę regresiją.
2. ***Maximum likelihood***. Faktoriai išskiriami naudojant didžiausiojo tikėtinumo analizės metodus. Turi būti tenkinama duomenų normaliojo skirstinio sąlyga. Išvesties lentelėje *Goodness-of-fit Test* (4.8 pav.) pateikiamas χ^2 kriterijaus rezultatas modelio tinkamumui įvertinti. Esant χ^2 kriterijaus *p*-reikšmei daugiau už nustatytą reikšmingumo lygmenį, tyrėjas gali didinti faktorių skaičių, kol bus pasiektas statistiškai reikšmingas rezultatas. Tačiau egzistuoja rizika išskirti per daug faktorių.

Rotated Component Matrix[a]

	Component			
	1	2	3	4
Statistika veda mane prie asaru		,496		
Mano draugai mano, kad as per kvailas(a) susidoroti su SPSS				,543
Standartinis nuokrypis mane dirgina		-,567		
Sapnuoju, kad Pirsonas atakuoja mane su koreliacijos koeficientu		,516		
Nesuprantu statistikos		,429		
Turiu menka darbo su kompiuteriais patirti	,800			
Visi kompiuteriai tiesiog nekencia manes	,638			
Niekada nebuvau stiprus (i) matematikoje			,833	
Mano draugai geriau uz mane ismano statistika				,648
Kompiuteriai skirti tik zaidimams	,550			
Mokykloje buvau blogas matematikas			,747	
Zmones nori jums iteigti, kad SPSS daro statistika paprastesne, bet taip nera	,473	,523		
Bijau, kad del savo nekompetencijos sugadinsiu kompiuteri	,647			
Kompiuteris yra nedraugiskas vartotojui	,579			
Kompiuteriai man neiveikiami	,459			
Aiskinimai apie centrine tendencija veda mane prie verksmo		,514		
Lygtys veda mane prie komos			,747	
SPSS visada luzta, kai tik as pradedu su juo dirbti	,684			
Visi spokso i mane kai as dirbu su SPSS				,428
Per tuos vektorius as negaliu miegoti		,677		
Sapnuoju normaluji pasiskirstyma		,661		
Mano draugai geriau ismano SPSS negu as				,645
Jeigu gerai ismanysiu statistika, draugai laikys mane nuobodziu(ia)				,586

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 8 iterations.

4.7.4 pav. Pagrindiniai faktorinės analizės rezultatai

Goodness-of-fit Test

Chi-Square	df	Sig.
1153,631	167	,000

4.8 pav. Išvesties lentelė Goodness-of-fit Test

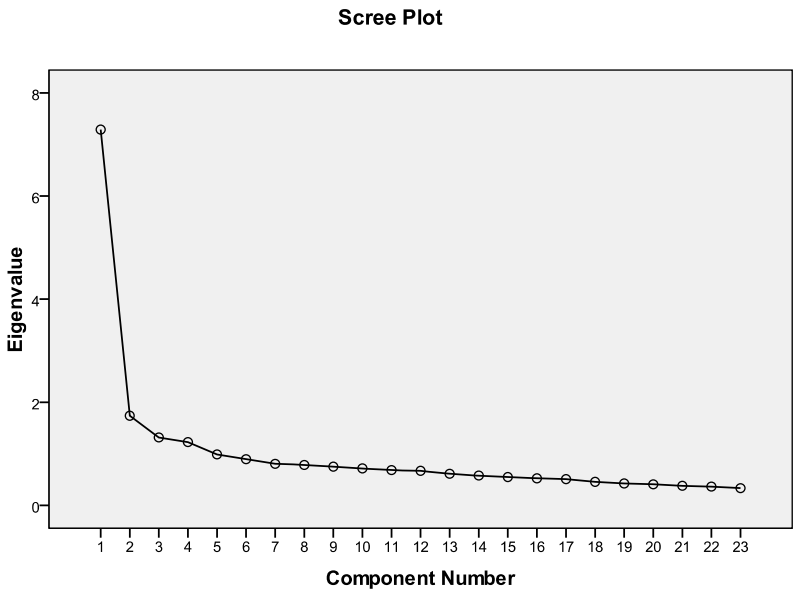
3. ***Alpha factoring***. Pagal šį metodą tiriami kintamieji yra traktuojami kaip atsitiktinai atrinkti iš kintamųjų aibės (pagal kitus metodus kintamieji yra nustatyti, stebėjimai – atrinkti).
4. ***Unweighted least squares***. Metodas grindžiamas skirtumų tarp stebimų duomenų koreliacijos matricos ir numatytų duomenų koreliacijos matricos kvadratų sumos (neįvertinant matricų įstrižajinėse esančių duomenų) minimizavimu.
5. ***Generalized least squares***. Tai modifikuotas ***Unweighted least squares*** metodas suteikiant kintamųjų koreliacijai atitinkamus svorius. Kaip ir ***Maximum likelihood*** metodo atveju, išvestyje pateikiamas χ^2 kriterijaus rezultatas modelio tinkamumui įvertinti. Esant χ^2 kriterijaus *p*-reikšmei daugiau už nustatytą reikšmingumo lygmenį, tyrėjas gali didinti faktorių skaičių, kol bus pasiektas statistiškai reikšmingas rezultatas.
6. ***Principal axis factoring***. Metodas taikomas, kai tyrėjas kelia sau uždavinį nustatyti latentinius faktorius, įtakojančius bendrą analizuojamų kintamųjų dispersiją, neįvertinant specifinių dispersijų, t. y. dispersijų, kurias lemia bendraisiais faktoriais nepaaaiškinama paties kintamojo variacija.

Dialogo langelyje ***Factor Analysis: Extraction*** (4.3 pav.) nustatytasis faktorių skaičius yra lygus ne mažesnių už vienetą koreliacijos matricos tikrinių (nuosavų) reikšmių skaičiui (***Eigenvalues greater than 1***). Tačiau tai nėra vienintelis galimas faktorių skaičiaus nustatymo variantas. Dar paminėtini šie, įdiegti SPSS 17.0 versijoje, bet rečiau naudojami metodai (Garson, 2009):

1. ***Scree plot***. Pažymėję dialogo langelyje ***Factor Analysis: Extraction*** (4.3 pav.), ***Display*** srityje laukelį ***Scree plot*** gausime 4.9 pav. pavaizduotą grafiką, kuriame X ašyje atidėtos komponentės pagal jų svarbą, o Y ašyje – atitinkamos tikrinės šių komponenčių reikšmės. Komponentės, kurios yra už taško, kuriame staigus kreivės kritimas pereina į palaipsninį mažėjimą, yra laikomos nereikšmingomis. Metodas turi esminį trūkumą, nes kreivė ne visada turi aiškiai išreikštą alkūnę ir didelį vaidmenį vaidina subjektyvusis faktorius.
2. Atkurtos koreliacijos liekanų (*Reproduced correlation residuals*) metodas. Pažymėję dialogo langelyje ***Factor Analysis: Descriptives*** (4.2 pav.) ***Correlation Matrix*** srityle laukelį ***Reproduced*** gausime atkurtąją koreliacinę matricą ir skirtuminę koreliacinę matricą tarp atkurtosios koreliacinės matricos ir stebimos koreliacinės matricos. Atkurtoji koreliacinė matrica yra skaičiuojama faktorių koreliacijos su kintamaisiais pagrindu. Taip, dviejų kintamųjų atkurtasis koreliacijos koeficientas yra lygus koreliacijos koeficientų tarp šių kintamųjų ir kiekvieno faktoriaus sandaugos sumai (Norušis, 2006), t. y.

$$r_{ij} = \sum_{f=1}^{m} r_{fi} r_{fj} = r_{1i} r_{1j} + r_{2i} r_{2j} + \cdots + r_{mi} r_{mj} . \qquad (4.12)$$

SPSS išvestyje, išnašoje po lentele *Reproduced Correlations* (lentelė, kuri yra dviejų dalių – viršutinėje dalyje pateikiamos atkurtosios koreliacines reikšmes, apatinėje dalyje pateikiamos koreliacijos liekanų reikšmės, dėl didelės apimties čia nepateikiama) yra nurodomas koreliacijos liekanų, didesnių už 0,05, procentas. Kuo jis mažesnis, tuo faktorinės analizės modelis yra geresnis. Operuodamas su skirtingais faktorių skaičiais, tyrėjas gali bandyti nustatyti tą priimtiną faktorių skaičių, kuris užtikrintų galimai mažesnį statistiškai reikšmingų koreliacijos liekanų procentą.



4.9 pav. Grafikas Scree Plot

Kaip pažymėta 4.4 poskyryje, tam, kad palengvinti faktorių diferencijaciją bei suteikti jiems lengviau interpretuojamą pavidalą, yra atliekama pradinės faktorių svorių matricos transformacijos (sukimo) procedūra. Sukimo išdavoje tikrinių reikšmių suma nepakinta, bet keičiasi atskirų faktorių tikrinės reikšmės (tuo pačiu ir paaiškinama tais faktoriais bendrosios kintamųjų dispersijos dalis) ir tų faktorių svoriai. Taikant skirtingus sukimo metodus, gaunami skirtingi faktorių svoriai (nekeičiant tikrinių reikšmių sumos). Tai gali turėti įtakos faktorių interpretacijai ir įvardijimui, todėl kartais tikslinga taikyti kelis alternatyvius sukimo metodus, kad gauti lengviausiai interpretuojamą faktorių struktūrą. SPSS pakete naudojami šie sukimo metodai, kuriuos galima pasirinkti dialogo langelyje ***Factor Analysis: Rotation*** (4.4 pav.):

1. ***Varimax***. Tai ortogonalusis faktorių ašių sukimo metodas, kurio rezultate pasiekiama didžiausia faktorių svorių susijusiems kintamiesiems dispersijos reikšmė, t. y. kiekvieno faktoriaus svoris kiekvienam kintamajam įgija galimai didesnę ar mažesnę reikšmę. Tai įgalina susieti kiekvieną kintamąjį su vienu išskirtu faktoriumi. ***Varimax*** yra dažniausiai naudojamas faktorių matricos sukimo metodas.
2. ***Quartimax***. Tai taip pat ortogonalusis faktorių ašių sukimo metodas, minimizuojantis faktorių, būtinų kiekvieno kintamojo paaiškinimui, skaičių. Tačiau šiuo metodu dažnai išskiriami bendri faktoriai, turintys didelius svorius daugumai kintamųjų.
3. ***Equamax***. Tai tarpinis variantas tarp ***Varimax*** ir ***Quartimax***.
4. ***Direct Oblimin***. Tai neortogonalusis (įžambusis) faktorių ašių sukimo metodas, po kurio naujieji bendrieji faktoriai jau koreliuoti. Metodas įgalina gauti didesnes tkrines

reikšmes, bet pasunkina faktorių interpretaciją. Kartais naudojamas realiam (ne matematiniam) faktorių ortogonalumui patvirtinti. Tai sprendžiama pagal faktorių koreliacijos lentelę *Component Correlation Matrix* (4.10 pav.). Faktoriai laikomi ortogonaliais, jeigu koreliacijos koeficiento reikšmė savo absoliutine išraiška ne didesnė už 0,32 (Garson, 2009). Diskriminantinis pagrįstumas (*Discriminant validity*) yra suprantamas, kad koreliacija tarp faktorių nėra per didelė. Jeigu koreliacijos koeficiento reikšmė yra didesnė už 0,85, laikoma, kad faktoriai pagal savo suvokimą persidengia (Garson, 2009).

5. ***Promax***. Metodas panašus į ***Direct Oblimin***, tik paprastesnis skaičiavimas, todėl gali būti taikomas didelės apimties duomenų rinkmenoms.

Taikant ortogonalųjį sukimo metodą pateikiama koreliacijos tarp faktorių prieš sukimą ir po sukimo lentelė *Component Transformation Matrix* (4.10 pav.)

Component Correlation Matrix

Component	1	2	3	4
1	1,000	-,153	,360	-,277
2	-,153	1,000	-,193	,093
3	,360	-,193	1,000	-,464
4	-,277	,093	-,464	1,000

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

Component Transformation Matrix

Component	1	2	3	4
1	,635	,585	,443	-,242
2	,137	-,167	,488	,846
3	,758	-,513	-,403	,008
4	,067	,605	-,635	,476

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

4.10 pav. Išvesties lentelės: Component Correlation Matrix ir Component Transformation Matrix

Nurodę dialogo langelyje ***Factor Analysis: Factor Scores*** (4.5 pav.) išsaugoti faktorių reikšmes visiems įrašams duomenų rinkmenoje (pažymėję laukelį ***Save as variables***), galime gauti taip pat faktorių reikšmių koeficientų kiekvienam kintamajam reikšmes, papildomai pažymėję laukelį ***Display factor score coefficient matrix***. Esant nustatytam ***Method*** variantui ***Regression*** šie koeficientai išvesties lentelėje *Component Score Coefficient Matrix* (4.11 pav.) reiškia regresijos koeficientų įverčius, t. y. išsaugota duomenų rinkmenoje *k*-tojo faktoriaus reikšmė *i*-tajam kintamajam yra apskaičiuojama pagal 4.11 formulę, kurioje b_{ij} ir yra lentelėje *Component Score Coefficient Matrix* nurodyti koeficientai. Reikia pabrėžti, kad faktorių reikšmių koeficientai ir faktorių svoriai yra skirtingi dalykai. Faktorių reikšmių koeficientai naudojami faktorių reikšmėms apskaičiuoti pagal 4.11 lygtį, tuo tarpu faktorių svoriai yra daugikliai 4.1 lygtyje, nusakančioje kintamųjų priklausomybę nuo faktorių.

Component Score Coefficient Matrix

	Component			
	1	2	3	4
Statistika veda mane prie asaru	-,053	,173	,089	,110
Mano draugai mano, kad as per kvailas(a) susidoroti su SPSS	,102	-,129	,086	,282
Standartinis nuokrypis mane dirgina	,087	-,195	,013	,137
Sapnuoju, kad Pirsonas atakuoja mane su koreliacijos koeficientu	-,011	,170	,045	,107
Nesuprantu statistikos	,021	,131	,014	,083
Turiu menka darbo su kompiuteriais patirti	,383	-,211	-,088	,014
Visi kompiuteriai tiesiog nekencia manes	,213	,004	-,078	,038
Niekada nebuvau stiprus(i) matematikoje	-,129	-,074	,460	,013
Mano draugai geriau uz mane ismano statistika	,025	-,029	,108	,354
Kompiuteriai skirti tik zaidimams	,244	-,161	-,021	-,036
Mokykloje buvau blogas matematikas	-,066	-,087	,379	-,059
Zmones nori jums iteigti, kad SPSS daro statistika paprastesne, bet taip nera	,097	,161	-,116	,051
Bijau, kad del savo nekompetencijos sugadinsiu kompiuteri	,224	-,065	-,019	,013
Kompiuteris yra nedraugiskas vartotojui	,180	,040	-,084	,043
Kompiuteriai man neiveikiami	,114	-,055	,061	-,058
Aiskinimai apie centrine tendencija veda mane prie verksmo	-,015	,146	,046	,014
Lygtys veda mane prie komos	-,057	-,067	,372	,005
SPSS visada luzta, kai tik as pradedu su juo dirbti	,242	-,001	-,104	,043
Visi spokso i mane kai as dirbu su SPSS	,048	-,115	,061	,199
Per tuos vektorius as negaliu miegoti	-,195	,359	-,061	-,002
Sapnuoju normaluji pasiskirstyma	-,039	,270	-,064	,059
Mano draugai geriau ismano SPSS negu as	-,036	,162	-,048	,382
Jeigu gerai ismanysiu statistika, draugai laikys mane nuobodziu(ia)	,032	,211	-,162	,379

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

4.11 pav. Išvesties lentelė Component Score Coefficient Matrix

Duomenų rinkmenoje išsaugotos faktorių reikšmės gali būti panaudotos hipotezėms apie faktorių homogeniškumą pagal įvairias tiriamas grupes tikrinti. Dažnai faktorių reikšmės suskirstomos kvartiliais, o hipotezės apie rangų skalei priklausančių kintamųjų homogeniškumą tikrinamos 2 skyriuje aprašytu χ^2 kriterijumi.

5. BINARINĖ LOGISTINĖ REGRESIJA

5.1. BINARINĖS LOGISTINĖS REGRESIJOS MODELIS IR STATISTINĖS IŠVADOS

Binarinė logistinė regresija (Čekanavičius, Murauskas, 2002), (Norušis, 2005), (Garson, 2009) taikoma priklausomų binarinių kintamųjų reikšmėms prognozuoti. Kadangi binarinis kintamasis paprastai reiškia įvykį, kuris gali įvykti arba neįvykti, naudojant binarinės logistinės regresiją, yra apskaičiuojama šio įvykio tikimybė, sąlygojama nepriklausomų kintamųjų, kurie gali būti tiek intervaliniai, tiek kategoriniai. Kategorinis kintamasis į regresijos modelį įtraukiamas ne tiesiogiai, o pakeičiant jį vienu ar keliais dvireikšmiais pseudokintamaisiais. Kai kategorinis kintamasis turi $m > 2$ kategorijų jis keičiamas $(m-1)$ pseudokintamuoju. Visi pseudokintamieji gali įgyti tik dvi reikšmes: 1 arba 0. Binarinės logistinės regresijos populiarumas dalinai paaiškinamas tuom, kad ji įgalina išvengti daugelį tiesinei regresijai keliamų reikalavimų. Binarinė logistinė regresija tinka galiojant gana bendroms prielaidoms – nepriklausomi kintamieji neturi būti multikolinearūs, t. y. tarp jų neturi būti stipriai koreliuojančių, nereikalaujama šių kintamųjų normalumo, tiesinės priklausomo kintamojo ir nepriklausomų kintamųjų priklausomybės bei priklausomo kintamojo homoskedastiškumo, t. y., kad su kiekviena fiksuota nepriklausomo kintamojo reikšme galimų priklausomo kintamojo reikšmių sklaida būtų vienoda. Tikimybė p_i, kad atsitiktinis dydis Y_i įgys reikšmę 1 (iš įmanomų 1 ir 0), apskaičiuojama pagal formulę

$$p_i = \frac{\exp\{z(\vec{x}_i)\}}{1+\exp\{z(\vec{x}_i)\}},\ z(\vec{x}_i) = a + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki}, \tag{5.1}$$

čia $x_{1i}, \cdots x_{ki}$ – fiksuotos nepriklausomų kintamųjų $X_1, \cdots, X_k$ reikšmės, $i = 1, 2, \cdots, n$, n – stebėjimų skaičius. Tikimybių santykis $p_i/(1-p_i)$, t. y. tikimybės, kad Y_i įgys reikšmę 1, santykis su tikimybe, kad Y_i įgys reikšmę 0, vadinamas galimybe įvykti įvykti įvykiui $Y_i = 1$. Galimybės logaritmas nuo kintamųjų reikšmių priklauso tiesiškai, t. y.

$$\ln\frac{p_i}{1-p_i} = a + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} = z(\vec{x}_i). \tag{5.2}$$

Parametrų $a, b_1, \cdots, b_k$ įverčiai $\hat{a}, \hat{b}_1, \cdots, \hat{b}_k$ parenkami taip, kad tikėtinumo funkcija (Čekanavičius, Murauskas, 2002)

$$L = \prod_{i:y_i=1} p_i \prod_{i:y_i=0} (1-p_i) \tag{5.3}$$

būtų didžiausia. Turėdami parametrų įverčius galime apskaičiuoti ir tikimybės $P(Y=1)$ įvertį, kai nepriklausomų kintamųjų reikšmių $x_1, x_2, \cdots, x_k$ vektorius yra $\vec{x} = (x_1, x_2, \cdots, x_k)$

$$\hat{P}(Y=1 \mid \vec{x}) = \frac{\exp\{\hat{z}(\vec{x})\}}{1+\exp\{\hat{z}(\vec{x})\}},\ \hat{z}(\vec{x}) = \hat{a} + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \cdots + \hat{b}_k x_k \tag{5.4}$$

Koeficientų $\hat{b}_j$ reikšmės parodo kiek pasikeis galimybių santykio logaritmas, padidėjus atitinkamo nepriklausomo kintamojo reikšmei vienu vienetu, kai likusieji nepriklausomi kintamieji yra fiksuoti.

Jeigu $\hat{P}(Y=1|\,\vec{x})$ įvertis yra daugiau negu 0,5, prognozuojama, kad *Y* reikšmė yra 1; jeigu $\hat{P}(Y=1|\,\vec{x})$ įvertis yra mažiau negu 0,5, prognozuojama, kad *Y* reikšmė yra 0. Kadangi $\hat{P}(Y=1\,|\,\vec{x})>0{,}5$ tada ir tik tada, kai $\hat{z}(\vec{x})>0$, *Y* reikšmėms prognozuoti patogiau taikyti šią taisyklę: jeigu $\hat{z}(\vec{x})>0$, tai prognozuojama, kad $Y=1$, jeigu $\hat{z}(\vec{x})<0$, tai prognozuojama, kad $Y=0$.

Ar binarinė logistinė regresija prognozėms tinka yra sprendžiama pagal teisingų prognozių procentą, t. y. kiek sutampa žinomos y_i reikšmės su reikšmėmis, gautomis remiantis logistine regresija. Jeigu duomenų prognozės netikslios, tai binarinė logistinė regresija netaikytina, net jei kiti rodikliai bei statistinės išvados rodytų, kad ji tinkama (Čekanavičius, Murauskas, 2002). Ar binarinės logistinės regresijos modelis apskritai yra taikytinas, t. y. bent vienas koeficientas $b_j\neq 0$, sprendžiama pagal modelio χ^2 suderinamumo kriterijų. χ^2 kriterijus remiasi tuom, kad didžiausio tikėtinumo funkcijos (5.3) maksimumas $L(\hat{a},\hat{b})$, apskaičiuotas parametrams $a, b_1, \cdots, b_k$ mažai skiriasi nuo didžiausio tikėtinumo funkcijos maksimumo $L(\hat{a},0)$, apskaičiuoto esant prielaidai, kad visi $b_j=0$. Kriterijaus statistika apskaičiuojama pagal formulę (Čekanavičius, Murauskas, 2002), (Norušis)

$$\chi^2=-2\ln L(\vec{a},0)+2\ln L(\hat{a},\hat{b}) \qquad (5.5)$$

Hipotezę H_0, teigiančią, kad visi $b_j=0$, atmetame, t. y. bent vienas $b_j\neq 0$, kai kriterijaus *p*-reikšmė mažesnė už reikšmingumo lygmenį ($p<\alpha$). Neatmesta nulinė hipotezė H_0 rodo, kad binarinės logistinės regresijos modelis netinka.

Alternatyvus χ^2 suderinamumo kriterijui yra Hosmer'io-Lemeshow'o kriterijus, kuris pagrįstas vertinimu, kiek sutampa padalintų į atskiras grupes pagal variacinę prognozuojamų tikimybių eilutę žinomos stebėjimų y_i reikšmės su reikšmėmis, gautomis remiantis logistine regresija. Žinomų ir tikėtinų reikšmių skirtumą aprašo χ^2 skirstinys su $(k-2)$ laisvės laipsnių, čia *k* – grupių skaičius. Kai Hosmer'io-Lemeshow'o statistikos *p*-reikšmė $p>\alpha$, gauname binarinės logistinės regresijos modelio suderinamumo su duomenimis patvirtinimą, t. y. nulinė hipotezė, teigianti, kad nėra skirtumo tarp stebimų ir binarinės logistinės regresijos modelio pagrindu prognozuojamų priklausomo kintamojo reikšmių, yra neatmestina. Hosmer'io-Lemeshow'o suderinamumo kriterijus taikytinas, kai stebėjimų yra pakankamai daug, t. y., į kiekvieną grupę (jų sudaroma apie 10) pakliūtų bent 5 stebėjimai.

Ar konkretus koeficientas $b_j\neq 0$ parodo Wald'o kriterijus. Sprendimo priėmimo taisyklė yra klasikinė: hipotezė H_0 atmetama (taigi $b_j\neq 0$), jeigu *p*-reikšmė $p<\alpha$.

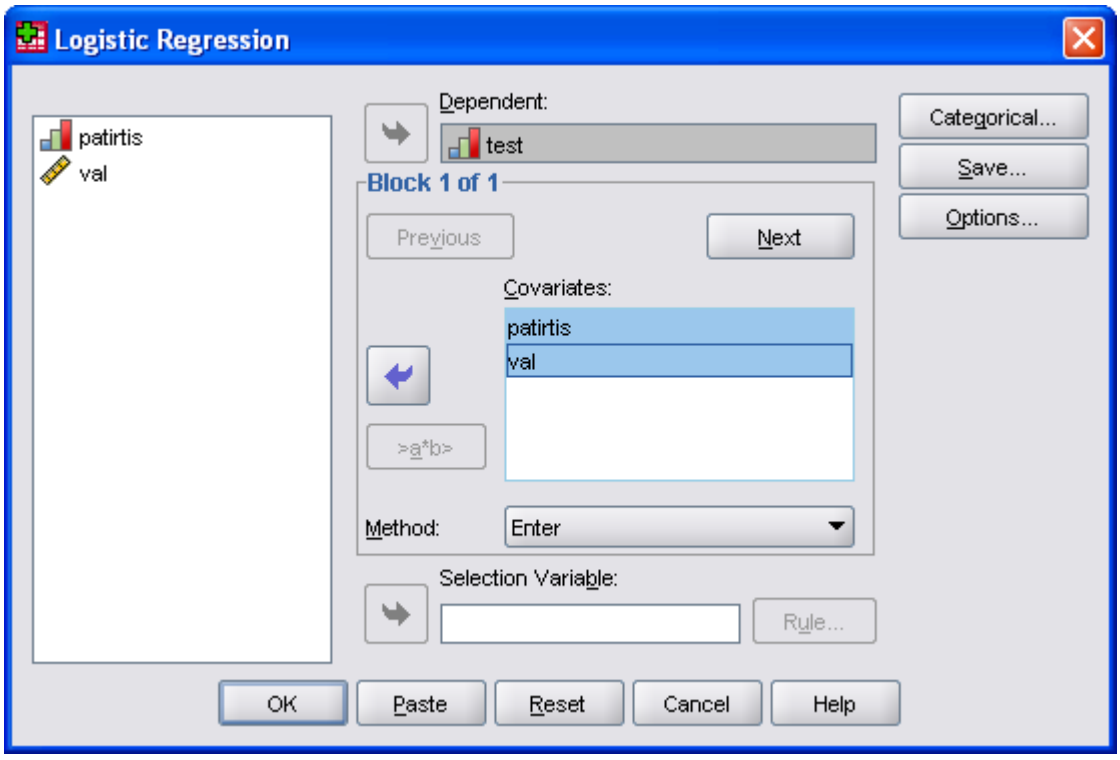
Logistinėje regresijoje taip pat naudojami tiesinės regresinės analizės determinacijos koeficiento analogai, nusakantys nepriklausomų kintamųjų ir priklausomo kintamojo priklausomybę. Tai Cox'o ir Snell'o pseudodeterminacijos koeficientas ir normuotas šio koeficiento variantas – Nagelkerke koeficientas (plačiau – Čekanavičius, Murauskas, 2002). Cox'o ir Snell'o pseudodeterminacijos koeficientas mažesnis už vienetą ir sunkiau interpretuojamas, tuo tarpu Nagelkerke koeficientas įgyja reikšmes diapazone nuo 0 iki 1.

5.2. BINARINĖS LOGISTINĖS REGRESIJOS MODELIO SUDARYMAS SU SPSS

Binarinės logistinės regresijos skaičiavimo SPSS paketu eiga yra ši:

- Atidarykite bylą su analizuojamais duomenimis.
- Nurodykite komandas ***Analyze*** → ***Regression...*** → ***Binary Logistic...***
- Dialogo langelyje ***Logistic Regression*** (5.1 pav.) priklausomą kintamąjį įkelkite į laukelį ***Dependent***, o nepriklausomus kintamuosius – į laukelį ***Covariates***. Nepriklausomus kintamuosius galima įkelti blokais – įkėlę vieną bloką, spragtelėkite ***Block*** mygtuką ***Next***.

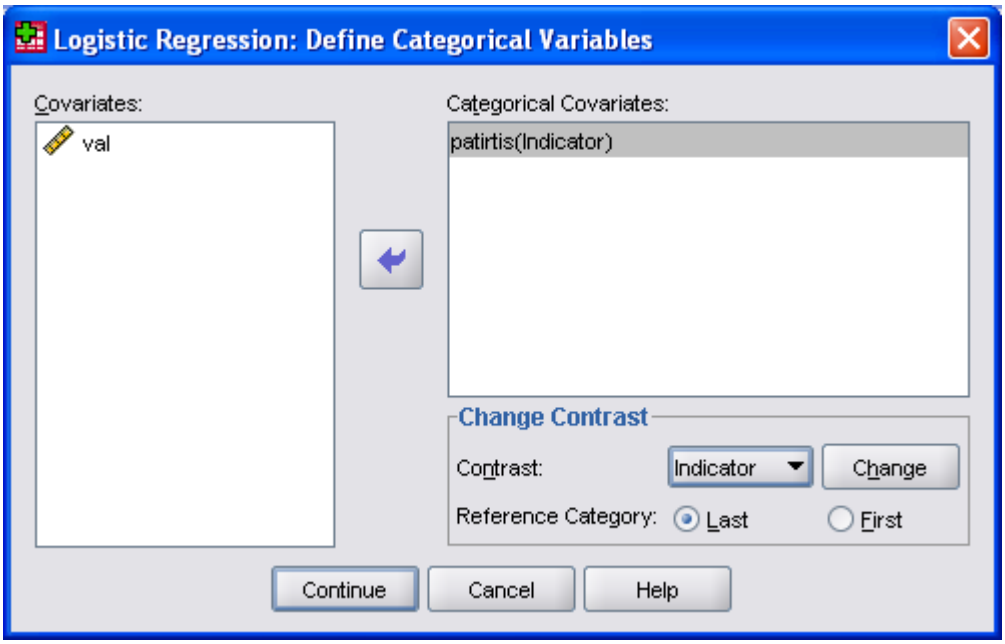
Išskleidžiamajame sąraše ***Method*** galima pasirinkti, kaip nepriklausomi kintamieji bus įtraukiami į skaičiavimą. Nustatytasis yra ***Enter*** variantas, kai visi atrinkti nepriklausomi kintamieji įtraukiami į skaičiavimą vienu metu. Alternatyva šiam metodui yra progresinė ir atgalinė atranka. Taikant tiesioginės atrankos (***Forward***) metodus, iš pradžių nustatomos konstantos, o toliau laipsniškai įtraukiami nepriklausomi kintamieji, turintys stiprų koreliacinį ryšį su priklausomu kintamuoju. Tada, remiantis atitinkamais kriterijais tikrinama, kurie kintamieji turi būti pašalinti iš regresijos lygties. Taikant atgalinės atrankos (***Backward***) metodus, skaičiavimas pradedamas su visais nepriklausomais kintamaisiais, pašalinant tolesnio darbo metu priklausomam kintamajam mažai įtakos turinčius nepriklausomus kintamuosius. Jeigu yra tik vienas nepriklausomas kintamasis, būtina pasirinkti ***Enter*** variantą, nors kai kurie autoriai (Garson, 2009) rekomenduoja šį metodą kaip pagrindinį ir daugelio kintamųjų atveju. Iš progresinės bei atgalinės atrankos metodų rekomenduojama taikyti ***Forward LR*** ir ***Backward LR*** metodus.



5.1 pav. Dialogo langelis ***Logistic Regression***

- Spragtelėkite mygtuką ***Categorical...***, jeigu tarp nepriklausomų kintamųjų yra kategorinių ir naujame dialogo langelyje ***Logistic Regression: Define Categorical Variables*** (5.2 pav.) įkelkite kategorinius kintamuosius į sąrašą ***Categorical***

Covariates. Išskleidžiamajame sąraše ***Contrast*** palikite nustatytąjį variantą ***Indicator*** arba pasirinkite taip pat dažnai naudojamą ***Deviation***. Kategorinis kintamasis bus suskaldytas į binarinius pseudokintamuosius, kurių skaičius bus lygus kategorijų skaičiui minus 1. Pasirinkite, kuriai kintamojo kategorijai suteikti nulinę pseudokintamojo reikšmę: palikite nustatytąjį ***Reference Category*** variantą ***Last*** arba pažymėkite ***First*** ir spragtelėkite mygtuką ***Change***. Spragtelėkite dialogo langelio ***Logistic Regression: Define Categorical Variables*** mygtuką ***Continue*** ir grįžkite į pagrindinį dialogo langelį.



5.2 pav. Dialogo langelis ***Logistic Regression: Define Categorical Variables***

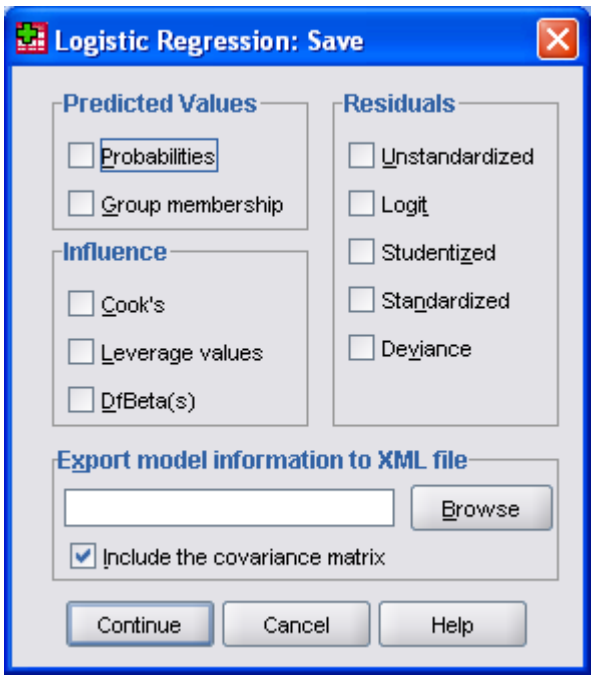
- Spragtelėję mygtuką ***Save...*** dialogo langelyje ***Logistic Regression: Save*** (5.3 pav.) galite nurodyti išsaugoti duomenų rinkmenoje papildomus kintamuosius: ***Predicted Values*** grupėje – prognozuojamas binarines reikšmes ***Group membership*** bei tų reikšmių tikimybes ***Probabilities***, ***Residuals*** grupėje – liekanas, ***Influence*** grupėje – įtakos įverčius.
- Spragtelėję mygtuką ***Options...*** dialogo langelyje ***Logistic Regression: Options*** (5.4 pav.) galite pasirinkti klasifikavimo grafiką ***Classification plots***, papildomą ***Hosmer-Lemeshow goodness-of-fit*** modelio suderinamumo kriterijų.
- Spragtelėkite dialogo langelio ***Logistic Regression*** mygtuką ***OK***.

Išvestyje pateikiamus binarinės logistinės regresijos rezultatus ir statistines išvadas paaiškinsime remdamiesi pavyzdžiu.

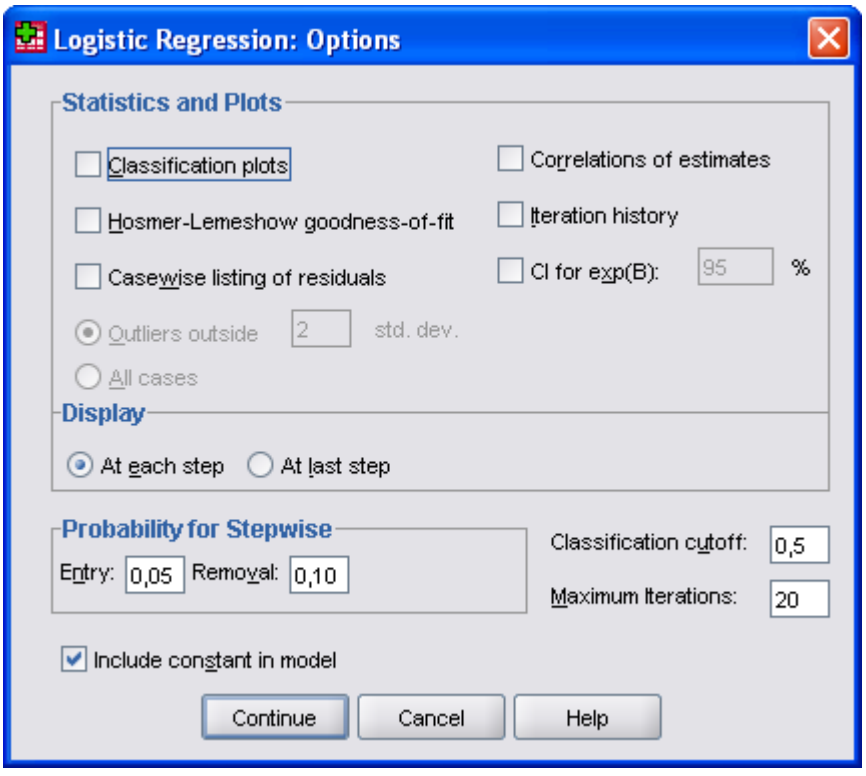
Pavyzdys. Studentai įvaldė naują kompiuterinę programą ir turėjo atlikti testą. Priklausomas kintamasis *test* lygus 1, jeigu testo rezultatai teigiami, ir lygus 0, jeigu testo rezultatai neigiami. Kiek valandų pratybų dirbo kiekvienas studentas, rodo intervalinis kintamasis *val*. Ar studentas turėjo darbo su panašiomis programomis patirtį, rodo kategorinis kintamasis *patirtis* (0 – ne, 1 – *taip*). Įvertinsime tikimybę, kad studento, kuris dirbo 30 val. ir turėjo darbo su panašiomis programomis patirtį, testo rezultatai bus teigiami. Pagrindiniai skaičiavimo rezultatai pateikti 5.5.1-5.5.2 pav. Skaičiavimams atlikti pasirinkta:

- Tiesioginės laipsniškos atrankos metodas ***Forward: Conditional***.

- ***Reference Category*** variantas ***First*** (5.2 pav.) kategoriniam nepriklausomam kintamajam *patirtis* perkoduoti į pseudokintamąjį. Esant nustatytajam ***Contrast*** variantui ***Indicator*** naujo binarinio pseudokintamojo pirmoji reikšmė bus 0. Tam, kad ši reikšmė atitiktų esamą kategorinio kintamojo *patirtis* kodavimą (0 – ne, 1 – *taip*) ir buvo pasirinktas ***First*** variantas, t. y. kad kintamojo 0 atitiktų pseudokintamojo 0.



5.3 pav. Dialogo langelis ***Logistic Regression: Save***



5.4 pav. Dialogo langelis ***Logistic Regression: Options***

Lentelėje *Categorical Variables Codings* (5.5.1 pav.) pateikiamos pseudokintamojo reikšmės. Kadangi buvo pasirinktas tiesioginės laipsniškos atrankos metodas ***Forward: Conditional***, rezultatai pateikiami pagal nepriklausomų kintamųjų įtraukimo į regresijos modelį etapus. Mus domina finalinis (šiuo atveju, antrasis etapas – *Step 2*). Lentelėje *Omnibus Tests of Model Coefficients* (5.5.1 pav.) pateikta χ^2 modelio suderinamumo kriterijaus reikšmė. Kaip paminėta aukščiau, hipotezė H_0 atmetama (taigi bent vienas $b_j \neq 0$), jeigu *p*-reikšmė $p < \alpha$. Hipotezė H_0 neatmetama (taigi visi $b_j = 0$), jeigu $p \geq \alpha$, čia α – reikšmingumo lygmuo. Šio pavyzdžio *p*-reikšmė (*Sig.*) $< 0{,}05$, todėl nulinę hipotezę atmetame. Be χ^2 modelio suderinamumo kriterijaus reikšmės (*Model*), kuri apskaičiuojama pagal (5.5) formulę lentelėje dar pateikiamos χ^2 kriterijaus reikšmės blokams (*Block*) ir finaliniam žingsniui (*Step*). χ^2 kriterijaus reikšmė blokams apskaičiuojama pagal (5.5) formulę, pakeitus joje $L(\hat{a}, 0)$ ir $L(\hat{a}, \hat{b})$ nepriklausomų kintamųjų gretimų blokų didžiausio tikėtinumo funkcijos maksimumais. Kai nepriklausomi kintamieji sudaro vieną bloką (taip yra pateikiamame pavyzdyje), χ^2 kriterijaus reikšmė blokams sutampa su χ^2 modelio suderinamumo kriterijaus reikšme. Pagal χ^2 kriterijaus reikšmę finaliniam žingsniui tikrinama nulinė hipotezė, kad finaliniame etape įtrauktų į modelį nepriklausomų kintamųjų koeficientai yra lygūs 0.

Lentelėje *Model Summary* (5.5.1 pav.) pateikiamos tiesinės regresijos determinacijos koeficiento analogų – Cox'o ir Snell'o pseudodeterminacijos koeficiento (*Cox & Snell R Square*) ir normuoto šio koeficiento varianto – Nagelkerke koeficiento (*Nagelkerke R Square*) reikšmės. Kuo šių koeficientų reikšmės didesnės, tuo geriau binarinė logistinė regresija suderinta su duomenimis. Tačiau reikia pažymėti, kad tos reikšmės paprastai būna gerokai mažesnės už tiesinės regresijos determinacijos koeficiento reikšmes. Šioje lentelėje taip pat pateikiama dviguba didžiausio tikėtinumo funkcijos maksimumo logaritmo reikšmė *–2 Log likelihood*.

Lentelėje *Hosmer and Lemeshow Test* (5.5.1 pav.) pateikiami Hosmer'io-Lemeshow'o testo rezultatai. Šiame pavyzdyje kriterijaus *p*-reikšmė $0{,}914 > 0{,}05$, t. y. kriterijus patvirtina, kad binarinės logistinės regresijos modelis gerai suderintas su duomenimis.

Lentelėje *Classification Table* (5.5.2 pav.) pateikiamos žinomos (*Observed*) priklausomo kintamojo reikšmės ir regresijos modelio pagrindu prognozuotinos (*Predicted*) reikšmės. Iš 13 kintamojo *test* reikšmių 0 (*ne*) testu buvo patvirtintos 10, o kitos 3 priskirtos prie 1 (*taip*). Iš 17 kintamojo *test* reikšmių 1 (*taip*) buvo patvirtintos 14, o kitos 3 priskirtos prie 0 (*ne*). Bendras teisingo atpažinimo procentas yra 80%. Nėra griežtų taisyklių, nusakančių, koks teisingo klasifikavimo procentas laikytinas patenkinamu. Paprastai reikalaujama, kad kiekvienos kategorijos jis būtų ne mažesnis kaip 50%, o logistinė regresija apskritai taikoma tik tuo atveju, kai $y_i = 0$ sudaro ne mažiau kaip 20% ir ne daugiau kaip 80% visų stebėjimų.

Lentelėje *Variables in the Equation* (5.5.2 pav.) pateikiamos koeficientų $\hat{a}, \hat{b}_1, \hat{b}_2$ įverčių reikšmės (*B* stulpelyje) ir jų nelygybės nuliui reikšmingumas. Wald'o kriterijus yra Stjudento kriterijaus tiesinėje regresijoje analogas, t. y. jis atsako į klausimą, ar konkretus koeficientas $b_j \neq 0$. Hipotezė H_0 atmetama (taigi $b_j \neq 0$), jeigu *p*-reikšmė $p < \alpha$. Hipotezė H_0 neatmetama, jeigu $p \geq \alpha$, čia α — reikšmingumo lygmuo. Visi šio pavyzdžio regresijos lygties koeficientai reikšmingai nelygūs nuliui, nes *p*-reikšmė (*Sig.*) yra mažesnė už 0,05. Šios lentelės *Exp(B)* stulpelyje pateikiama galimybių santykio reikšmė. Galimybių santykis rodo, kaip kinta priklausomo kintamojo galimybė įgyti reikšmę 1. Šiuo pavyzdžiu galime teigti, kad studentas, turintis darbo su analogiškomis programomis patirtį padidino savo galimybę

išlaikyti testą 8,444 karto. Analogiškai viena papildoma pratybų valanda padidina galimybę išlaikyti testą 1,305 karto.

Pasirinkus tiesioginės laipsniškos atrankos metodą ***Forward: Conditional*** papildomai pateikiama lentelė *Model if Term Removed* (5.5.2 pav.). Joje nurodoma χ^2 suderinamumo kriterijaus reikšmė pagal kiekvieną nepriklausomą kintamąjį, šios reikšmės pokytis pašalinus kintamąjį iš regresijos modelio ir šio pokyčio statistinis reikšmingumas.

Categorical Variables Codings

		Frequency	Parameter coding (1)
patirtis	ne	15	,000
	taip	15	1,000

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	11,465	1	,001
	Block	11,465	1	,001
	Model	11,465	1	,001
Step 2	Step	4,669	1	,031
	Block	16,134	2	,000
	Model	16,134	2	,000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	29,589[a]	,318	,426
2	24,920[b]	,416	,558

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than ,001.

b. Estimation terminated at iteration number 6 because parameter estimates changed by less than ,001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	8,904	7	,260
2	3,298	8	,914

5.5.1 pav. Pagrindiniai binarinės logistinės regresijos skaičiavimo rezultatai

Classification Table[a]

Observed			Predicted		
			test		Percentage Correct
			neigiamas	teigiamas	
Step 1	test	neigiamas	8	5	61,5
		teigiamas	4	13	76,5
	Overall Percentage				70,0
Step 2	test	neigiamas	10	3	76,9
		teigiamas	3	14	82,4
	Overall Percentage				80,0

a. The cut value is ,500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	val	,241	,089	7,294	1	,007	1,273
	Constant	-7,231	2,749	6,920	1	,009	,001
Step 2[b]	val	,266	,113	5,513	1	,019	1,305
	patirtis(1)	2,133	1,084	3,872	1	,049	8,444
	Constant	-8,866	3,613	6,020	1	,014	,000

a. Variable(s) entered on step 1: val.

b. Variable(s) entered on step 2: patirtis.

Model if Term Removed[a]

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	val	-20,643	11,697	1	,001
Step 2	patirtis	-15,025	5,131	1	,024
	val	-17,773	10,626	1	,001

a. Based on conditional parameter estimates

5.5.2 pav. Pagrindiniai binarinės logistinės regresijos skaičiavimo rezultatai

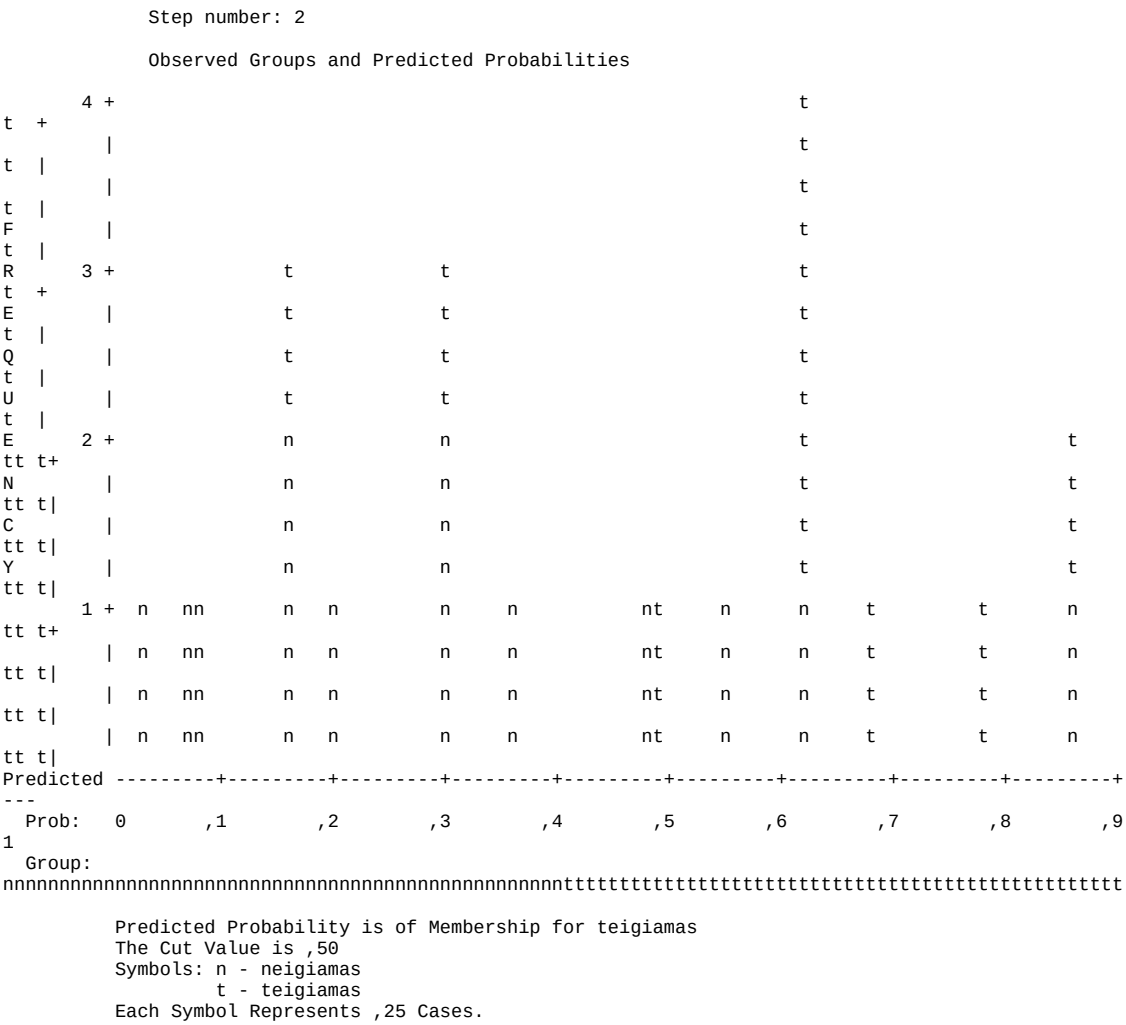
Turėdami koeficientų reikšmes, galime apskaičiuoti dydžio $\hat{z}$ reikšmę ir norimo įvykio tikimybę p pagal 5.4 formulę. Įrašę koeficientų reikšmes į $\hat{z}$ išraišką, gauname

$$\hat{z} = -8{,}866+2{,}133*1+0{,}266*30 = 1{,}247;$$

$$\hat{P}\left(testas\ teigiamas \middle| 30{,}1\right) = \exp\{1{,}247\}/\left(1+\exp\{1{,}247\}\right) = 0{,}78\,.$$

Gautas rezultatas visada reiškia įvykio tikimybę, kad priklausomas kintamasis iš dviejų galimų kodavimo reikšmių turės didesnę reikšmę.

Be klasifikacinės lentelės *Classification Table* (5.5.2 pav.) duomenų klasifikavimą galima įvertinti ir grafiškai. Tam dialogo lentelėje ***Logistic Regression: Options*** (5.4 pav.) reikia pažymėti laukelį ***Classification plots***. Klasifikacijos grafiko *x* ašyje atidėtas prognozuojamos tikimybės įvertis $\hat{P}(Y=1|\vec{x})$ – *Prob.* Visi stebėjimai, kurių tikimybė mažesnė už 0,5, prognozuojami kaip neigiami testo rezultatai (n), stebėjimai, kurių tikimybė didesnė už 0,5 – kaip teigiami testo rezultatai (t). Virš *x* ašies pažymėta, ar respondentas iš tikrųjų išlaikė testą, ar ne. Kuo daugiau simbolių "n" yra į kairę nuo 0,5 ir kuo daugiau simbolių "t" į dešinę nuo 0,5, tuo geriau prognozė sutampa su turimais stebėjimais. Kiek simbolių turi vaizduoti vieną stebėjimą programa nustato automatiškai – šiame pavyzdyje vieną stebėjimą atitinka keturi simboliai.



5.6 pav. Klasifikacijos grafikas

Binarinėje logistinėje regresijoje, kaip ir tiesinėje regresijoje, yra daug rodiklių, skirtų liekamųjų paklaidų analizei. Juos galima nurodyti dialogo langelyje ***Logistic Regression: Save*** (5.3 pav.). Trumpai paminėsime kai kuriuos iš jų.

- Liekamosios paklaidos (***Residuals: Unstandardized***). Tai skirtumas tarp stebimos priklausomo kintamojo tikimybės (kuri laikoma lygi 1, kai kintamasis įgyja reikšmę 1) ir prognozuojamos pagal logistinės regresijos modelį tikimybės.
- Standartizuotos liekamosios paklaidos (***Residuals: Standardized***). Jos apskaičiuojamos liekamąsias paklaidas padalijus iš standartinio nuokrypio. Esant pakankamai didelei imčiai, standarizuotos liekamosios paklaidos pasiskirsto pagal artimą normaliajam pasiskirstymo dėsnį. Jų imties vidurkis lygus nuliui, o standartinis nuokrypis – vienetui.
- Nuokrypis (***Residuals: Deviance***). Apskaičiuojamas pagal formulę (Norušis, 2005)

$$\hat{d}_i = -\sqrt{-\ln\left(\hat{P}_i\right)}. \qquad (5.6)$$

Minuso ženklas prieš šaknį dedamas, kai priklausomas kintamasis įgyja nulinę reikšmę.

- Stebėjimo įtakos koeficientas (***Influence: Leverage values***) savo esme panašus į mažiausiųjų kvadratų metodo stebėjimo įtakos koeficientą. Dažnai naudojamas nustatyti didelę įtaką prognozuojamoms reikšmėms turinčius stebėjimus. Koeficiento reikšmės yra diapazone nuo nulio iki vieneto.
- Kuko matas (***Influence: Cook's***). Apskaičiuojamas pagal formulę (Norušis, 2005)

$$D_i = \frac{Z_i^2 h_i}{1 - h_i}, \qquad (5.7)$$

čia Z_i – standartizuotoji liekana, h_i – stebėjimo įtakos koeficientas.

- ***Influence: DfBeta(s)***. DfBeta koeficientas apibrėžia binarinės logistinės regresijos koeficientų pokytį, kai stebėjimas pašalinamas.

Visos dialogo langelyje ***Logistic Regression: Save*** nurodytų parametrų reišmės bus išsaugotos duomenų rinkmenoje. Jų analizei galima pasitelkti grafikus.

6. SPSS SPRENDIMŲ MEDŽIAI (*SPSS Decision Trees*)

SPSS programų paketo sprendimų medžių (*Decision Trees*) modulis įgalina klasifikuoti turimus duomenis pagal grupes bei prognozuoti priklausomus kintamuosius pagal žinomus nepriklausomus kintamuosius. Tai svarbus tiriančiosios bei patvirtinančiosios diskriminantinės analizės įrankis. Sprendimų medžių moduliu Jūs galite:

- Identifikuoti tiriamuosius pagal jų galimą priklausomybę tam tikrai klasifikacinei grupei.
- Priskirti tiriamuosius tam tikrai kategorijai, pvz., mažos, vidutinės bei didelės rizikos grupėms.
- Pagal sudarytą modelį prognozuoti būsimus įvykius, pvz., kredito negrąžinimą.
- Suspausti turimus duomenis, iš didelės nepriklausomų kintamųjų grupės palikdami tik turinčius statistiškai reikšmingą įtaką priklausomo kintamojo prognozei.
- Identifikuoti sąveiką tarp atskirų tiriamųjų grupių.

Sprendimų medžių modelyje yra naudojami klasikiniai statistiniai kriterijai (χ^2 ir kt.), tačiau vaizdingas analizės rezultatų pateikimas klasifikacinių ar sprendimų medžių pavidalu leidžia lengvai suvokti hierarchinę kintamųjų priklausomybę bei nustatyti specifines kategorijas.

Tiriami duomenys gali priklausyti vardinei (*Nominal*), rangų (*Ordinal*) ir intervalų (*Scale*) skalėms. Kategoriniams (vardinės ir rangų skalės) priklausomiems kintamiesiems turi būti nurodytos visų analizuojamų kategorijų žymenos.

6.1. SPRENDIMŲ MEDŽIO SUDARYMAS

Sprendimų medžių modulio taikymą pailiustruosime šiuo pavyzdžiu, pateikiamu su SPSS 17.0 versijos programų paketu – bankas, remdamasis sukauptais duomenimis apie kredito gavėjus ir jų įsipareigojimą vykdymą, nori nustatyti kredito gavėjų, turinčių problemų su kredito grąžinimu, rizikos grupes. Priklausomas kintamasis yra kreditingumo įvertinimas – *Credit rating* (0 – "blogas", 1 – "geras"), nepriklausomi kintamieji – amžius – *Age*, pajamos – *Income level* (1 – "mažos",2 – "vidutinės",3 – "didelės"), kredito kortelių skaičius – *Number of credit cards* (1 – "mažiau 5",2 – 1 – "5 ir daugiau"), išsilavinimas – *Education* (1 – "aukštasis", 2 – "vidurinis"), nuomojami automobiliai – *Car loans* (1 – "nenomuoja arba 1",2 – "daugiau kaip 2").

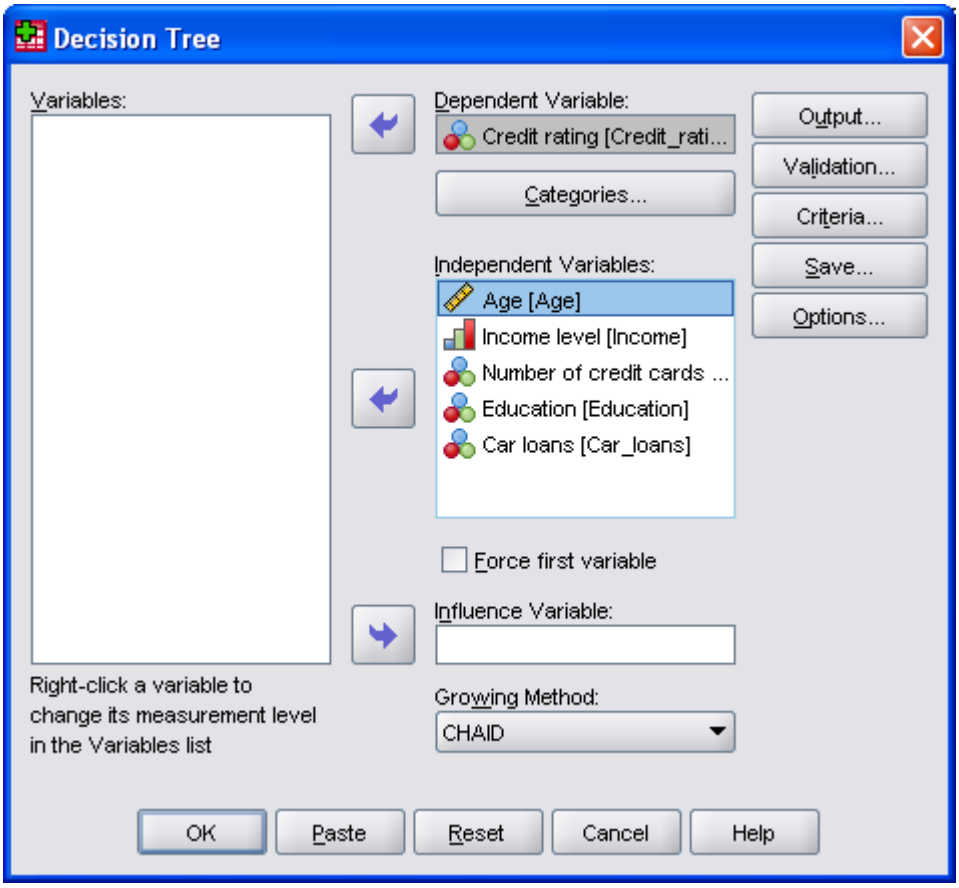
- Nurodykite komandas ***Analyze → Classify →Tree...***
- Įkelkite į dialogo langelio ***Decision Tree*** (6.1 pav.) laukelį ***Dependent Variable*** priklausomą kintamąjį, o nepriklausomus kintamuosius – į laukelį ***Independent Variables***.
- Pasirinkite sprendimų medžio sudarymo (auginimo) metodą ***Growing Method***:
  - CHAID (*Chi-squared Automatic Interaction Detection*). Tai nustatytasis metodas. Pagal šį metodą kiekviename žingsnyje nustatomas stipriausią sąveiką su priklausomu kintamuoju turintis nepriklausomas kintamasis (*predictor*). Savo įtaka priklausomam kintamajam mažai besiskiriančios nepriklausomų kintamųjų kategorijos apjungiamos.
  - Exhaustive CHAID yra CHAID metodo modifikacija.
  - CRT *(Classification and Regression Trees)*.
  - QUEST *(Quick, Unbiased, Efficient Statistical Tree)*. Įgalina išvengti kitiems metodams būdingų paklaidų, kai pirmenybė teikiama daug kategorijų turintiems nepriklausomiem kintamiesiems. Taikomas tik vardinės skalės priklausomiems kintamiesiems.

Sprendimų medžio sudarymo metodų palyginimas pateiktas 6.1 lentelėje

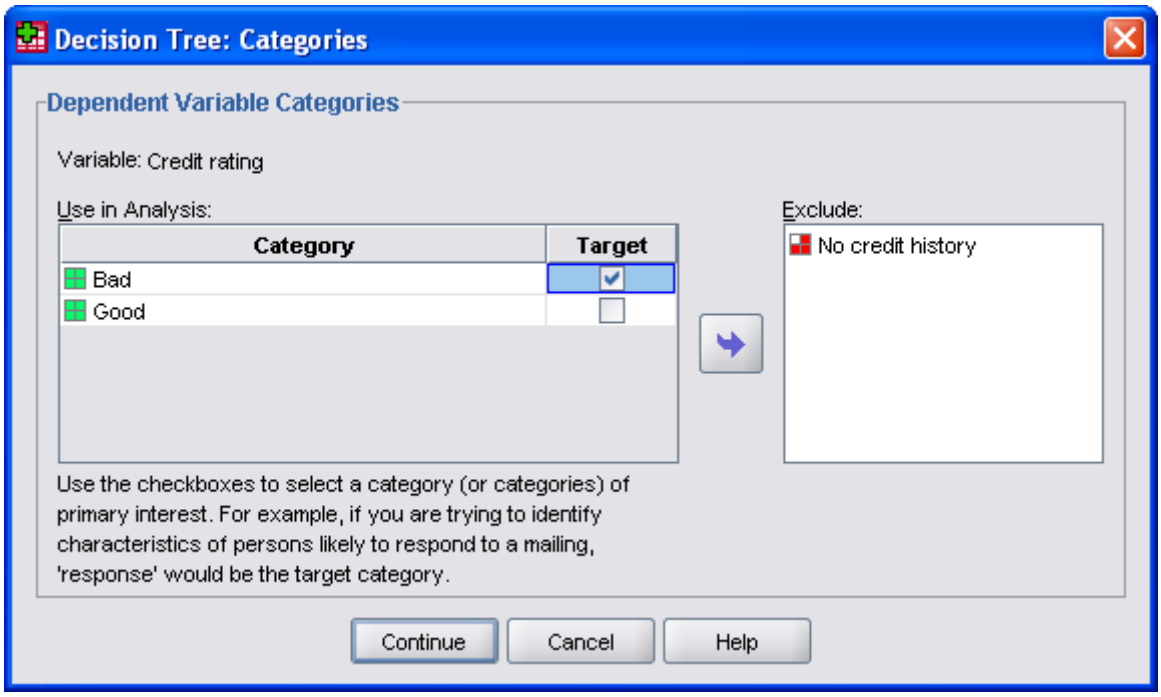
6.1 lentelė. Sprendimų medžių sudarymo metodų palyginimas

	CHAID (Exhaustive CHAID)	CRT	QUEST
Chi-kvadratu pagrindu	X		
Pakaitinių (surogatinių) nepriklausomų kintamųjų naudojimas		X	X
Redukuotas ("apgenėtas") sprendimų medis		X	X
Daugialypis mazgo šakojimasis	X		
Binarinis mazgo šakojimasis		X	X
Įtakos kintamieji	X	X	
Pirminė priklausomų kintamųjų kategorijų tikimybė		X	X
Klaidingos klasifikacijos kaina	X	X	X
Spartus skaičiavimas	X		X

- Dialogo langelio ***Decision Tree*** laukelis ***Force first variable*** pažymimas tada, kai norima, kad atsišakojimas prasidėtų pagal pirmą nepriklausomų kintamųjų sąraše esantį kintamąjį.
- Į laukelį ***Influence Variable*** gali būti įkeltas įtakos kintamasis, kurio įtaka visiems stebėjimams priklauso nuo kintamojo reikšmių – didesnės įtakos kintamojo reikšmės labiau įtakoja stebėjimus, mažesnės įtakos kintamojo reikšmės mažiau įtakoja stebėjimus. Įtakos kintamojo reikšmės turi būti teigiamos.
- Spragtelėkite dialogo langelio ***Decision Tree*** mygtuką ***Categories...*** ir atsidariusiame dialogo langelyje ***Decision Tree: Categories*** (6.2 pav.) nurodykite vardinės arba rangų skalės priklausomo kintamojo dominančią kategoriją – šiuo atveju tokia kategorija bus *bad, t. y.* blogas kredito grąžinimas. Galima apsiriboti mažesnio priklausomo kintamojo kategorijų skaičiaus analize, perkeliant nepageidaujamas kategorijas į laukelį ***Exclude***. Pasirinkus kelias kategorijas, rezultatai bus pateikti kiekvienai nurodytai kategorijai. Vartotojo nustatytos praleistos vardinės skalės priklausomo kintamojo reikšmės į laukelį ***Exclude*** patalpinamos automatiškai.

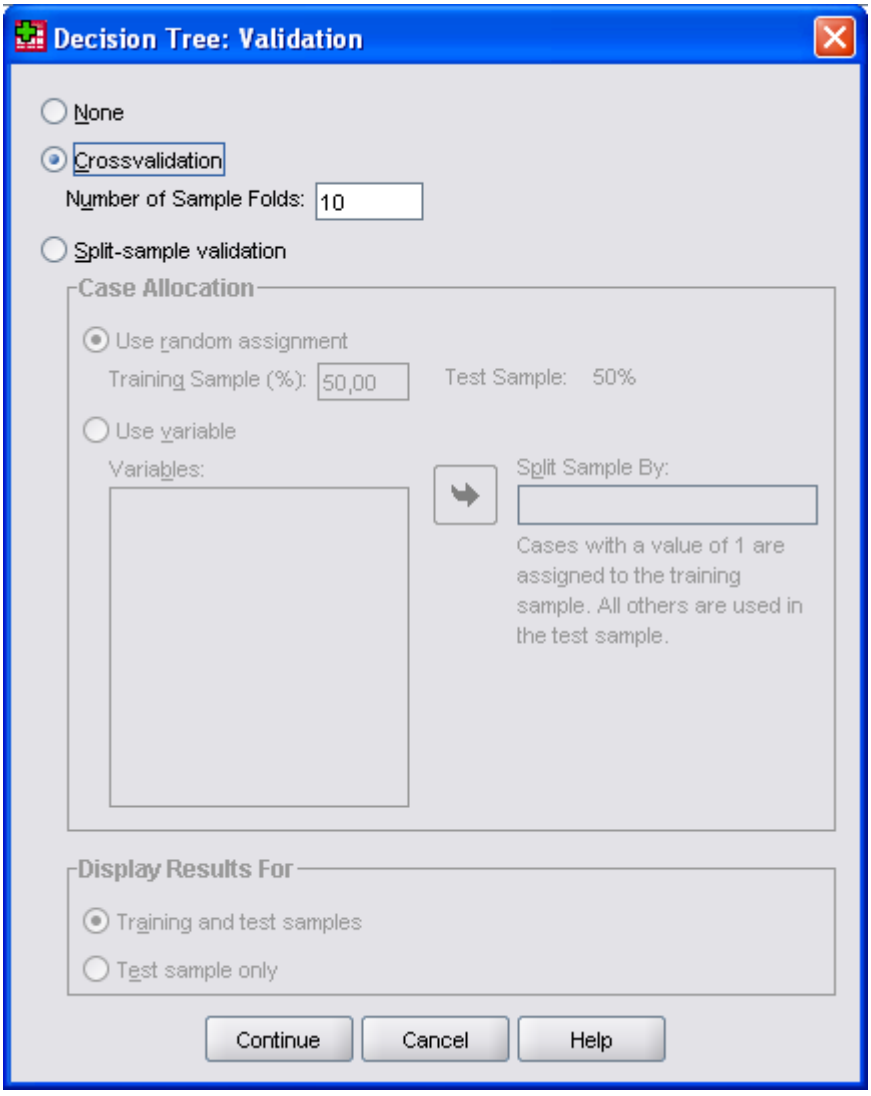


6.1 pav. Dialogo langelis ***Decision Tree***



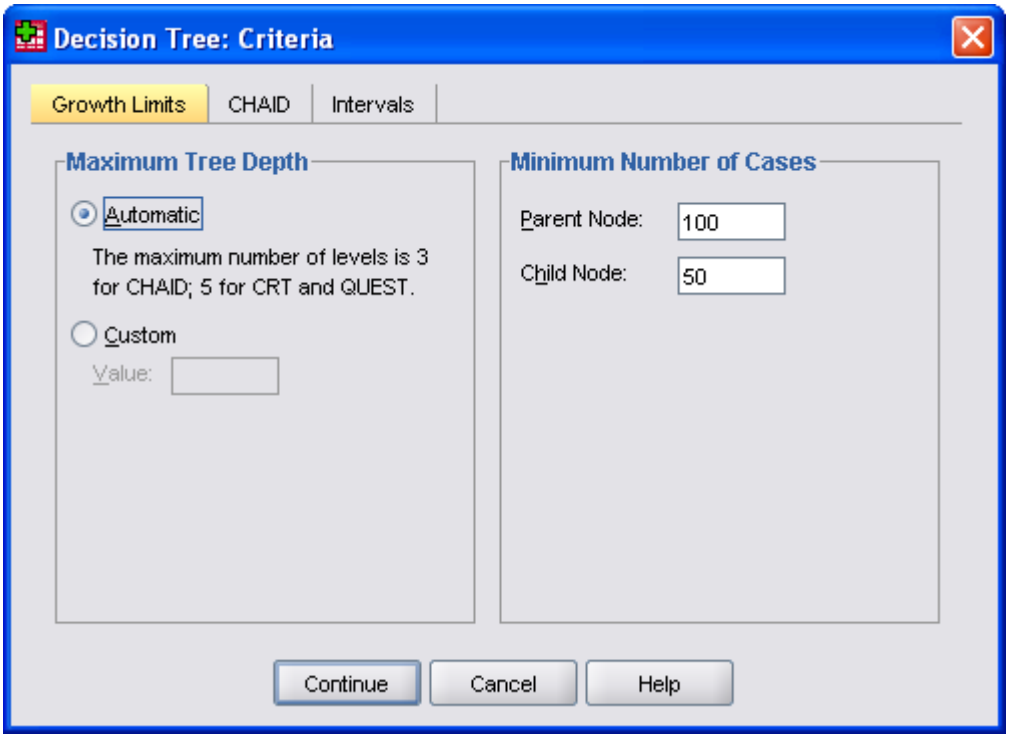
6.2 pav. Dialogo langelis ***Decision Tree: Categories***

- Spragtelėkite dialogo langelio ***Decision Tree*** mygtuką ***Validation...*** ir atsidariusiame dialogo langelyje ***Decision Tree: Validation*** (6.3 pav.) pasirinkite vieną iš dviejų gautų rezultatų įverčio metodų – ***Crossvalidation*** ar ***Split-sample validation***. Pagal ***Crossvalidation*** metodą imtis suskaldomą į mažesnio dydžio dalines imtis ir sudaromi sprendimų medžio modeliai: pirmas modelis – be pirmos dalinės imties duomenų, antras modelis – be antros dalinės imties duomenų ir t. t. Pagal sudarytus modelius apskaičiuojama klaidingos klasifikacijos rizika kiekvienai pašalintai iš modelio sudarymo dalinei imčiai ir pateikiamas galutinis suvidurkintas rezultatas. Nustatytasis dalinių imčių skaičius yra 10, didžiausias galimas – 25.
  Pagal ***Split-sample validation*** metodą modelis yra sudaromas pagal apmokomąją imtį ir tikrinamas pagal pagrindinę imtį. Apmokomąją imtį galima sudaryti nurodžius procentinę bendros imties dalį arba įvedus kintamąjį, kuris padalintų bendrą imtį į apmokomąją ir testuojamą. Duomenys, atitinkantys šio kintamojo reikšmę 1, priskiriami apmokomajai imčiai. Nedidelėms imtims ***Split-sample validation*** metodas gali būti taikomas įsitikinus, kad pakanka duomenų apmokomajai imčiai sudaryti.



6.3 pav. Dialogo langelis ***Decision Tree: Validation***

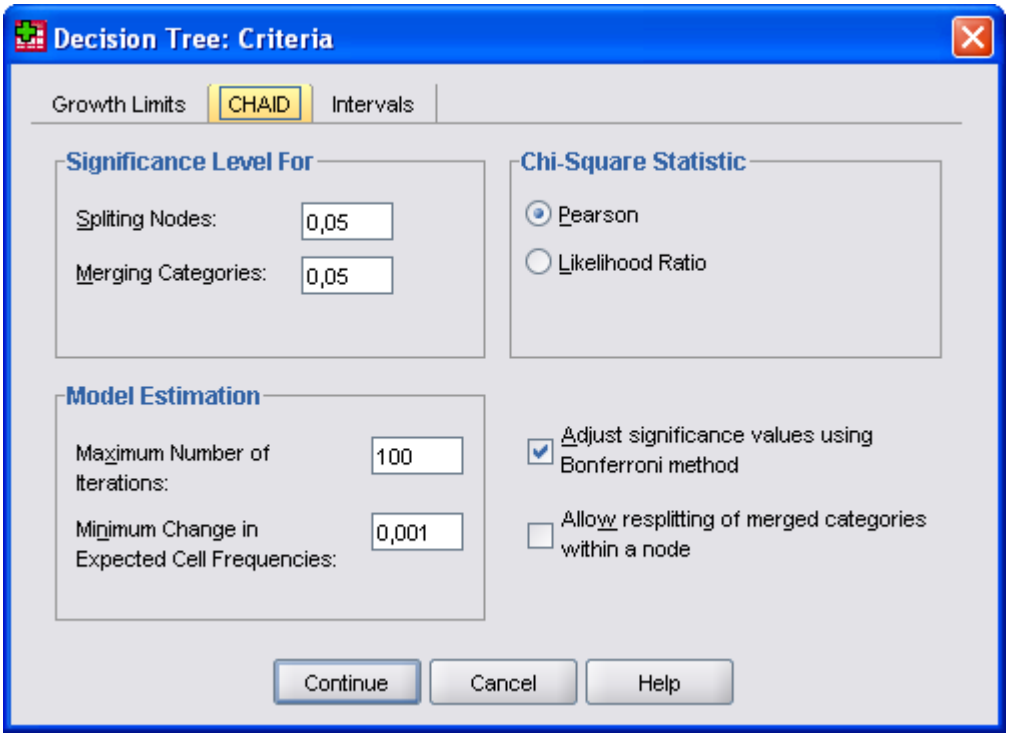
- Spragtelėkite dialogo langelio ***Decision Tree*** mygtuką ***Criteria...*** sprendimų medžio sudarymo kriterijams nustatyti. Šie kriterijai priklauso nuo pasirinkto metodo bei nuo priklausomo kintamojo skalės.
- Pasirinkus CHAID (Exhaustive CHAID) metodą dialogo langelis ***Decision Tree: Criteria*** atrodys kaip parodyta 6.4 pav. Šio langelio kortelės ***Growth Limits*** laukelyje ***Maximum Tree Depth*** nurodomas sprendimų medžio lygių skaičius žemiau pagrindinio mazgo (*root node*). Nustatytasis automatinis (***Automatic***) režimas numato daugiausia 3 lygius CHAID metodui bei daugiausia 5 lygius CRT ir QUEST metodams. Laukelyje ***Minimum Number of Cases*** nurodomas mažiausias reikšmių, sudarančių pagrindinius (motininius) mazgus (***Parent Node***) ir atžalų mazgus (***Child Node***) skaičius. Šių skaičių didinimas veda prie sprendimų medžių su mažesniu mazgų skaičiumi sudarymo ir atvirkščiai. Nustatytosios ***Parent Node*** ir ***Child Node*** reikšmės – atitinkamai 100 ir 50 – gali būti per didelės duomenų rinkmenoms su nedideliu stebėjimų skaičiumi, todėl gali tekti jas mažinti. Nagrinėjamame pavyzdyje pasirenkam mažiausią ***Parent Node*** sudarančių reikšmių skaičių lygų 400 ir mažiausią ***Child Node*** sudarančių reikšmių skaičių lygų 200.



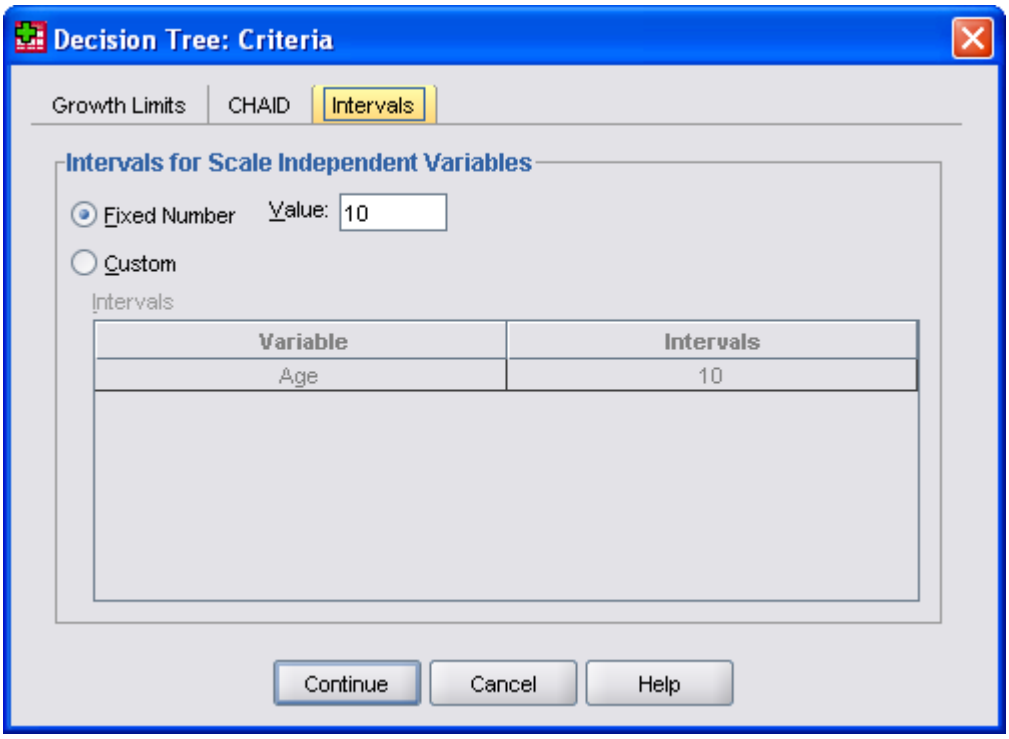
6.4 pav. Dialogo langelis ***Decision Tree: Criteria***

Dialogo langelio ***Decision Tree: Criteria*** kortelėje ***CHAID*** (6.5 pav.) Jūs galite pasirinkti CHAID (Exhaustive CHAID) metodo parametrus (pasikliautinąjį intervalą, χ^2 statistiką ir kt.). Rekomenduojama palikti nustatytuosius parametrus, išskyrus mažas imtis, kada vardinės skalės priklausomiems kintamiesiems tikslinga pasirinkti ***Likelihood Ratio*** statistiką.

Pagal CHAID metodą nepriklausomi intervalų skalės kintamieji suskirstomi į diskretines grupes. Nustatytasis tokių grupių skaičius yra 10. Pažymėję dialogo langelio ***Decision Tree: Criteria*** kortelėje ***Intervals*** (6.6 pav.) variantą ***Custom***, Jūs galite pasirinkti norimą grupių skaičių kiekvienam nepriklausomam intervalų skalės kintamajam atsižvelgdami į to kintamojo reikšmių sklaidą.



6.5 pav. Dialogo langelio ***Decision Tree: Criteria*** kortelė ***CHAID***

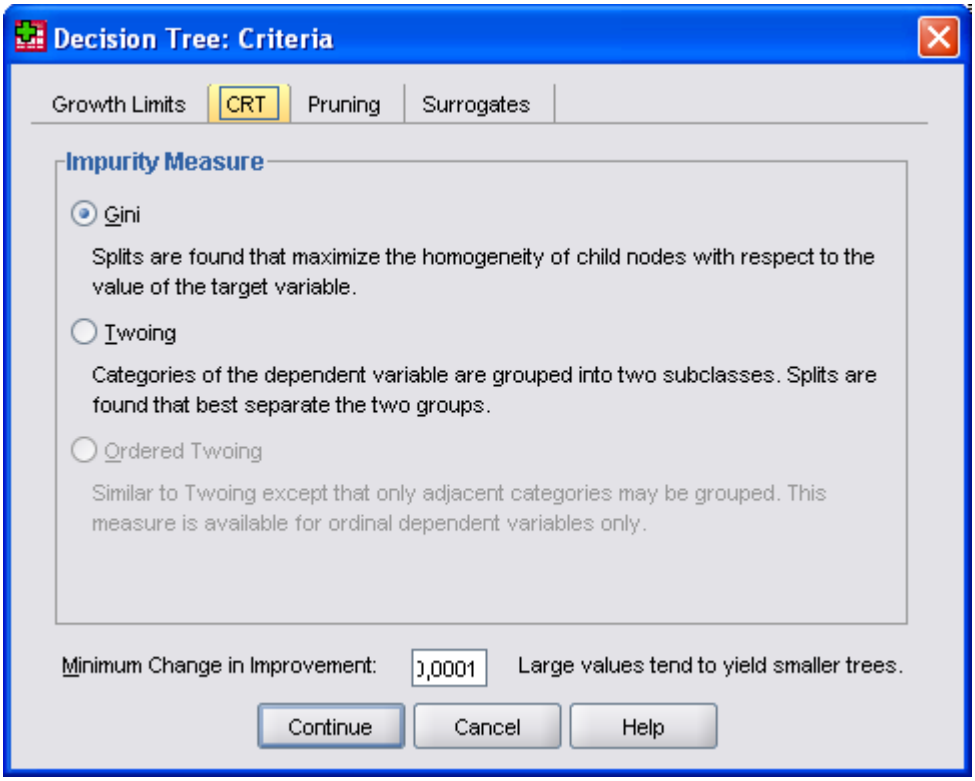


6.6 pav. Dialogo langelio ***Decision Tree: Criteria*** kortelė ***Intervals***

- Pasirinkus CRT metodą dialogo langelio ***Decision Tree: Criteria*** vaizdas pasikeis – jame atsiras naujos kortelės ***CRT, Pruning ir Surrogates*** (6.7 pav.). CRT metodu sudaromas sprendimų medis, siekiant didžiausio mazgų (*nodes*) homogeniškumo.

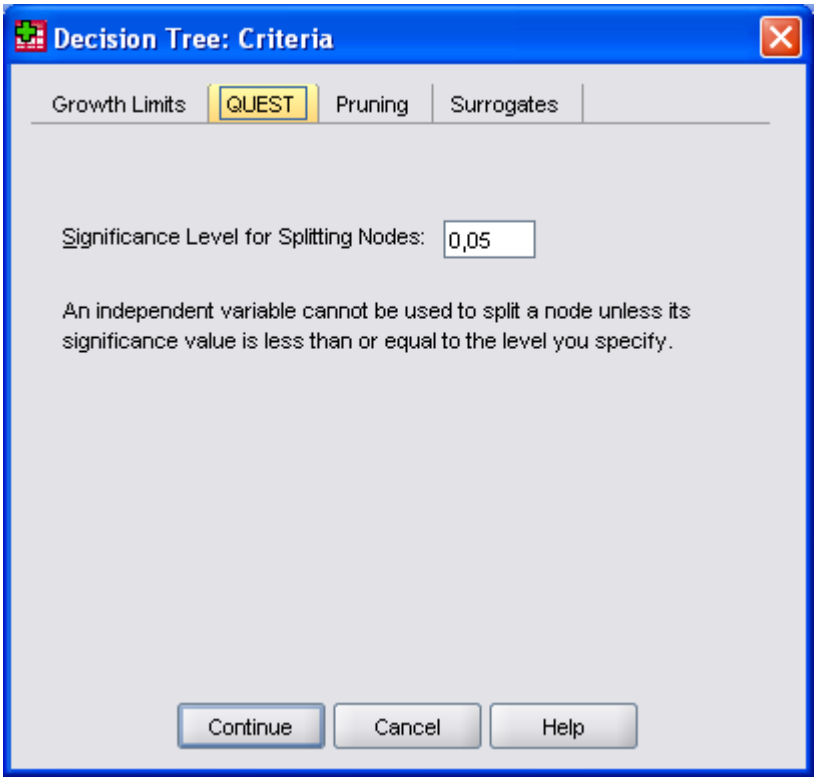
Kategoriniams (vardinės ir rangų skalės) priklausomiems kintamiesiems Jūs galite pasirinkti šiuos homogeniškumo laipsnio įverčio matus:

- ***Gini*** (nustatytasis variantas). Duomenys skeliami taip, kad mazgus sudarytų galimai didesnio homogeniškumo (atžvilgiu priklausomo kintamojo) duomenys. Gini matas paremtas tikimybe priklausyti kiekvienai priklausomo kintamojo kategorijai. Mažiausia reikšmė (nulis) pasiekiama, kai visi mazgo duomenys patenka į vieną kategoriją.
- ***Twoing***. Priklausomo kintamojo kategorijos grupuojamos į dvi grupes. Duomenys skeliami taip, kad abi grupės būtų atskirtos geriausiai.
- ***Ordered Twoing***. Panašus į ***Twoing***, išskyrus tai, kad grupuojamos tik gretimos priklausomo kintamojo kategorijos. Taikomas tik rangų skalės priklausomiems kintamiesiems.

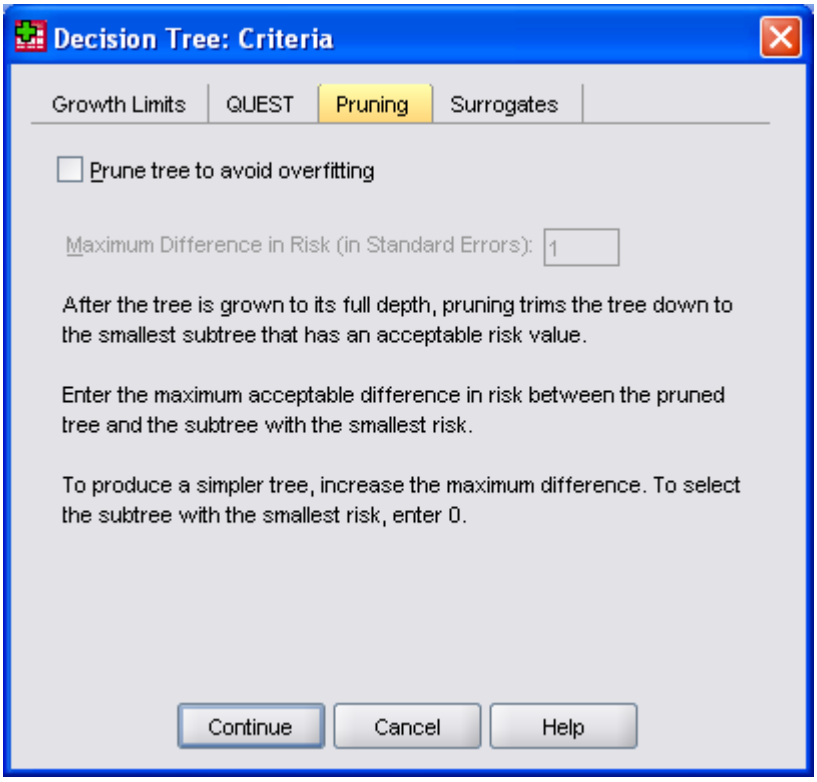


6.7 pav. Dialogo langelio ***Decision Tree: Criteria*** kortelė ***CRT***

- Pasirinkę QUEST metodą dialogo langelyje ***Decision Tree: Criteria*** turėsite ***QUEST*** kortelę (6.8 pav.), kurioje galima pakeisti nustatytąjį reikšmingumo lygmenį (0,05). Mazgas gali būti skeliamas pagal nepriklausomą kintamąjį, jeigu reikšmingumo lygmuo lygus ar mažesnis už nustatytą. Taigi, reikšmingumo lygmens mažinimas veda prie mažesnio nepriklausomų kintamųjų skaičiaus galutiniame modelyje.
- Pasirinkus CRT ar QUEST metodą numatyta galimybė išvengti t. v. modelio permokymo (*overfitting*) taikant sprendimų medžio "apgenėjimą" (*pruning*) – pirminis sprendimų medis apkarpomas iki galimai mažiausio remiantis didžiausia rizika. Tam reikia dialogo langelio ***Decision Tree: Criteria*** kortelėje ***Pruning*** (6.9 pav.) pažymėti laukelį ***Prune tree to avoid overfitting***. Nustatyta rizikos reikšmė yra 1, norint gauti apkarpytą medį su mažiausia rizika, laukelyje ***Maximum Difference in Risk (in Standart Errors)*** reikia nurodyti 0.

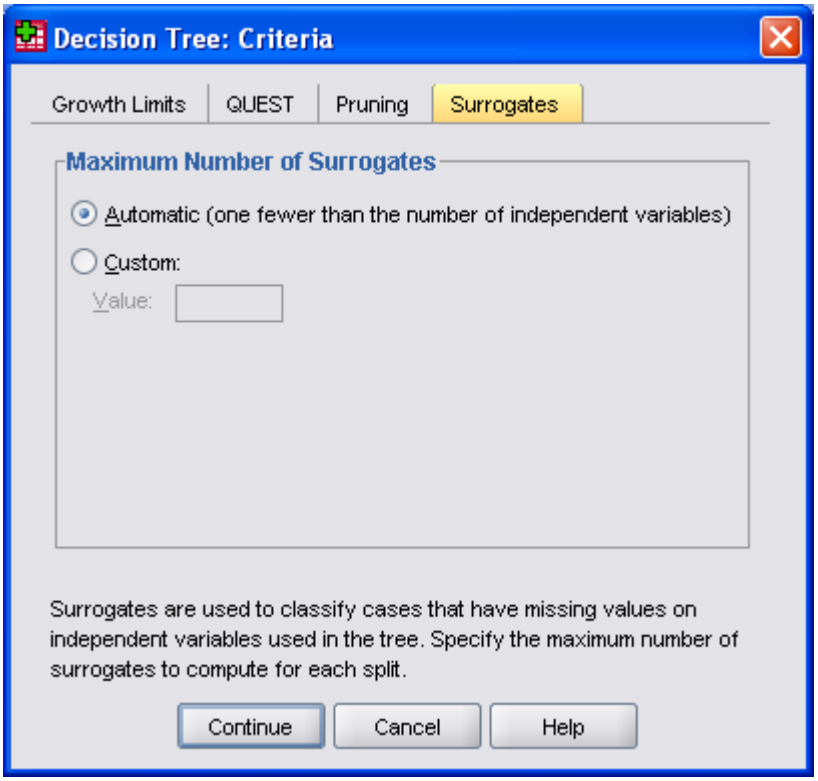


6.8 pav. Dialogo langelio ***Decision Tree: Criteria*** kortelė ***QUEST***



6.9 pav. Dialogo langelio ***Decision Tree: Criteria*** kortelė ***Pruning***

- Jeigu nepriklausomas kintamasis turi praleistų duomenų, tai CRT ir QUEST metodai klasifikacijai naudoja kitą alternatyvų nepriklausomą kintamąjį, t. v. surogatinį kintamąjį, turintį stiprų ryšį su originaliu kintamuoju. Nustatytasis surogatinių kintamųjų skaičius (žiūr. dialogo langelio ***Decision Tree: Criteria*** kortelėje ***Surrogates*** (6.10 pav.) yra vienetu mažesnis už bendrą nepriklausomų kintamųjų skaičių. Jeigu Jūs nenorite naudoti savo tyrime surogatinių kintamųjų, pasirinkite variantą ***Custom*** ir laukelyje ***Value*** įrašykite 0.

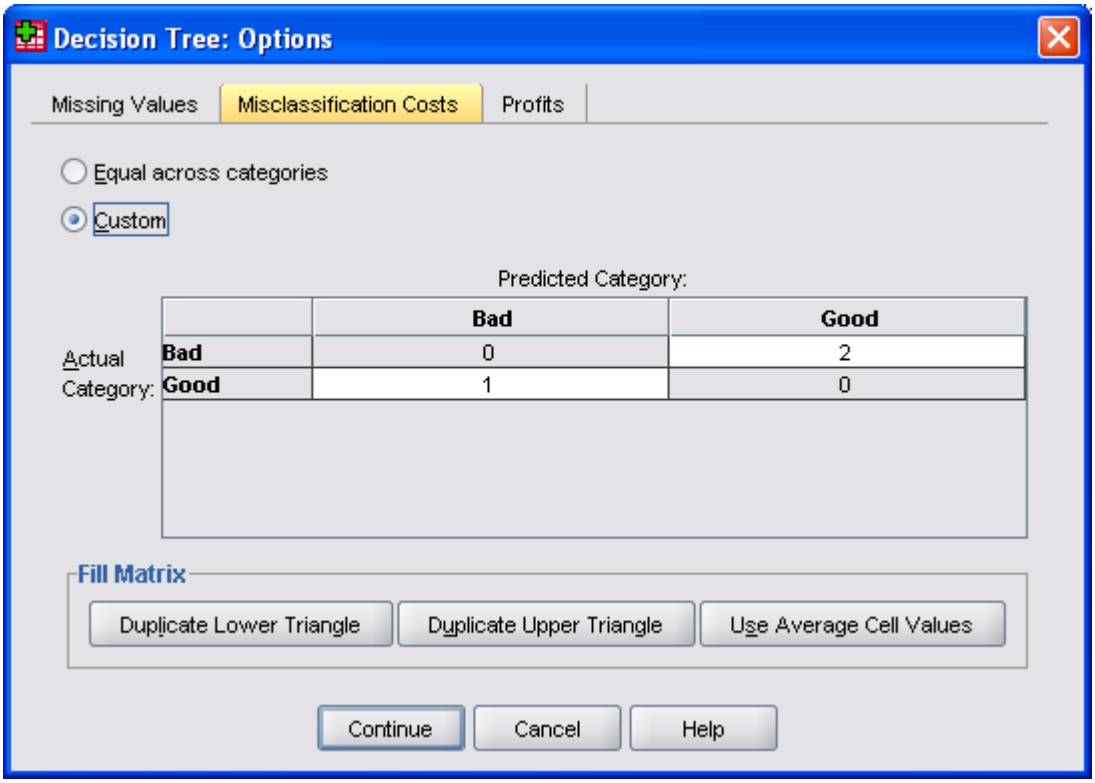


6.10 pav. Dialogo langelio ***Decision Tree: Criteria*** kortelė ***Surrogates***

- Spragtelėkite dialogo langelio ***Decision Tree*** mygtuką ***Options....*** Įvairios pasirinktys priklauso nuo metodo bei nuo priklausomo kintamojo skalės. Paminėsime svarbiausias.
- Pasirinkę CHAID (Exhaustive CHAID) metodą vardinės skalės priklausomo kintamojo analizei turėsite 6.11 pav. parodytą trijų kortelių dialogo langelį ***Decision Tree***: ***Options...*** Šio langelio kortelėje ***Misclassification Costs*** Jūs galite nurodyti santykinę vardinės ar rangų skalės priklausomo kintamojo klaidingos klasifikacijos kainą. Klaidinga klasifikacija gali turėti skirtingą kainą (pasekmes), pavyzdžiui:
  - Bankas, nesuteikęs kredito patikimam klientui ir suteikęs kreditą klientui, kuris negalės jo grąžinti, turės skirtingų nuostolių.
  - Klaidingos išvados, kada pacientui su didele širdies ligos tikimybe šios ligos tikimybė yra laikoma maža, kaina gali būti žymiai didesnė negu tada, kai pacientui su maža širdies ligos tikimybe šios ligos tikimybė yra laikoma didelė.

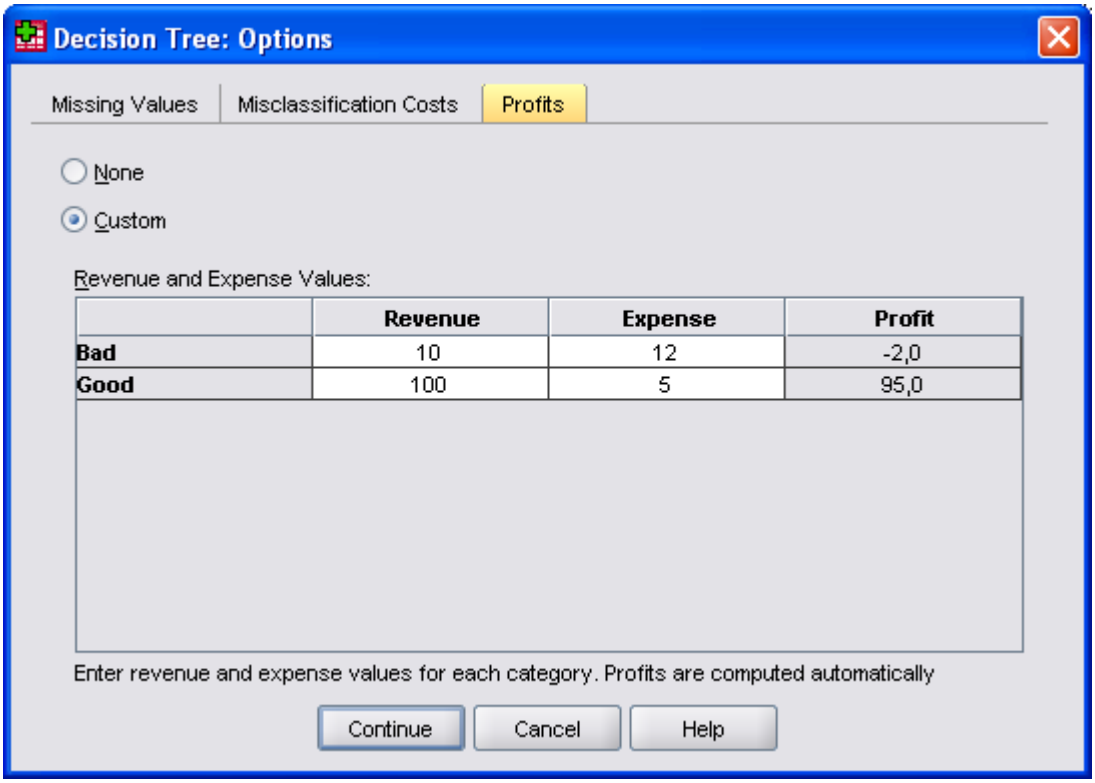
  Nustatytasis yra lygios kainos variantas – ***Equal across categories***. Norėdami nurodyti savo variantą, pasirinkite ***Custom*** ir įrašykite į lentelę savo skaičius (teigiamus). Šiame pavyzdyje laikysime, kad *bad* kategorijos palaikymas *good* kategorija kainuos dvigubai daugiau negu atvirkščiai. Tuo atveju, kai klaidingos klasifikacijos kaina yra

simetrinė, matricos sudarymą palengvina mygtukai ***Duplicate Lover Triangle***, ***Duplicate Upper Triangle***, ***Use Average Cell Values***.

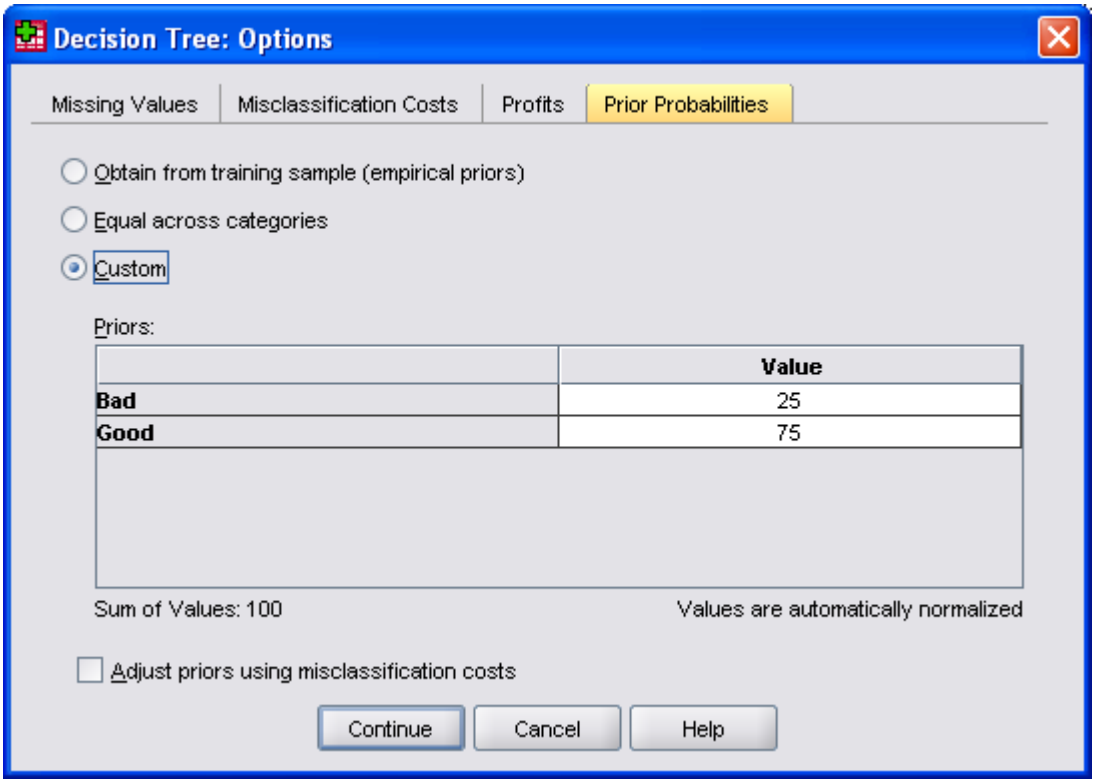


6.11 pav. Dialogo langelio ***Decision Tree: Options*** kortelė ***Misclassification Cost***

- Dialogo langelio ***Decision Tree***: ***Options...*** kortelėje ***Profits*** (6.12 pav.) pasirinkę variantą ***Custom*** kategoriniams (vardinės ar rangų skalės) kintamiesiems Jūs galite nurodyti numatomas pajamas (***Revenue***) ir išlaidas (***Expense***) pagal šių kintamųjų kategorijas. Pelnas (***Profit***) yra apskaičiuojamas iš pajamų atimant išlaidas. Pelno dydis nelemia sprendimų medžio struktūros, jis tik įtakoja investicijų grąžos rodiklius.
- CRT ir QUEST sprendimų medžio sudarymo metodai numato galimybę nurodyti kategoriniams (vardinės ar rangų skalės) kintamiesiems pirminę tikimybę įgyti konkrečios kategorijos reikšmę. Ši pirminė tikimybė nustatoma pagal bendrą kiekvienos kategorijos dažnį, neatsižvelgiant į nepriklausomus kintamuosius ir gali būti sėkmingai taikoma, kai imtis nėra reprezentatyvi. Pirminei tikimybei nurodyti atidarykite dialogo langelio ***Decision Tree: Options*** kortelę ***Prior Probabilities*** (6.13 pav.). Nustatytasis variantas yra ***Obtain from training sample (empirical priors)***, t. y. tikimybė nustatoma pagal apmokymo imties duomenys. Ši nuostata naudojama, kai imtis yra reprezentatyvi. Jūs galite pasirinkti lygių tikimybių variantą (***Equal across categories***) arba įrašyti savo reikšmes pasirinkę variantą ***Custom***. Variantą ***Adjust priors using misclassification costs*** galite pasirinkti, jei prieš tai nurodėte klaidingos klasifikacijos kainą.
- CHAID (Exhaustive CHAID) sprendimų medžio sudarymo metodas numato galimybę naujai įvertinti rangų skalės priklausomų kintamųjų kategorijų reikšmes, tuo būdu patikslinant tiek kategorijų eilės tvarką, tiek atstumą tarp kategorijų. Atidarykite dialogo langelio ***Decision Tree: Options*** kortelę ***Scores*** (6.14 pav. Nustatytasis variantas yra ***Use ordinal rank for each category***, t. y. priklausomas kintamasis suranguojamas pagal kategorijų reikšmes.

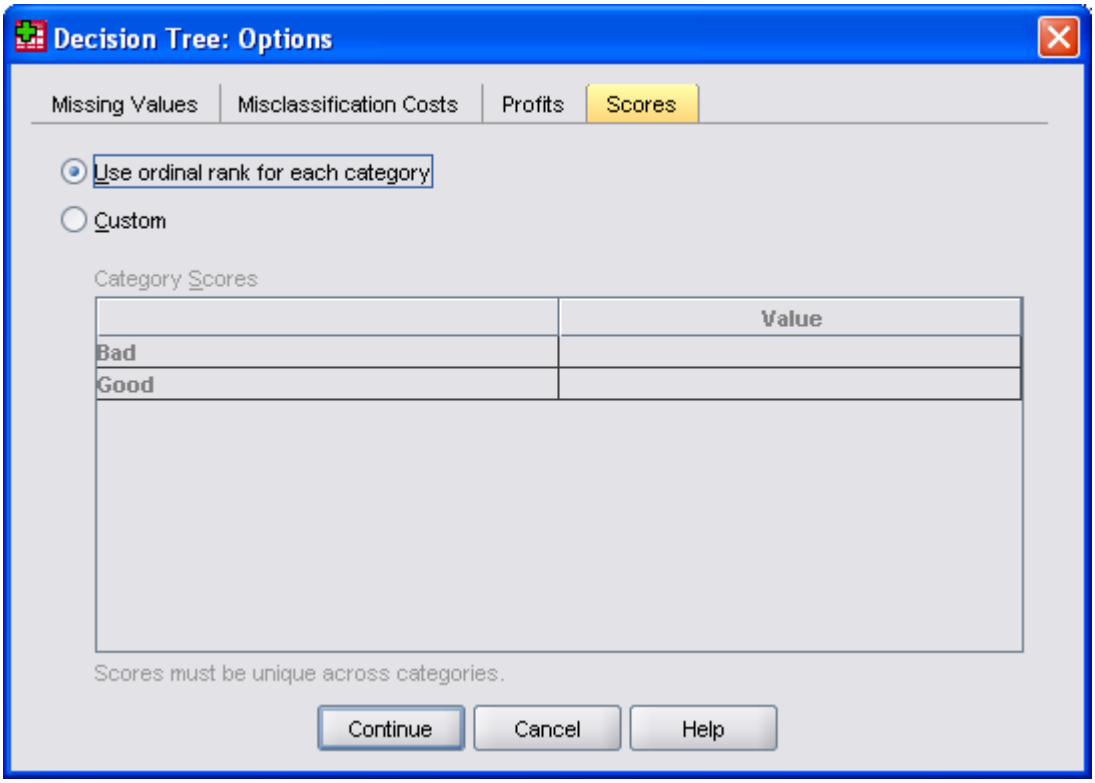


6.12 pav. Dialogo langelio ***Decision Tree: Options*** kortelė ***Profits***



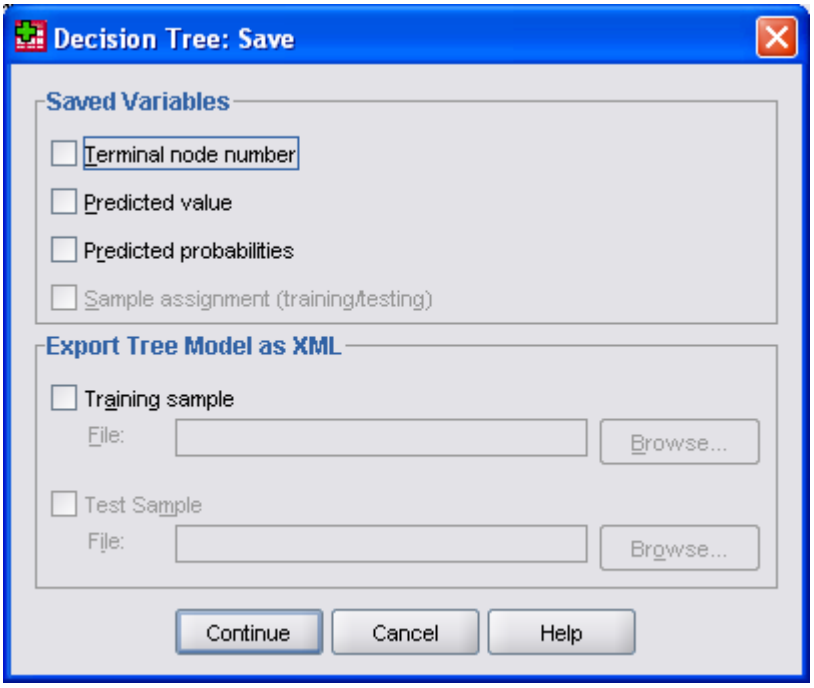
6.13 pav. Dialogo langelio ***Decision Tree: Options*** kortelė ***Prior Probabilities***

Norėdami įrašyti kategorijų reikšmes savo nuožiūra, pasirinkite variantą ***Custom*** ir įrašykite į kategorijų lentelę reikiamas reikšmes.

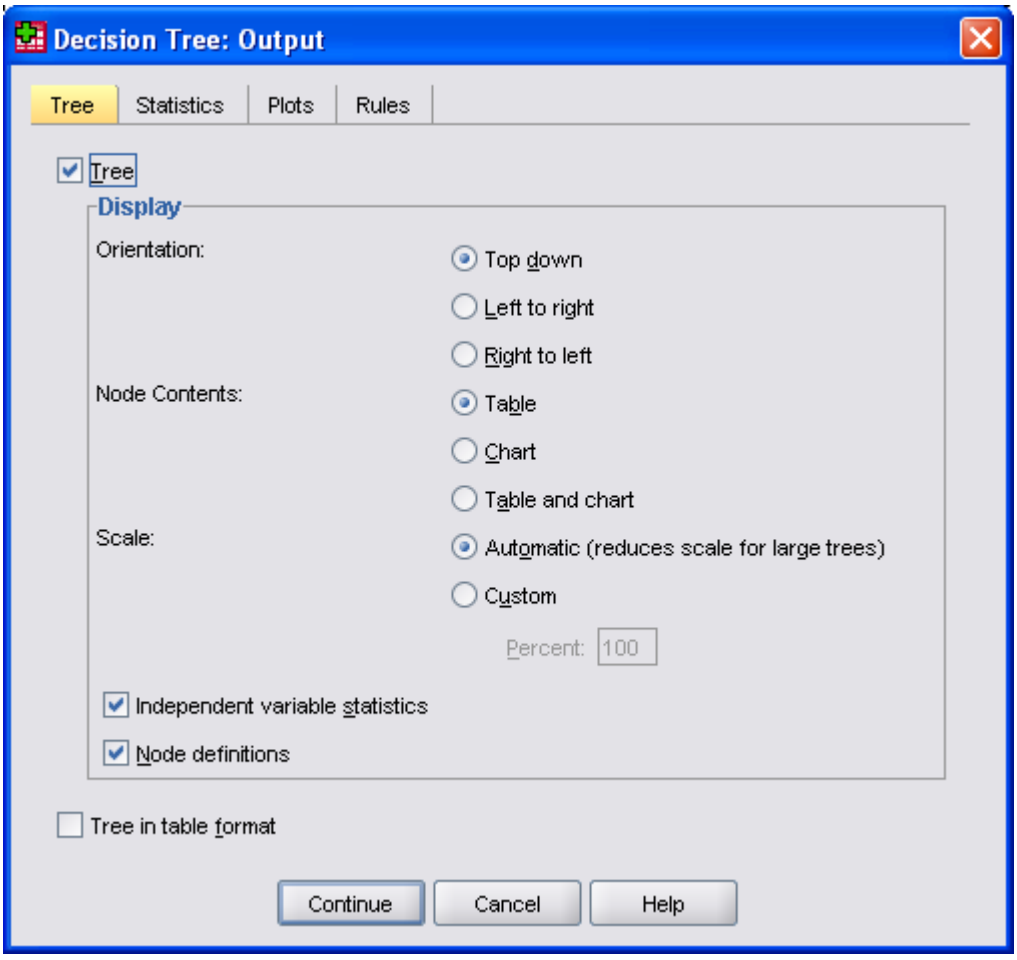


6.14 pav. Dialogo langelio ***Decision Tree: Options*** kortelė ***Scores***

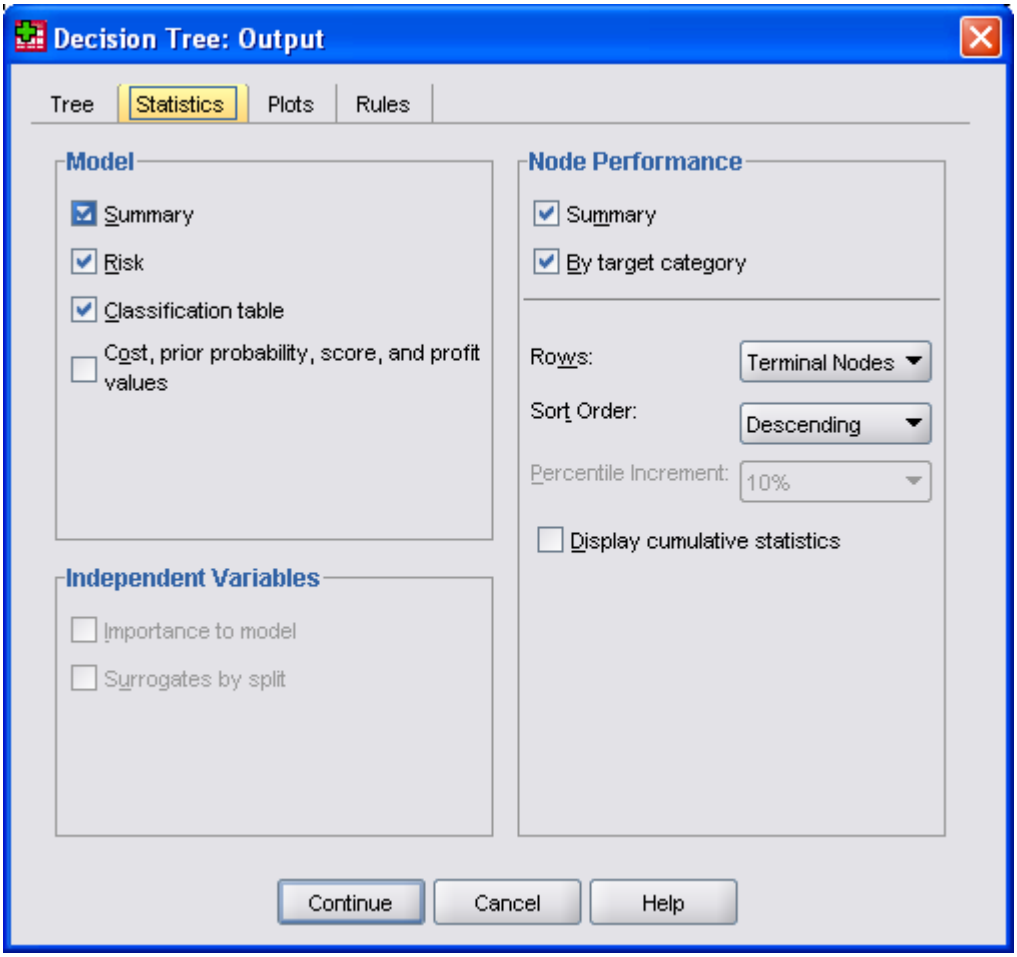
- Spragtelėkite dialogo langelio ***Decision Tree*** mygtuką ***Save...*** Dialogo langelyje ***Decision Tree: Save*** (6.15 pav.) Jūs galite nurodyti išsaugoti duomenų rinkmenoje kiekvienai priklausomo kintamojo reikšmei priskiriamą sprendimų medžio mazgą (***Terminal node number***), pritaikyto modelio numatytas priklausomo kintamojo reikšmes (***Predicted value***) bei tų reikšmių tikimybes (***Predicted probabilities***).
- Spragtelėkite dialogo langelio ***Decision Tree*** mygtuką ***Output....*** Įvairių pasirinčių galimybė priklauso nuo metodo bei nuo priklausomo kintamojo skalės. Paminėsime svarbiausias.
- Dialogo langelio ***Decision Tree***: ***Output....***kortelėje ***Tree*** (6.16 pav.) nustatytieji sprendimų medžio išvesties variantai daugeliu atvejų nereikalauja kokių nors korekcijų. Tačiau Jūs savo nuožiūra galite orientuoti sprendimų medį (komandų grupė ***Orientation***), keisti informacijos pateikimo medžio mazguose pobūdį (komandų grupė ***Node Contens***) bei mastelį (komandų grupė ***Scale***). Taip pat galite pasirinkti sprendimų medį lentelės pavidalu, pažymėję ***Tree in table format***.
- Dialogo langelio ***Decision Tree***: ***Output....***kortelėje ***Statistics*** (6.17 pav.) nustatytieji sprendimų medžio išvesties variantai taip pat yra išsamūs. Trumpai paaiškinsime jų paskirtį:
  Model komandų grupėje:
  - ***Summary.*** Suvestinėje pateikiamas metodas, į modelį įtraukti kintamieji, o taip pat nurodyti, bet į modelį neįtraukti kintamieji, kitos bendro pobūdžio charakteristikos.



6.15 pav. Dialogo langelis ***Decision Tree: Save***



6.16 pav. Dialogo langelio ***Decision Tree: Output*** kortelė ***Tree***



6.17 pav. Dialogo langelio ***Decision Tree: Save*** kortelė ***Statistics***

- ***Risk***. Rizikos įvertis ir jo standartinė paklaida. Tai sprendimų medžio prognozės tikslumo matas. Vardinės bei rangų skalės priklausomiems kintamiesiems apskaičiuojamas kaip klaidingai klasifikuotų duomenų dalis (įvertinant klaidingos klasifikacijos kainą ir pirminę tikimybę).
- ***Classification table***. Vardinės ir rangų skalės priklausomiems kintamiesiems lentelėje pateikiamas teisingai ir klaidingai suklasifikuotų duomenų skaičius pagal kiekvieną priklausomo kintamojo kategoriją.
- ***Cost, prior probability, score, and profit values***. Išvestyje pateikiamos modelio sudaryme pasirinktos klaidingos klasifikacijos kainos, pirminės tikimybės, kintamojo kategorijų vertinimo ir pelno reikšmės.

Independent Variables komandų grupėje:

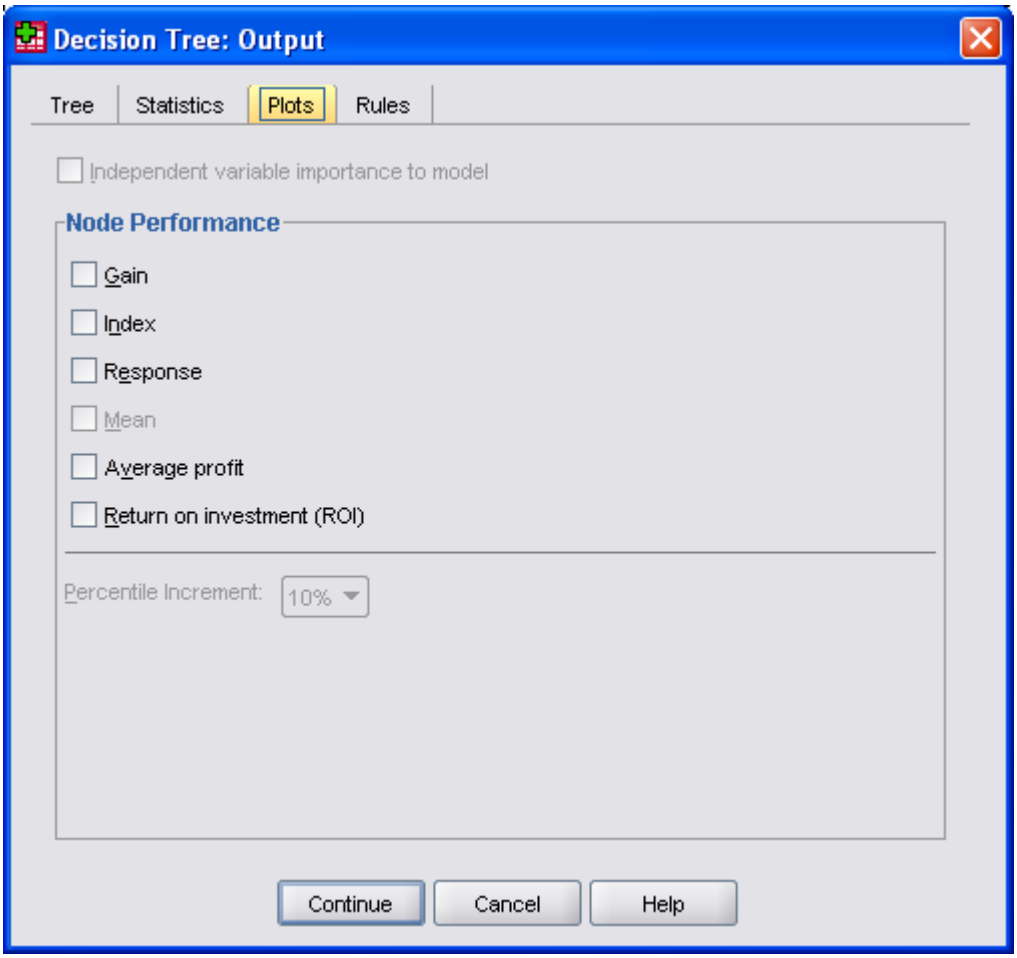
- ***Importance to model***. Esant CRT metodui, nepriklausomi kintamieji suranguojami pagal jų įtaką priklausomam kintamajam. QUEST ir CHAID metodams ši pasirinktis yra neaktyvi.
- ***Surrogates by split***. Esant CRT ir QUEST metodams, pateikiami surogatiniai kintamieji, jeigu jie buvo nurodyti.

Node Performance komandų grupėje:

- ***Summary***. Intervaliniams kintamiesiems lentelėje pateikiamas sprendimų medžio mazgo numeris, duomenų skaičius mazge, priklausomo kintamojo vidurkis. Vardinės ir rangų skalės kintamiesiems (esant nurodytai pelno reikšmei)

pateikiamas sprendimų medžio mazgo numeris, duomenų skaičius mazge, pelno vidurkis, investicijų grąža.

– ***By target category***. Vardinės ir rangų skalės kintamiesiems pagal pasirinktą (*target*) kategoriją pateikiama procentinė pasirinktos kategorijos duomenų dalis (*percentage gain*) atžvilgiu visų pasirinktos kategorijų duomenų, atsako procentas, indeksas tiek pagal sprendimų medžio mazgus, tiek pagal procentiles.
– ***Rows***. Pasirenkama rezultatų pateikimo forma: ***Terminal Nodes*** – pagal sprendimų medžio galinius mazgus, ***Percentiles*** – pagal procentiles, ***Both*** – pateikiami abu variantai.
– ***Percentile Increment***. Nurodomas procentilių prieaugis: 1, 2, 5, 10, 20, 25. Nustatytoji reikšmė – 10.



6.18 pav. Dialogo langelio ***Decision Tree: Save*** kortelė ***Plots***

- Įvairių pasirinčių galimybė dialogo langelio ***Decision Tree***: ***Output...***kortelėje ***Plots*** (6.18 pav.) priklauso nuo metodo bei nuo priklausomo kintamojo skalės:
  – Pasirinktis ***Independent variable importance to model*** galima tikCRT modeliui. Pateikiama nepriklausomų kintamųjų įtakos laipsnio stulpelinė diagrama.

  Node Performance komandų grupėje:
  – ***Gain***. Vardinės ir rangų skalės priklausomiems kintamiesiems koeficientas *gain* yra apibrėžiamas kaip procentinė pasirinktos kategorijos duomenų dalis sprendimų medžio mazge atžvilgiu visų pasirinktos kategorijų duomenų. Papildomai skaitmeninėms reikšmėms (esant pažymėtam ***Statistics*** kortelėje variantui ***By***

target category) pateikiamas linijinis sukauptų pagal procentiles koeficiento reikšmių grafikas.

- ***Index***. Vardinės ir rangų skalės priklausomiems kintamiesiems indeksas yra apibrėžiamas kaip pasirinktos kategorijos atsako procentinės dalies sprendimų medžio mazge santykis su visos imties atsako procentine dalimi. Papildomai skaitmeninėms reikšmėms (esant pažymėtam ***Statistics*** kortelėje variantui ***By target category***) pateikiamas linijinis sukauptų pagal procentiles indekso reikšmių grafikas.
- ***Response***. Vardinės ir rangų skalės priklausomiems kintamiesiems atsakas yra apibrėžiamas kaip procentinė pasirinktos kategorijos duomenų dalis sprendimų medžio mazge atžvilgiu visų mazgo duomenų. Papildomai skaitmeninėms reikšmėms (esant pažymėtam ***Statistics*** kortelėje variantui ***By target category***) pateikiamas linijinis sukauptų pagal procentiles koeficiento reikšmių grafikas.
- ***Mean***. Intervaliniams priklausomiems kintamiesiems pateikiamas linijinis sukauptų pagal procentiles vidurkio reikšmių grafikas.
- ***Average profit***. Vardinės ir rangų skalės priklausomiems kintamiesiems pateikiamas (esant nurodytam pelno rodikliui) linijinis sukaupto pagal procentiles pelno grafikas.
- ***Return on investment (ROI)***. Vardinės ir rangų skalės priklausomiems kintamiesiems pateikiamas (esant nurodytam pelno rodikliui) linijinis sukauptos pagal procentiles investicinės grąžos grafikas. Investicinė grąža yra apibrėžiama kaip pelno ir išlaidų santykis.

- Spragtelėkite mygtuką ***OK*** pagrindiniame dialogo langelyje ***Decision Tree***. Padrindiniai sprendimų medžio rezultatai parodyti 6.19.1–6.19.6 pav.

Lentelėje *Model Summary* (6.19.1 pav.) pateikiama bendro pobūdžio informacija apie modelio sudarymo nuostatas bei gautus rezultatus: *Specifications* dalyje nurodomas pasirinktas metodas, naudojami kintamieji ir kt., *Results* dalyje – bendras ir galinių mazgų skaičius, sprendimų medžio lygių skaičius žemiau pagrindinio mazgo, į modelį įtraukti statistiškai reikšmingi nepriklausomi kintamieji. Nagrinėjamame pavyzdyje, tai nepriklausomi kintamieji – pajamos (*Income level*), kredito kortelių skaičius (*Number of credit cards*) ir amžius (*Age*). Tuo tarpu, nepriklausomi kintamieji: išsilavinimas (*Education*) ir nuomojami automobiliai (*Car loans*) pripažinti neturinčiais esminės įtakos priklausomam kintamajam ir į modelį neįtraukti.

Lentelėse *Misclassification Costs* ir *Profits* (6.19.1 pav.) atitinkamai nurodoma santykinė priklausomo kintamojo klaidingos klasifikacijos kaina ir pajamos, išlaidos bei pelnas pagal priklausomo kintamojo kategorijas.

Sprendimų medžio diagrama (6.19.2 pav.) yra grafinė gauto modelio išraiška. Iš diagramos seka, kad:

- Pagal CHAID metodą nepriklausomas kintamasis *Income level* turi didžiausią įtaką priklausomam kintamajam *Credit rating*.
- Esant pajamų kategorijai "mažos", nepriklausomas kintamasis *Income level* yra vienintelis, rurintis esminę įtaką priklausomam kintamajam. 82,1% šios kategorijos banko klientų turi problemų su kredito grąžinimu. Kadangi šis sprendimų medžio mazgas (*Node* 1) neturi tęsinio, jis vadinamas galiniu (*Terminal node*).
- Esant pajamų kategorijai "vidutinės" ir "didelės", kitas nepriklausomas kintamasis, turintis esminę įtaką priklausomam kintamajam, yra kredito kortelių skaičius.
- Klientams, turintiems vidutines pajamas ir penkias ar daugiau kredito kortelių, nurodomas dar vienas statistiškas reikšmingas nepriklausomas kintamasis – amžius. 80,8 % jaunenių kaip 28 metai šių kategorijų klientų turi problemų su kredito grąžinimu, kai tuo tarpu vyreniems kaip 28 metai šis procentas yra 43,7 %.

Dukart spragtelėjus pele sprendimų medžio diagramą, galima ją redaguoti: paslėpti atskiras šakas, keisti spalvas, šriftą ir t. t.

Model Summary

Specifications	Growing Method	CHAID
	Dependent Variable	Credit rating
	Independent Variables	Age, Income level, Number of credit cards, Education, Car loans
	Validation	Cross Validation
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	400
	Minimum Cases in Child Node	200
Results	Independent Variables Included	Income level, Number of credit cards, Age
	Number of Nodes	10
	Number of Terminal Nodes	6
	Depth	3

Misclassification Costs

Observed	Predicted	
	Bad	Good
Bad	,000	2,000
Good	1,000	,000

Dependent Variable: Credit rating

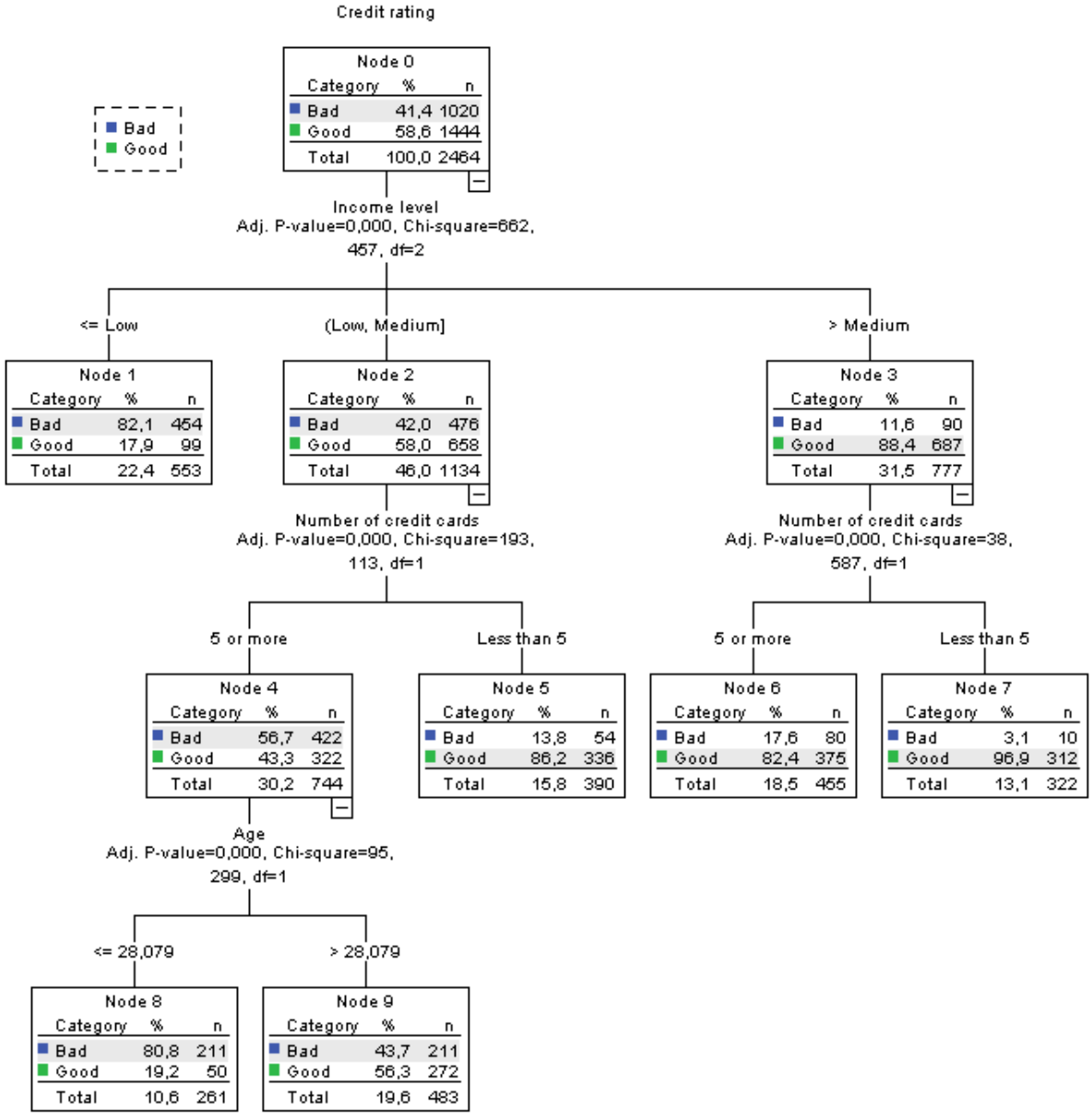
Profits

Credit rating	Revenue	Expense	Profit
Bad	10,000	12,000	-2,000
Good	100,000	25,000	75,000

6.19.1 pav. Pagrindiniai sprendimų medžio rezultatai

Esant pažymėtam laukeliui ***Tree in table format*** dialogo langelio ***Decision Tree: Output*** kortelėje ***Tree***, sprendimų medžio rezultatai papildomai pateikiami ir lentelės (*Tree Table*) forma (6.19.3 pav.). Kiekvienam mazgui lentelėje nurodoma:

- Kiekvienos priklausomo kintamojo kategorijos duomenų skaičius ir procentinė dalis.
- Prognozuojama priklausomo kintamojo kategorija (*Predicted Category*). Nagrinėjamame pavyzdyje, tai kategorija, turinti mazge daugiau 50 % duomenų (viso yra dvi kategorijos).
- Pirminis mazgas (*Parent Node*).
- Nepriklausomas kintamasis, kuriuo pagrindu sudaromas mazgas.
- Mazgo sudarymo χ^2 kriterijaus reikšmė (CHAID modeliui), laisvės laipsnių skaičius ir kriterijaus *p*-reikšmė.
- Nepriklausomų kintamųjų, pagal kuriuos sudaromas mazgas, reikšmės.



6.19.2 pav. Pagrindiniai sprendimų medžio rezultatai

Lentelėje *Gain Summary for Nodes* (6.19.3 pav.) vardinės ir rangų skalės kintamiesiems (esant nurodytai pelno reikšmei) pateikiamas sprendimų medžio mazgo numeris, duomenų skaičius mazge, pelno vidurkis (*Profit*), investicijų grąža (*ROI*). Lentelėje *Gain Summary for Percentiles* (6.19.3 pav.) duomenų skaičius, pelno vidurkis ir investicijų grąža pateikiama pagal procentiles. Šalia lentelių taip pat pateikiami pelno vidurkio ir investicijų grąžos priklausomybės nuo procentilių grafikai.

Toliau seka informacija pagal kiekvieną pasirinktą priklausomo kintamojo kategoriją (*Target Category*). Lentelėje *Gains for Nodes* (6.19.4 pav.) pateikiama išsami informacija apie galinius sprendimų medžio mazgus (*Terminal Nodes*), t. y. mazgus, ties kuriais baigiasi priklausomo kintamojo klasifikacija pagal atitinkamą nepriklausomą kintamąjį. Nagrinėjamame pavyzdyje informacija pateikiama apie vieną pasirinktą priklausomo kintamojo kategoriją.

Tree Table

Node	Bad		Good		Total		Predicted Category	Parent Node	Primary Independent Variable				
	N	Percent	N	Percent	N	Percent			Variable	Sig.[a]	Chi-Square	df	Split Values
0	1020	41,4%	1444	58,6%	2464	100,0%	Bad						
1	454	82,1%	99	17,9%	553	22,4%	Bad	0	Income level	,000	662,457	2	<= Low
2	476	42,0%	658	58,0%	1134	46,0%	Bad	0	Income level	,000	662,457	2	(Low, Medium]
3	90	11,6%	687	88,4%	777	31,5%	Good	0	Income level	,000	662,457	2	> Medium
4	422	56,7%	322	43,3%	744	30,2%	Bad	2	Number of credit cards	,000	193,113	1	5 or more
5	54	13,8%	336	86,2%	390	15,8%	Good	2	Number of credit cards	,000	193,113	1	Less than 5
6	80	17,6%	375	82,4%	455	18,5%	Good	3	Number of credit cards	,000	38,587	1	5 or more
7	10	3,1%	312	96,9%	322	13,1%	Good	3	Number of credit cards	,000	38,587	1	Less than 5
8	211	80,8%	50	19,2%	261	10,6%	Bad	4	Age	,000	95,299	1	<= 28,079
9	211	43,7%	272	56,3%	483	19,6%	Bad	4	Age	,000	95,299	1	> 28,079

Growing Method: CHAID
Dependent Variable: Credit rating

a. Bonferroni adjusted

Gain Summary for Nodes

Node	N	Percent	Profit	ROI
7	322	13,1%	48,385	196,7%
5	390	15,8%	42,800	184,5%
6	455	18,5%	40,857	179,9%
9	483	19,6%	27,284	141,2%
8	261	10,6%	7,962	54,9%
1	553	22,4%	7,309	51,0%

Growing Method: CHAID
Dependent Variable: Credit rating

Gain Summary for Percentiles

Percentile	Nodes	N	Profit	ROI
10	7	246	48,385	196,7%
20	7 ; 5	493	46,449	192,6%
30	5 ; 6	739	45,161	189,8%
40	6	986	44,085	187,4%
50	6 ; 9	1232	42,724	184,3%
60	9	1478	40,150	178,1%
70	9 ; 8	1725	37,474	171,4%
80	8 ; 1	1971	33,765	161,2%
90	1	2218	30,826	152,6%
100	1	2464	28,474	145,1%

Growing Method: CHAID
Dependent Variable: Credit rating

6.19.3 pav. Pagrindiniai sprendimų medžio rezultatai

Lentelėje pateikiama:

- Duomenų skaičius mazge (*Node N*) ir jo procentinė dalis atžvilgiu bendro duomenų skaičiaus.
- Pasirinktos kategorijos duomenų skaičius mazge (*Gain N*) ir procentinė pasirinktos kategorijos duomenų dalis sprendimų medžio mazge atžvilgiu bendro pasirinktos kategorijų duomenų skaičiaus.
- Procentinė pasirinktos kategorijos duomenų dalis sprendimų medžio mazge atžvilgiu bendro duomenų skaičiaus mazge (*Response*).
- Procentinės pasirinktos kategorijos duomenų dalies sprendimų medžio mazge atžvilgiu bendro duomenų skaičiaus mazge santykis su procentine bendra pasirinktos kategorijos duomenų dalimi atžvilgiu bendro duomenų skaičiaus, t. v. indeksas (*Index*). Kitaip tariant, indeksas yra mazgo *Response* santykis su pagrindinio mazgo (*root node*) *Response*. Indekso reikšmės, didesnės nei 100 %, reiškia, kad pasirinktos kategorijos duomenų dalis konkrečiame mazge yra didesnė nei bendra pasirinktos kategorijos duomenų dalis. Ir atvirkščiai, indekso reikšmės, mažesnės nei 100 %, reiškia, kad pasirinktos kategorijos duomenų dalis konkrečiame mazge yra mažesnė nei bendra pasirinktos kategorijos duomenų dalis.

Lentelėje *Gains for Percentiles* (6.19.4 pav.) pateikiama analogiška informacija procentilių atžvilgiu.

Gains for Nodes

Node	Node		Gain		Response	Index
	N	Percent	N	Percent		
1	553	22,4%	454	44,5%	82,1%	198,3%
8	261	10,6%	211	20,7%	80,8%	195,3%
9	483	19,6%	211	20,7%	43,7%	105,5%
6	455	18,5%	80	7,8%	17,6%	42,5%
5	390	15,8%	54	5,3%	13,8%	33,4%
7	322	13,1%	10	1,0%	3,1%	7,5%

Growing Method: CHAID
Dependent Variable: Credit rating

Gains for Percentiles

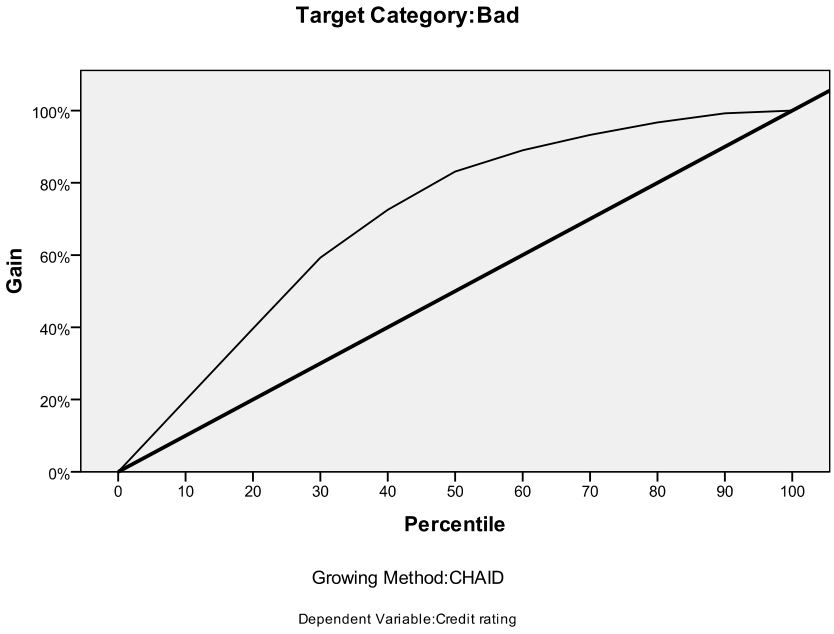
Percentile	Nodes	N	Gain		Response	Index
			N	Percent		
10	1	246	202	19,8%	82,1%	198,3%
20	1	493	405	39,7%	82,1%	198,3%
30	1 ; 8	739	605	59,3%	81,8%	197,6%
40	8 ; 9	986	740	72,5%	75,1%	181,4%
50	9	1232	848	83,1%	68,8%	166,2%
60	9 ; 6	1478	908	89,0%	61,4%	148,3%
70	6	1725	951	93,3%	55,1%	133,2%
80	6 ; 5	1971	986	96,7%	50,0%	120,9%
90	5 ; 7	2218	1012	99,2%	45,7%	110,3%
100	7	2464	1020	100,0%	41,4%	100,0%

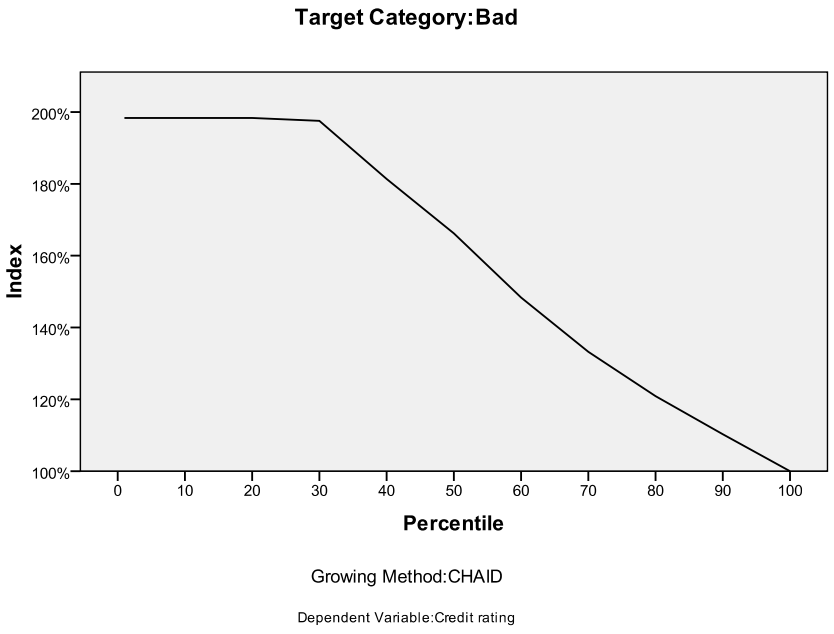
Growing Method: CHAID
Dependent Variable: Credit rating

6.19.4 pav. Pagrindiniai sprendimų medžio rezultatai

Kiek modelis yra geras galima spręsti pagal koeficiento *Gain* priklausomybės nuo procentilių grafiką (6.19.5 pav.). Kuo greičiau priklausomybės kreivė priartėja prie 100 % ir nusistovi ties ta reikšme, tuo modelis geresnis. Jei priklausomybės kreivė yra artima atraminei įstrižainei, modelis nesuteikia jokios naujos informacijos.

Kitas modelio tinkamumo rodiklis yra indekso priklausomybės nuo procentilių grafikas (6.19.5 pav.). Modelis tuo geresnis, kuo indekso reikšmė ilgiau išlieka artima pradinei reikšmei (didesnei negu 100 %) ir po to staigiai krenta iki 100 % atžymos. Modelis nesuteikia naujos informacijos, jeigu indekso kreivė svyruoja apie 100 %.





6.19.5 pav. Pagrindiniai sprendimų medžio rezultatai

Lentelėje *Risk* (6.19.5 pav.) pateikiamas rizikos klaidingai numatyti priklausomo kintamojo kategoriją įvertis. Šiame pavyzdyje rizikos įvertis yra 0,288, t. y. beveik 29 % vertinimo ("blogas", "geras") atvejų bus klaidingi. Lentelėje *Classification* (6.19.5 pav.) pateikiamas bendras klasifikacijos tikslumas (*Overall Percentage*) bei pagal atskiras priklausomo kintamojo kategorijas. Kadangi dialogo langelio ***Decision Tree***: ***Options...*** kortelėje ***Misclassification Costs*** (6.11 pav.) mes nurodėme dvigubai didesnę santykinę klaidingos klasifikacijos kainą kategoriją "blogas" priimant kaip "geras" negu kategoriją "geras" priimant kaip "blogas", modelyje teikiamas prioritetas būtent tikslinės kategorijos "blogas" prognozės tikslumui – šiai kategorijai klasifikacijos tikslumas sudaro 85,9 %. Tačiau šis tikslumas pasiekiamas kategorijos "geras" prognozės tikslumo sąskaita.

Risk

Method	Estimate	Std. Error
Resubstitution	,288	,011
Cross-Validation	,288	,011

Growing Method: CHAID

Dependent Variable: Credit rating

Classification

Observed	Predicted		
	Bad	Good	Percent Correct
Bad	876	144	85,9%
Good	421	1023	70,8%
Overall Percentage	52,6%	47,4%	77,1%

Growing Method: CHAID

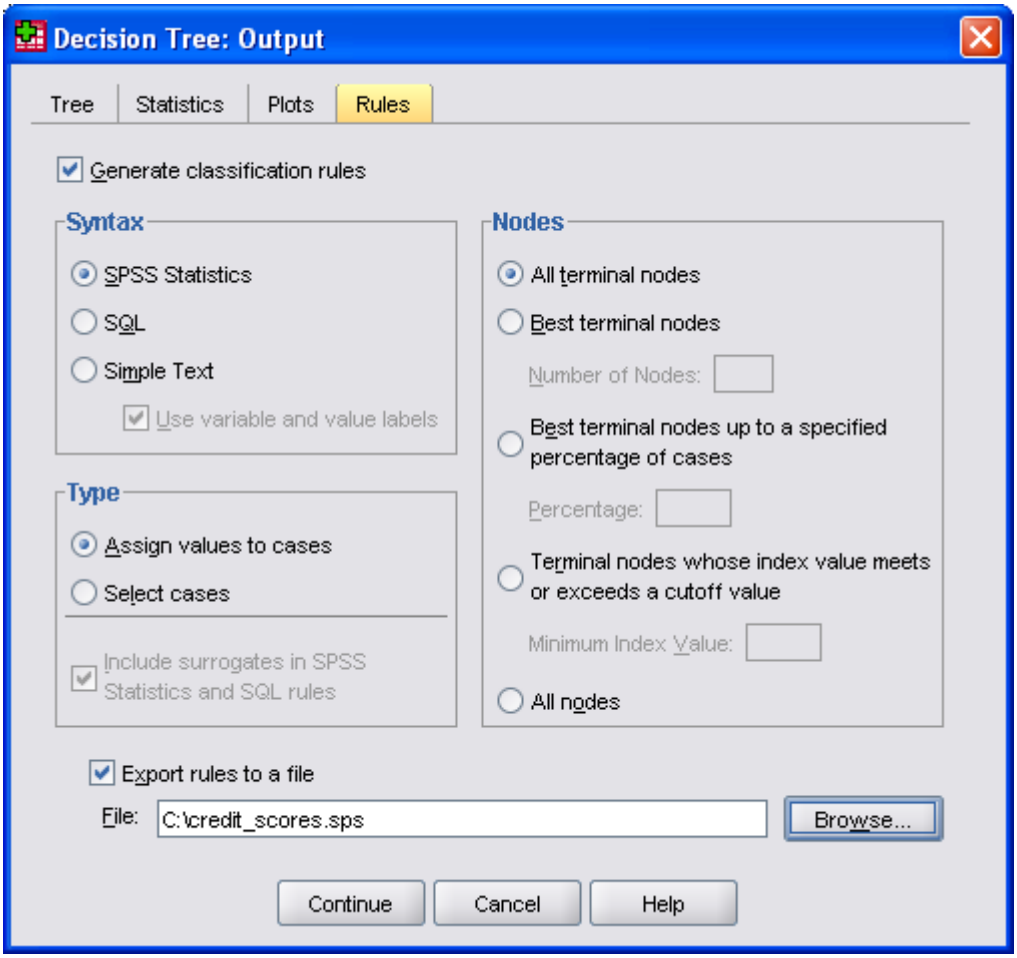
Dependent Variable: Credit rating

6.19.6 pav. Pagrindiniai sprendimų medžio rezultatai

6.2. MODELIO TAIKYMAS KITIEMS DUOMENIMS

Viena iš svarbiausių sprendimų medžio galimybių, kuri dažnai panaudojama praktikoje yra sudaryto pagal turimus statistinius duomenis modelio pritaikymas naujų duomenų prognozei kai priklausomo kintamojo reikšmės yra nežinomos. Pavyzdžiui, remdamasis turimais kreditingumo duomenimis bankas gali sudaryti kreditingumo vertinimo modelį, kurį gali naudoti naujų klientų galimybių prognozei. Norėdami pritaikyti sudarytą modelį naujų duomenų prognozei:

- Atlikite sprendimų medžio sudarymo veiksmus, nurodytus 6.1 poskyryje.
- Papildomai aprašytiems veiksmams, atidarykite dialogo langelio ***Decision Tree***: ***Output...***kortelę ***Rules*** (6.20 pav.).
- Pažymėkite laukelį ***Generate classification rules***.
- ***Syntax*** komandų grupėje, kur galima pasirinkti modelio klasifikacijos taisyklių, kurios bus išsaugotos atskiroje duomenų rinkmenoje, formą (***SPSS Statistics***, ***SQL*** ar ***Simple Text***) palikite nustatytąjį variantą ***SPSS Statistics***.
- ***Type*** komandų grupėje palikite nustatytąjį variantą ***Assign values to cases***.
- Pažymėkite laukelį ***Export rules to a file*** ir nurodykite duomenų rinkmenos pavadinimą bei direktoriją. Direktorijai nurodyti pasinaudokite peržiūros mygtuku ***Browse...***
- ***Nodes*** komandų grupėje palikite dažniausiai naudojamą nustatytąjį variantą ***All terminal nodes***.



6.20 pav. Dialogo langelio ***Decision Tree: Output*** kortelė ***Rules***

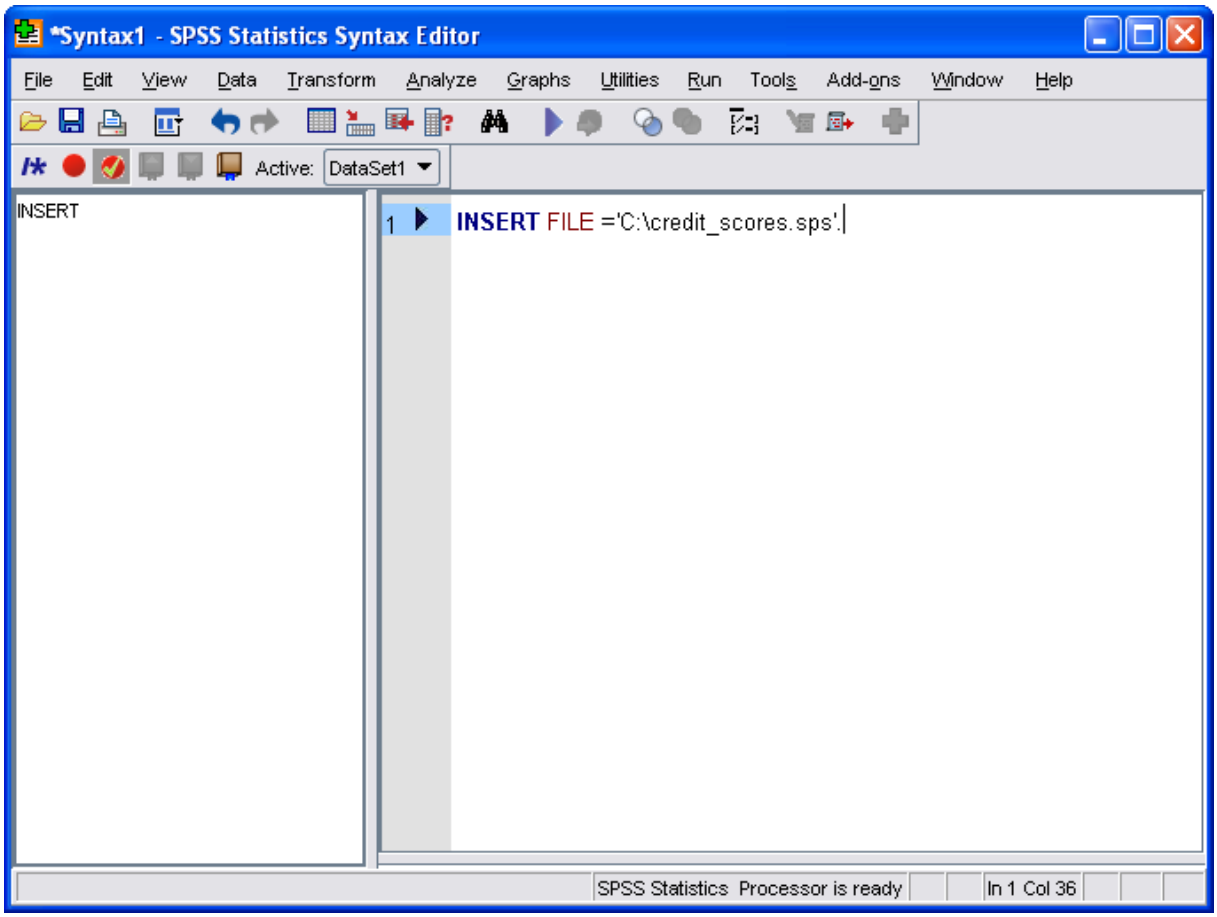
- Spragtelėkite dialogo langelio ***Decision Tree***: ***Output...*** mygtuką ***Continue***, po to – dialogo langelio ***Decision Tree*** mygtuką ***OK***.
- Sprendimų medžio rezultatai bus pateikti išvesties lange ***SPSS Statistics Viewer*** , o duomenų klasifikacijos taisyklės pagal sudarytą sprendimų medžio modelį – nurodytoje rinkmenoje.

Įsitikinus, kad sprendimų medžio modelis yra tinkamas, galima jį pritaikyti naujiems duomenims. Naujų duomenų rinkmenoje turi būti tie nepriklausomi kintamieji, kurie sprendimų medžio modelyje pripažinti turintys statistiškai reikšmingą įtaką priklausomam kintamajam. Mūsų nagrinėjamame pavyzdyje tai – pajamos (*Income level*), kredito kortelių skaičius (*Number of credit cards*) ir amžius (*Age*). Analizės rezultate bus gautas naujas kintamasis, apibrėžiantis priklausomo kintamojo priklausomybę vienai iš kategorijų (“blogas” arba “geras”).

- Atidarykite rinkmeną su naujais duomenimis.
- Nurodykite komandas ***File → New →Syntax***.
- Komandų sintaksės lange ***SPSS Statistics Syntax Editor*** (6.21 pav.) nurodykite klasifikacijos taisyklių rinkmenos pilną adresą ir pavadinimą.
- Nurodykite komandas ***Run → All***.

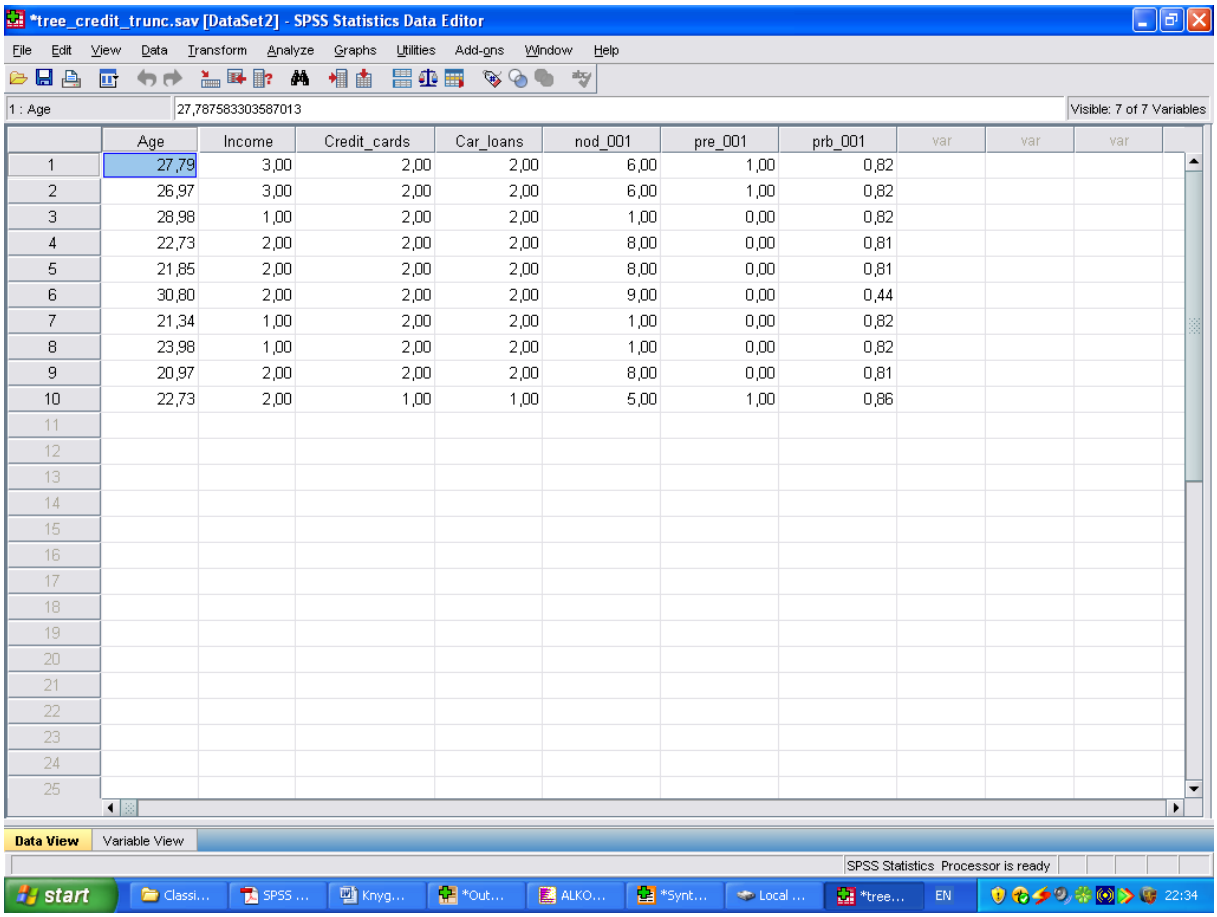
Duomenų rinkmenoje bus sukurti šie nauji kintamieji (6.22 pav.): *nod_001*, *pre_001* ir *prb_001*:

– *nod_001* nurodo galinio mazgo numerį kiekvienam atvejui;



6.21 pav. Komandos užrašymo pavyzdys SPSS sintaksės lange

– *pre_001* nurodo priklausomo kintamojo reikšmę kiekvienam atvejui (kreditingumo įvertinimas nagrinėjamame pavyzdyje 0 – "blogas", 1 – "geras");
– *prb_001* nurodo priklausomo kintamojo reikšmės tikimybę kiekvienu atveju;

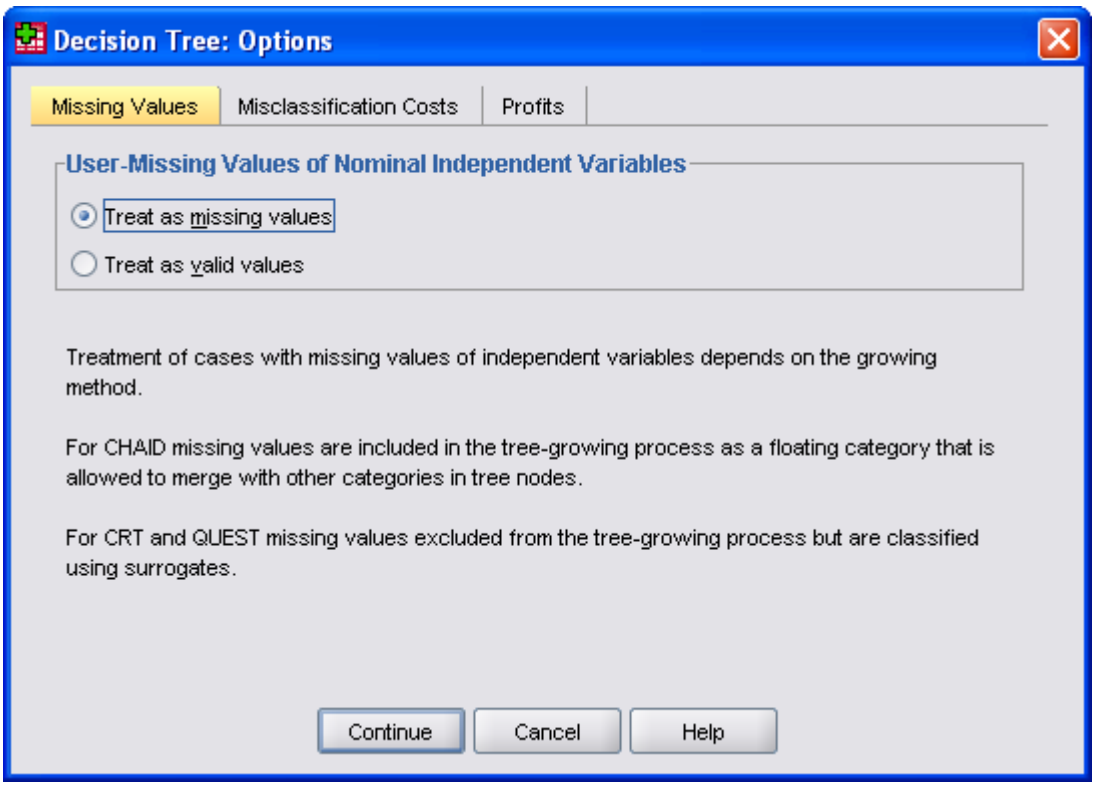


6.22 pav. Naujai sukurti klasifikaciniai duomenys duomenų rinkmenoje

6.3. PRALEISTI DUOMENYS SPRENDIMŲ MEDŽIO MODELIUOSE

Vartotojo nustatytos praleistos vardinės skalės priklausomo kintamojo reikšmės į. dialogo langelio ***Decision Tree: Categories*** laukelį ***Exclude*** (6.2 pav.) patalpinamos automatiškai. Tačiau jas taip pat galima įtraukti į analizę. Intervalų ir rangų skalės priklausomų kintamųjų sisteminiai bei vartotojo nustatytos praleistos reikšmės visuomet pašalinamos iš analizės.

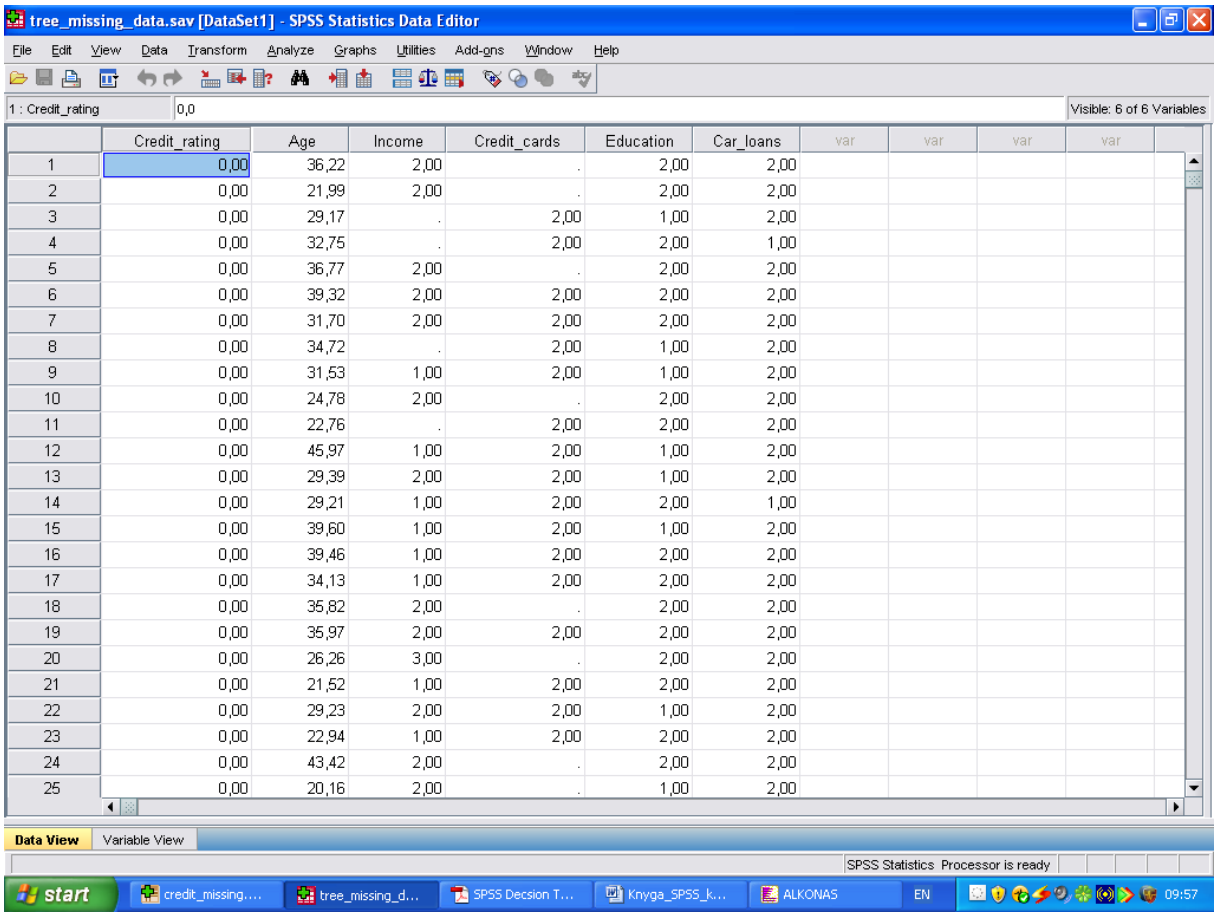
Vardinės skalės nepriklausomų kintamųjų vartotojo nustatytos praleistos reikšmės traktuojamos sutinkamai su dialogo langelio ***Decision Tree: Options*** kortelės ***Missing Values*** (6.23 pav.) nuostata. Nustatytasis variantas yra traktuoti kaip praleistas reikšmes (***Treat as missing values***). Esant šiai nuostatai vartotojo nustatytos praleistos reikšmės traktuojamos kaip sisteminiai praleistos, kuriuos, savo ruožtu traktuojamos priklausomai nuo pasirinkto sprendimų medžio sudarymo metodo. Pažymėjus laukelį ***Treat as valid values*** vardinės skalės nepriklausomų kintamųjų vartotojo nustatytos praleistos reikšmės yra traktuojamos kaip įprasti duomenys sprendimų medžio klasifikacijoje.



6.23 pav. Dialogo langelio ***Decision Tree: Options*** kortelė ***Missing Values***

- Taikant CHAID (Exhaustive CHAID) metodą visos nepriklausomo kintamojo sisteminiai bei vartotojo nustatytos praleistos reikšmės yra traktuojamos kaip viena kategorija. Intervalų ir rangų skalės nepriklausomiems kintamiesiems algoritmas generuoja kategorijas pagal galiojančias reikšmes ir po to sprendžia, ar sujungti praleistų reikšmių kategoriją su artimiausia galiojančių reikšmių kategorija ar traktuoti ją kaip atskirą.
- Taikant CRT ir QUEST metodus, praleistoms nepriklausomų kintamųjų reikšmėms pakeisti gali būti naudojami surogatiniai duomenys, t. y. nauji kintamieji, turintys glaudų ryšį su originaliu kintamuoju.

Skirtumą tarp CHAID ir CRT metodų, operuojant su praleistomis reikšmėmis, pademonstruosime aukščiau išnagrinėtu pavyzdžiu, kuriame dalį nepriklausomų kintamųjų reikšmių praleisime (6.24 pav.)

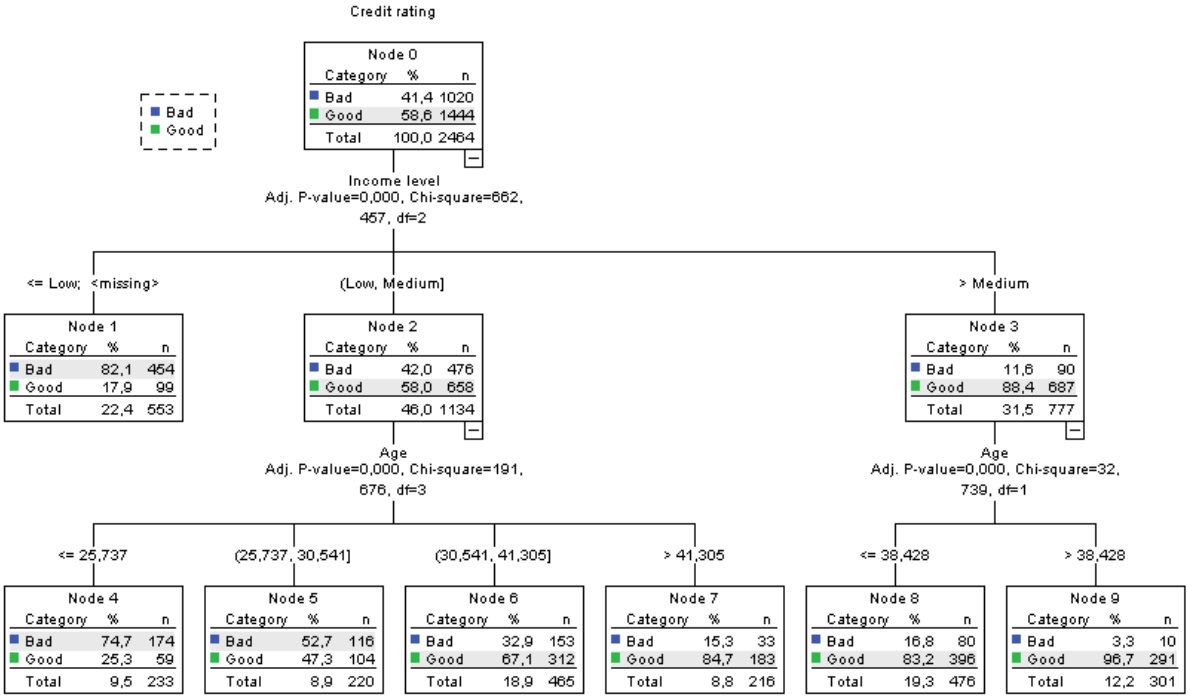


6.24 pav. Praleistos nepriklausomų kintamųjų reikšmės duomenų rinkmenoje

Supaprastintam sprendimų medžio variantui pagal CHAID metodą:

- Nurodome komandas ***Analyze → Classify →Tree...***
- Priklausomą kintamąjį *Credit rating* įkeliame į dialogo langelio ***Decision Tree*** (6.1 pav.). laukelį ***Dependent Variable***, o nepriklausomus kintamuosius – į laukelį ***Independent Variables***.
- Paliekame nustatytąjį CHAID metodą (laukelyje ***Growing Method***).
- Spragtelime dialogo langelio ***Decision Tree*** mygtuką ***Criteria...*** ir šio langelio kortelės ***Growth Limits*** (6.4 pav.) laukelyje ***Minimum Number of Cases*** pasirenkam mažiausią pagrindinius (motininius) mazgus ***Parent Node*** sudarančių reikšmių skaičių lygų 500 ir mažiausią atžalų mazgus ***Child Node*** sudarančių reikšmių skaičių lygų 200. Tokius skaičius pasirenkame norėdami gauti galimai paprastesnį, apibendrintą sprendimų medį.
- Spragtelime dialogo langelio ***Decision Tree*** mygtuką ***OK***.

Sudarytas pagal CHAID metodą sprendimų medis, esant praleistų nepriklausomų kintamųjų reikšmių, parodytas 6.25 pav. Šiame sprendimų medžio modelyje praleistos reikšmės priskirtos pirmam (galiniam) mazgui kartu su pajamų kategorija “mažos”. Pagal lenteles *Risk* ir *Classification* (6.26 pav.) bendras klasifikacijos tikslumas sudaro 78,2 %, tačiau kategorijos “blogas” klasifikacijos tikslumas 72,9 % laikytinas nepakankamu.



6.25 pav. Sudarytas pagal CHAID metodą sprendimų medis, esant praleistų nepriklausomų kintamųjų reikšmių

Risk

Estimate	Std. Error
,218	,008

Growing Method: CHAID
Dependent Variable: Credit rating

Classification

Observed	Predicted		
	Bad	Good	Percent Correct
Bad	744	276	72,9%
Good	262	1182	81,9%
Overall Percentage	40,8%	59,2%	78,2%

Growing Method: CHAID
Dependent Variable: Credit rating

6.26 pav. Sudaryto pagal CHAID metodą sprendimų medžio lentelės Risk ir Classification

Supaprastintam sprendimų medžio variantui pagal CRT metodą:

- Nurodome komandas ***Analyze → Classify →Tree...***
- Priklausomą kintamąjį *Credit rating* įkeliame į dialogo langelio ***Decision Tree*** (6.1 pav.). laukelį ***Dependent Variable***, o nepriklausomus kintamuosius – į laukelį ***Independent Variables***.
- Llaukelyje ***Growing Method*** nurodome CRT metodą.
- Paliekame dialogo langelio ***Decision Tree: Criteria...*** kortelės ***Growth Limits*** (6.4 pav.) laukelyje ***Minimum Number of Cases*** nustatytą mažiausią pagrindinius (motininius) mazgus ***Parent Node*** sudarančių reikšmių skaičių lygų 500 ir mažiausią atžalų mazgus ***Child Node*** sudarančių reikšmių skaičių lygų 200.
- Dialogo langelio ***Decision Tree: Criteria*** kortelėje ***Surrogates*** (6.10 pav.) paliekame nustatytąjį variantą ***Automatic (one fewer than the number of independent variables).***
- Dialogo langelio ***Decision Tree: Output....***kortelėje ***Statistics*** (6.17 pav.) pažymime laukelį ***Surrogates by split***.
- Spragtelime dialogo langelio ***Decision Tree*** mygtuką ***OK***.

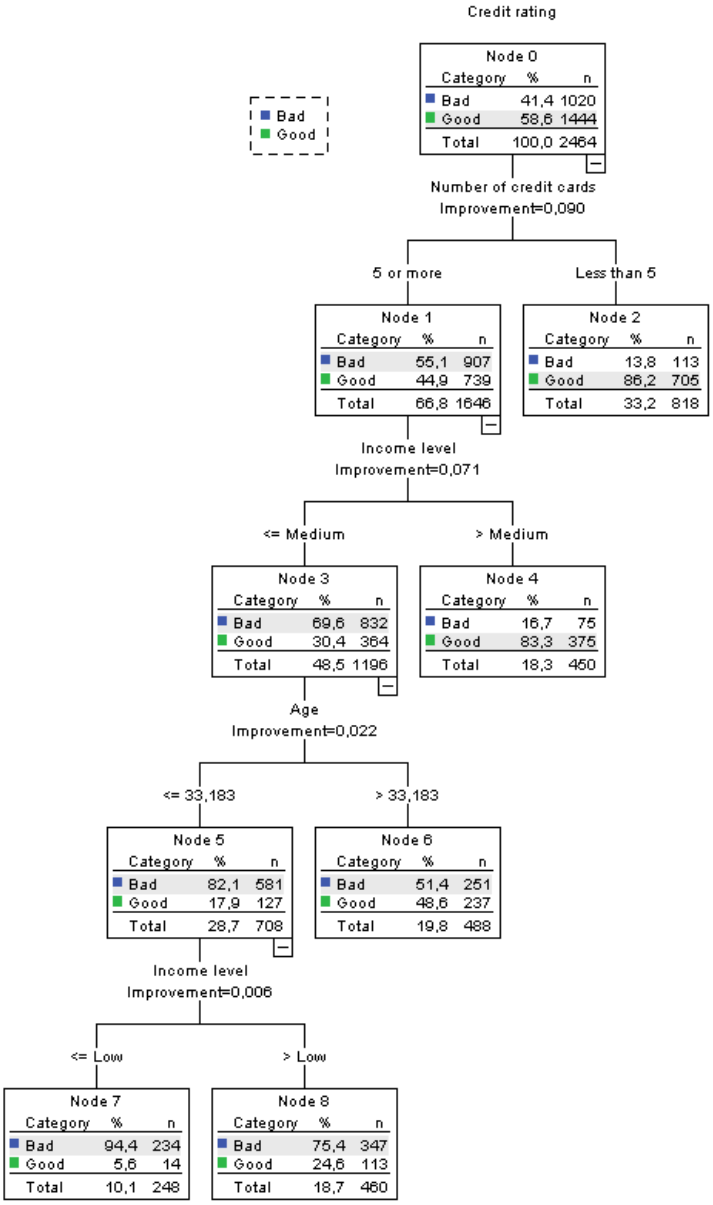
Sudarytas pagal CRT metodą sprendimų medis, esant praleistų nepriklausomų kintamųjų reikšmių, parodytas 6.27 pav. Kaip matome, šis sprendimų medis savo išvaizda esminiai skiriasi nuo sudaryto pagal CHAID metodą sprendimų medžio. Pagal CRT modelį visi mazgo skėlimai yra binariniai, t. y. kiekvienas pagrindinis mazgas skeliamas tik į du atžalų mazgus. Nežiūrint skirtingos išvaizdos abu sprendimų medžiai atspindi tą patį modelį. Tačiau yra ir tam tikrų skirtumų:

- Didžiausią įtaką priklausomam kintamajam turintis nepriklausomas kintamasis pagal CRT metodą yra kredito kortelių skaičius – *Number of credit cards*, kai tuo tarpu pagal CHAID metodą didžiausią įtaką priklausomam kintamajam turi nepriklausomas kintamasis pajamos – *Income level*.
- Pagal CRT metodą nepriklausomas kintamasis pajamos – *Income level* yra antras pagal įtaką priklausomam kintamajam.
- Nėra mazgų, turinčių praleistų reikšmių kategoriją, nes CRT metodas naudoja surogatinius kintamuosius.

Pagal lenteles *Risk* ir *Classification* (6.28 pav.) bendras klasifikacijos tikslumas pasikeitė nežymiai ir sudaro 77,6 %, tačiau pastebimai pagerėjo kategorijos “blogas” klasifikacijos tikslumas 81,6 %, nors kategorijos “geras” klasifikacijos tikslumas pablogėjo nuo 81,9 % iki 74,8 %.

Lentelėje *Surrogates* (6.28 pav.) nurodyti modelio sudaryme panaudoti surogatiniai duomenys. Iš lentelės seka, kad:

- Išeities mazge (*root node*) didžiausią įtaką priklausomam kintamajam turi kredito kortelių skaičius – *Number of credit cards*. Vietoje praleistų kintamojo *Number of credit cards* reikšmių kaip surogatinis kintamasis yra naudojamas nuomojamų automobilių skaičius – *Car loans*, turintis pakankamai stiprų koreliacinį ryšį (0,643) su kintamuoju *Number of credit cards*.
- Tuo atveju, kai ir kintamojo *Car loans* reikšmės yra praleistos, kaip surogatinis kintamasis yra naudojamas amžius – *Age*, nors jo koreliacinis ryšys su kintamuoju *Number of credit cards* yra silpnas (0,04).
- Kintamasis *Age* naudojamas kaip surogatinis taip pat mazguose 1 ir 5.



6.27 pav. Sudarytas pagal CRT metodą sprendimų medis, esant praleistų nepriklausomų kintamųjų reikšmių

Risk

Estimate	Std. Error
,224	,008

Growing Method: CRT
Dependent Variable: Credit rating

Classification

Observed	Predicted		
	Bad	Good	Percent Correct
Bad	832	188	81,6%
Good	364	1080	74,8%
Overall Percentage	48,5%	51,5%	77,6%

Growing Method: CRT
Dependent Variable: Credit rating

Surrogates

Parent Node	Independent Variable		Improvement	Association
0	Primary	Number of credit cards	,090	
	Surrogate	Car loans	,052	,643
		Age	,001	,004
1	Primary	Income level	,071	
	Surrogate	Age	,001	,004
3	Primary	Age	,022	
5	Primary	Income level	,006	
	Surrogate	Age	3,929E-5	,009

Growing Method: CRT
Dependent Variable: Credit rating

6.28 pav. Sudaryto pagal CRT metodą sprendimų medžio lentelės Risk, Classification ir Surrogates

DALYKINĖ RODYKLĖ

LITERATŪRA

1. Norušis M.-J. SPSS14.0 Statistical Procedures Companion: Prentice Hall Inc., 2005.
2. Norušis M.-J. SPSS 14.0 Advanced Statistical Procedures Companion: Prentice Hall Inc., 2005.
3. Bryman A., Cramer D. *Quantitative Data Analysis with SPSS 12 and 13: A Guide for Social Scientists*. Routledge, 2005.
4. Čekanavičius V., Murauskas G. *Statistika ir jos taikymai 1*. Vilnius: TEV, 2000.
5. Čekanavičius V., Murauskas G. *Statistika ir jos taikymai 2*. Vilnius: TEV, 2002.
6. Gonestas E., Strielčiūnas R. R. *Taikomoji statistika*. Vadovėlis kūno kultūros ir sporto specialybių studentams. Kaunas: LKKA, 2003.
7. Sakalauskas V. *Duomenų analizė su STATISTICA*. Vilnius: Margi raštai, 2003.
8. R. A. Yaffee**.** *Common Correlation and Reliability Analysis with SPSS for Windows*. [elektroninis išteklius] http://www.nyu.edu/its/statistics/Docs/correlate.html
9. Dattalo P. *Determining Sample Size: Balancing Power, Precision, and Practicality.* Oxford University Press, 2008.
10. *Designing, Conducting and Analysing Surveys and Questionnaires (ASQ.)* [elektroninis išteklius] www.library.nhs.uk/nlhdocs/FOLIO13_choosing_a_sample.doc
11. Garson D. *Chi-Square Significance Tests* [elektroninis išteklius, atnaujintas 2009 m.] http://faculty.chass.ncsu.edu/garson/PA765/chisq.htm
12. Garson D. *Ordinal Association: Gamma, Kendall's tau-b and tau-c, Somers' d* [elektroninis išteklius, atnaujintas 2009 m.] http://faculty.chass.ncsu.edu/garson/PA765/assocordinal.htm#taub
13. Garson D. *Reliability Analysis* [elektroninis išteklius, atnaujintas 2009 m.] http://www2.chass.ncsu.edu/garson/pa765/reliab.htm#concepts
14. Garson D. *Factor Analysis* [elektroninis išteklius, atnaujintas 2009 m.] http://www2.chass.ncsu.edu/garson/pa765/factor.htm
15. Garson D. *Logistic Regression* [elektroninis išteklius, atnaujintas 2009 m.] http://faculty.chass.ncsu.edu/garson/PA765/logistic.htm