TARTU ÜLIKOOL

Ene-Margit Tiit, Martin Viil

ANDMEANALÜÜS PERSONAALARVUTIL PROGRAMMIPAKI STATGRAPHICS ABIL

II osa

TARTU 1992

TARTU ÜLIKOOL

Ene-Margit Tiit, Martin Viil

ANDMEANALÜÜS PERSONAALARVUTIL PROGRAMMIPAKI STATGRAPHICS ABIL

II osa

STATISTIKAPROTSEDUURID

Tartu 1992

Kinnitatud matemaatikateaduskonna nõukogus
1. novembril 1991.a.

Ene-Margit T i i t, Martin V i i l.
ANDMEANALÜÜS PERSONAALARVUTIL PROGRAMMIPAKI
STATGRAPHICS ABIL. II osa. Statistikaprotseduurid.
Eesti keeles.
Tartu Ülikool.
EV, 202400 Tartu, Ülikooli, 18.
Vastutav toimetaja A.-M. Parring.
19,30.19,10.20,75. T. 568. 500.
Hind rbl. 17.-
TÜ trükikoda. EV, 202400 Tartu, Tiigi, 78.

SISUKORD

SAATEKS

Käesolev õppevahend on jätkuks õppevahendile "Andmeanalüüs personaalarvutil programmipaki STATGRAPHICS abil I" ning tugineb sellele oluliselt. Kui õppevahendi esimeses osas tutvustati programmipaki *STATGRAPHICS* (edaspidi SG) üldist ülesehitust, põhiprintsiipe, samuti selle abil teostatavaid andmeteisendusi, siis käesolev II osa on pühendatud põhilistele statistilise analüüsi protseduuridele. Samuti kui I osas ei lähtu me aine esituses ja järjestuses mitte süsteemi SG loogilisest struktuurist, vaid kasutame andmeanalüüsile omast lähenemisviisi ja materjali järjestust.

Oleme tänulikud oma nõudlikule ja heatahtlikule retsensendile Anne-Mai Parringule rohkete täpsustuste, paranduste ja soovituste eest. Avaldame tänu oma noortele kolleegidele Anneli Bachausile ja Tarmo Kollile operatiivse abi eest raamatus sisalduvate graafikute trükkimisel ja korrektuuri lugemisel, samuti Lyvia Jõesaarele tehnilisel tööl nähtud vaeva eest.

5. MATEMAATILISE STATISTIKA PÕHIPROTSEDUURID

§ 5.1. Matemaatilise statistika põhimõisted

Andmeanalüüsi teostamisel tuginetakse tavaliselt matemaatilise statistika metoodikale, mis lähtub järgmistest eeldustest (vt [7], § 1.2, 1.3):

- meid huvitavat nähtust või protsessi iseloomustab uuritavate tunnuste jaotus üldkogumis;
- meile kättesaadav informatsioon nende tunnuste kohta on antud valimiga, kusjuures me eeldame, et valim on esindav;
- meie ülesandeks on teha järeldusi üldkogumi kohta, kasutades selleks valimis sisalduvat informatsiooni;
- tehtud järelduste usaldatavust iseloomustatakse kvantitatiivselt.

Kõik üldkogumi kohta tehtavad järeldused sisaldavad väiteid

- üldkogumi jaotuse,
- üldkogumi jaotust iseloomustavate arvkarakteristikute,
- üldkogumi jaotust kirjeldavate mudelite,
- nende mudelite arvparameetrite kohta.

Järelduste võimalikeks vormideks on:

- arvkarakteristiku või mudeli parameetri hinnang (st sellele teatud mõttes optimaalse väärtuse omistamine);
- arvkarakteristiku või mudeli parameetri vahemikhinnang (st piirkonna leidmine, milles see etteantud tõenäosusega paikneb);
- hüpoteeside kontrollimine arvkarakteristikute või mudeli parameetrite kohta;
- hüpoteeside kontrollimine jaotuste kohta.

Järelduste usaldusväärsuse kirjeldamiseks hindamisel kasutatakse järgmisi mõisteid:

- olulisuse nivoo, s.o (varem kokku lepitud) maksimaalne lubatav eksimise tõenäosus tõestatava hüpoteesi vastuvõtmisel;

- usaldusnivoo, s.o varem kokku lepitud tõenäosus, millega vahemikhinnang katab hinnatava parameetri õige väärtuse;

- olulisuse tõenäosus, s.o näitaja, mis iseloomustab andmestiku kooskõla nullhüpoteesiga; kui olulisuse tõenäosus on küllalt väike (mitte suurem kui olulisuse nivoo), siis võib nullhüpoteesi kummutatuks lugeda.

Käesolevas paragrahvis esitame ülalmärgitud põhimõistete kohta kõige olulisema teabe. Asjasse süveneda soovijail soovitame pöörduda raamatu [10] ja õppevahendite [3, 4, 6, 8, 9] poole.

Järgmistes peatükkides esitame statistika meetodite kirjelduse ning neid realiseerivate programmide üldiseloomustused teemade kaupa.

§ 5.2. Parameetrite hindamine

5.2.1. Hinnang. Niihästi jaotusi kui ka mudeleid iseloomustavate parameetrite (ka arvkarakteristikute) hindamiseks valimi põhjal kasutatakse mitmesuguseid valimi elementide funktsioone ehk hinnangfunktsioone. Iga konkreetse valimi korral omandab selline funktsioon konkreetse väärtuse, mida nimetatakse (konkreetseks) hinnanguks. Juhime tähelepanu sellele, et samast üldkogumist on võimalik saada palju erinevaid valimeid (mis kõik on esindavad) ning nende valimite põhjal saadud hinnangud ühele ja samale parameetrile on üldiselt erinevad. Seega: teoreetiliselt on hinnang juhuslik suurus (või vektor), mille jaotus sõltub üldkogumi jaotusest. Käesolevas peatükis käsitleme me hinnangut kui juhuslikku suurust (vektorit), edaspidi aga huvitab meid enamasti hinnangu väärtus, mis on saadud konkreetse valimi puhul - nn konkreetne hinnang, mida me lihtsuse mõttes samuti hinnanguks hakkame nimetama.

Ühe ja sama valimi põhjal on ühele ja samale parameetrile erinevate hinnangfunktsioonide abil võimalik saada erinevaid hinnanguid. Matemaatilises statistikas uuritakse põhjalikult hinnangute omadusi, kuid käesolevas õppevahendis pöörame me tähelepanu ainult ühele hinnangu omadusele, nimelt nihketusele.

5.2.2. Hinnangu nihe. Me ütleme, et hinnang on nihketa, kui ta süstemaatiliselt ei üle- ega alahinda hinnatavat parameetrit.

Matemaatiliselt on nihketuse nõue sõnastatav järgmiselt:

Olgu ϑ hinnatav parameeter, mille õiget väärtust (mida tähistagu samuti ϑ) me ei tea. Olgu $\mathcal{T}$ hinnang selle parameetri ϑ jaoks. Hinnangu väärtust konkreetse valimi korral tähistame sümbolitega $\hat{\vartheta}$, $\bar{\vartheta}$ jne. Sümboleid ϑ_0, ϑ_1 jne kasutame edaspidi parameetri ϑ mingite etteantud väärtuste jaoks.

Hinnang $\mathcal{T}$ on nihketa, kui hinnangu $\mathcal{T}$ keskväärtus langeb ühte hinnatava parameetri õige väärtusega ϑ, st

$$E\mathcal{T} = \vartheta, \tag{5.1}$$

kus E on keskväärtuse võtmise tehte sümbol.

Kui seos (5.1) ei kehti, siis sisaldab hinnang süstemaatilist viga, kusjuures juhul $E\mathcal{T} < \vartheta$ hinnang alahindab, juhul $E\mathcal{T} > \vartheta$ aga ülehindab hinnatavat parameetrit. Vahet $E\mathcal{T} - \vartheta$ nimetatakse hinnangu nihkeks.

Praktiliseks kasutamiseks sobivad tavaliselt kõige paremini nihketa hinnangud. Hinnangu absoluutväärtuselt suur nihe võib põhjustada olulisi valejäreldusi selle hinnangu kasutamisel.

5.2.3. Hinnangu hajuvus ja efektiivsus. Et hinnang $\mathcal{T}$ on juhuslik suurus, siis iseloomustab teda ka suurem või väiksem hajuvus. Selle suurus sõltub järgmistest teguritest:

- mõõdetava(te) tunnus(t)e jaotusest üldkogumis,
- hinnatavast parameetrist,
- kasutatavast hinnangfunktsioonist,
- valimi mahust.

Loomulikult on hindamine seda tõhusam, mida väiksem on hinnangu hajuvus; liialt suure hajuvusega hinnangute põhjal osutub sisukate järelduste vastuvõtmine koguni võimatuks. Hinnangu hajuvuse mõõduks kasutatakse tavaliselt selle standardhälvet $\sqrt{D\mathcal{T}}$ (siin $D\mathcal{T}$ on hinnangu dispersioon).

Hinnangu hajuvuse vähendamiseks ei ole üldiselt võimalik muuta üldkogumi jaotust ega ka hinnatavat parameetrit. Küll aga on otstarbekas valida selline hinnangfunktsioon, mis annab võimalikult väikese hajuvusega hinnangu. Märgime, et nihketa ja võimalikult väikese hajuvusega hinnanguid nimetatakse efektiivseteks, ning alati, kui vähegi võimalik, kasutatakse praktilises töös selliseid hinnanguid.

Kõige lihtsam (aga enamasti mitte kõige odavam!) on hinnangu hajuvust vähendada valimi suurendamise teel. Siin toimib üldjuhul nn rusikareegel: valimi mahu suurendamisel n korda väheneb hinnangu hajuvus $\sqrt{n}$ korda, st näiteks hajuvuse kahekordseks vähendamiseks tuleb valimi mahtu neli korda suurendada.

§ 5.3. Vahemikhindamine

5.3.1. Usalduspiirkond. Nagu eelmises paragrahvis märgitud, on hinnang $\mathcal{T}$ juhuslik ja seetõttu ei lange tema konkreetne väärtus tavaliselt ühte hinnatava parameetri õige väärtusega ϑ. Siit tuleneb vahemikhindamise ülesande püstitus: leida (valimi põhjal) selline parameetri võimalike väärtuste piirkond $\mathcal{U}$, milles hinnatava parameetri õige väärtus ϑ sisalduks küllalt suure tõenäosusega.

Seda uurija poolt ette antavat "küllalt suurt" tõenäosust nimetatakse usaldusnivooks, ning tavaliselt on selle väärtuseks 0,95 või 0,99 (öeldakse ka 95 % või 99 %). Piirkonda $\mathcal{U}$ nimetatakse usalduspiirkonnaks, kusjuures sageli näidatakse ära ka usaldusnivoo, kõneldes näiteks 95 % usalduspiirkonnast.

Traditsiooniliselt kasutatakse usaldusnivoo tähisena sümbolit $1-\alpha$ (kus α on vastavalt 0,05 või 0,01).

Matemaatiliselt defineeritakse usalduspiirkond $\mathcal{U}$ tingimusega

$$P(\vartheta \in \mathcal{U}) \geq 1 - \alpha. \qquad (5.2)$$

Arusaadavalt ei määra tingimus (5.2) piirkonda $\mathcal{U}$ üheselt; kasutamiseks on soodsaim võimalikult väikese ulatuse (pikkuse, pindala, mahuga) usalduspiirkond, ning enamasti taotletaksegi sellise leidmist.

5.3.2. Usalduspiirid. Kui hinnatav parameeter paikneb arvsirgel, siis on usalduspiirkonnaks (lahtine või kinnine) vahemik ehk intervall. Sel juhul nimetatakse piirkonda $\mathcal{U}$ usaldusvahemikuks ehk usaldusintervalliks. Usaldusvahemikke on kolme tüüpi:

1) lõplik,
2) alt tõkestamata,
3) ülalt tõkestamata.

Lõplikku usaldusvahemikku piiravad alt ja ülalt usalduspiirid $\underline{u}$ ja $\bar{u}$. Usalduspiirkonda tähistab sümbol

$$\mathcal{U} = \langle \underline{u}, \bar{u} \rangle,$$

kus sulud $\langle\rangle$ võidakse asendada nurk- või ümarsulgudega vastavalt sellele, kas vahemiku otspunktid loetakse sellesse kuuluvaks või mitte, vt ka [4].

Punkte $\underline{u}$ ja $\bar{u}$ nimetatakse vastavalt alumiseks ja ülemiseks usalduspiiriks.

Alt tõkestamata usalduspiirkonna puhul loetakse $\underline{u} = -\infty$ ja $\bar{u}$ nimetatakse ühepoolseks ülemiseks usalduspiiriks. Ülalt tõkestamata usalduspiirkonna puhul $\bar{u} = \infty$ ja $\underline{u}$ on ühepoolne alumine usalduspiir.

§ 5.4. Statistiliste hüpoteeside kontrollimine parameetrite kohta

5.4.1. Statistilise hüpoteesi mõiste. Statistiline hüpotees sisaldab teatavat väidet üldkogumi kohta. Käesolevas paragrahvis vaatleme üldkogumi jaotuse või ka mudeli parameetrite kohta sõnastatud väiteid. Väited sõnastatakse üld-

juhul võrratus-, võrdsus- ja mittevõrdsussuhete abil. Kontrolliniseks esitatakse alati kaks teineteist välistavat hüpoteesi, millest parajasti üks saab olla tõene.

Hüpotees, mida uurija soovib tõestada, kannab sisuka hüpoteesi (ka alternatiivi) nimetust ja see tähistatakse tavaliselt sümboliga H_1. Sisuka hüpoteesi vastandväide on nullhüpotees, mille tähiseks on H_0.

Statistilises mõttes ei ole sisukas hüpotees ja nullhüpotees samaväärsed - statistika meetodid võimaldavad tõestada ainult sisukat hüpoteesi. Nullhüpoteesi ei ole võimalik tõestada. Kui tekib vajadus tõestada väidet, mis on nullhüpoteesi sisuks, tuleb (kui see on võimalik) sõnastada uus hüpoteesipaar.

5.4.2. Vead hüpoteeside kontrollimisel. Olulisuse nivoo ja olulisustõenäosus. Viga, mis tekib siis, kui ekslikult võetakse vastu sisukas hüpotees H_1, loetakse raskeks (nn I liiki viga). Sellise vea esinemise tõenäosus tõkestatakse olulisuse nivooga, mille väärtus valitakse vabalt vastavalt lahendatavale ülesandele. Olulisuse nivood tähistatakse tähega α. Traditsioonilise α väärtuse 0,05 kõrval (mis tähendab, et sagedasemat eksimist kui keskmiselt ühel juhul kahekümnest ei lubata) kasutatakse näiteks meditsiiniuuringutes tihti α väärtust 0,01 (vastu võetud hüpotees osutub juhuslikkuse tagajärjel ekslikuks mitte sagedamini kui ühel juhul sajast).

Hüpoteeside kohta otsuse langetamiseks arvutatakse valimi põhjal välja sobivalt valitud test-statatistiku väärtus (selleks on näiteks hii-ruut χ^2, F jne). Selle statistiku jaotus nullhüpoteesile vastaval eeldusel on teada ning selle põhjal on võimalik arvutada väikseim olulisuse nivoo, mille korral saaks antud valimi jaoks sisuka hüpoteesi vastu võtta. Seda väikseimat olulisuse nivood nimetatakse olulisuse tõenäosuseks ja tähistatakse tähega p (SG: sõna *significance*).

Sisuka hüpoteesi saab tõestatuks lugeda (olulisuse nivooga α) parajasti siis, kui olulisuse tõenäosus pole suurem kui olulisuse nivoo:

$$p \leq \alpha.$$

Kui aga kehtib vastupidine võrratus ($p > \alpha$), siis võetakse vastu nullhüpotees H_0. Pange tähele, et see toimub ka olukorras, kus olulisuse tõenäosus on üsna väike - ainult pisut suurem kui α, näiteks 0,06. Sellepärast ei saagi nullhüpoteesi vastuvõtmist lugeda selle tõestamiseks, vaid see väljendab asjaolu, et antud valimi puhul ei ole võimalik sisukat hüpoteesi tõestada.

5.4.3. Hüpoteeside tüübid. Kui uuritava nähtuse kohta on olemas teatav eelinformatsioon, siis sõnastatakse hüpoteesid sageli ühepoolsetena, näiteks

$$H_1: \vartheta > \vartheta_0,$$

$$H_0: \vartheta \leq \vartheta_0,$$

kus ϑ_0 on mingi ette antud arvuline väärtus. Näiteks võib olla eesmärgiks tõestada, et Eestis 1970. aastal sündinud noormeeste keskmine kasv on üle 180 cm.

Analoogiliselt saab sõnastada ka vastupidiste märkidega hüpoteesipaari.

Teine tüüpiline hüpoteesipaar on järgmine:

$$H_1: \vartheta \neq \vartheta_0,$$

$$H_0: \vartheta = \vartheta_0.$$

Siin on eesmärgiks tõestada, et uuritav parameeter erineb mingist antud väärtusest ϑ_0. Tihti on siin $\vartheta_0 = 0$, niisugune on probleemiseade näiteks siis, kui ϑ iseloomustab mingit muutust ja püütakse tõestada muutuse olemasolu, ilma et selle suund (positiivne või negatiivne) oleks teada.

Vastupidist hüpoteesipaari, mille puhul sisukaks hüpoteesiks oleks väide $\vartheta = \vartheta_0$, ei saa standardse metoodika abil tõestada. See on ka loogiliselt mõistetav, sest valim üldiselt erineb üldkogumist ning seetõttu ei saa valimi põhjal väita, et üldkogumi parameetri väärtuseks on täpselt ϑ_0.

§ 5.5. Jaotuse hindamine ning empiirilise ja teoreetilise jaotuse kooskõla kontrollimine

5.5.1. Empiiriline jaotus. Iga valim defineerib alati empiirilise jaotuse $P_n(X)$ eeskirjaga

$$P(X = x_i) = \frac{1}{n}, \quad i = 1, \ldots, n, \tag{5.3}$$

mille sisuks on üksikute valimi punktide (vaatluste) samaväärsus ja võrdtõenäosus. Valemiga (5.3) määratud jaotust saab käsitleda diskreetse jaotusena ning kirjutada selle jaoks tavalisel viisil välja tõenäosus- ja jaotusfunktsiooni.

Empiirilist jaotust iseloomustavad arvkarakteristikud (mille kohta öeldakse kas empiiriline või valimkarakteristik) sobivad enamasti hästi vastavate üldkogumi karakteristikute hinnanguks.

Et teha vahet empiiriliste ja üldkogumi (öeldakse ka: teoreetiliste) karakteristikute vahel, kasutatakse tihti sümboolikas paralleelselt ladina ja kreeka tähti. Nii näiteks tähistatakse üldkogumi momente sümboliga μ_n, vastavaid valimmomente aga sümboliga m_n jne.

5.5.2. Empiirilise jaotuse võrdlemine teoreetilise jaotusega. Tihti esitatakse küsimus - kas uuritava tunnuse jaotuseks üldkogumis saab olla mingi tuntud jaotus P? Sellist ülesannet nimetatakse (teatud mõttes tinglikult) ka jaotuse sobitamise ülesandeks või empiirilise ja teoreetilise jaotuse võrdlemise ülesandeks.

Ilmselt on siin tegemist hüpoteeside kontrollimise ülesandega, kusjuures kontrollimist vajav hüpoteesipaar on järgmine (P_X tähistab uuritava tunnuse jaotust):

$$\begin{aligned} H_1&: P_X \neq P \\ H_0&: P_X = P \end{aligned} \tag{5.4}$$

Täpsustamist vajab veel teoreetiline jaotus P. Tavaliselt on tegemist parameetrilise jaotuste perega, milles üksikjaotus identifitseeritakse parameetrite järgi. Seega on meil põhimõtteliselt võimalikud kaks ülesannet:

1. Mingil viisil on teada teoreetilise jaotuse parameetri ϑ väärtus ϑ_0 ning empiirilist jaotust soovitakse võrrelda (sobitada) teoreetilise jaotusega P_{ϑ_0}

2. Teoreetilise jaotuse parameetri ϑ kohta ei ole mingit eelinformatsiooni. Lahendada tuleb kaks osaülesannet:

a) leida valimi põhjal hinnang $\tilde{\vartheta}$ teoreetilise jaotuse parameetrile ϑ;

b) võrrelda (sobitada) empiirilist jaotust teoreetilise jaotusega $P_{\tilde{\vartheta}}$.

Märkigem, et valimi põhjal leitud parim parameetri väärtus $\tilde{\vartheta}$ ei taga alati empiirilise jaotusega $P_n(X)$ kõige paremas kooskõlas oleva teoreetilise jaotuse $P_{\tilde{\vartheta}}$ leidmist, sest parameetrite hindamisel ja jaotuste võrdlemisel kasutatakse üldiselt erinevaid kooskõlakriteeriume; enamasti pole aga tekkinud erinevus tähtis.

Hüpoteesipaari kontrollimine toimub standardsel viisil. Valimi põhjal arvutatakse välja teststatistik (see on tavaliselt χ^2 või Kolmogorov-Smirnovi Δ), selle väärtus määrab olulisuse tõenäosuse p, s.o nullhüpoteesi (jaotuste kooskõla) tõenäosuse.

Kui $p \leqslant \alpha$, siis loetakse H_1 tõestatuks.

Kui $p > \alpha$, siis võetakse vastu H_0.

5.5.3. Jaotuste kooskõla interpreteerimine. Olgu antud rida teoreetilisi jaotusi P_1, P_2, ..., P_h ja empiiriline jaotus $P_n(X)$. Empiirilist jaotust võrreldakse igaühega antud teoreetiliste jaotuste seast. On võimalikud järgmised tulemused:

1) kõigi jaotuste korral tõestatakse hüpotees H_1;

2) leidub üks jaotus P_i, mille korral võetakse vastu H_0, kõigi ülejäänud jaotuste korral kehtib tulemus H_1;

3) leidub mitu jaotust (võib-olla isegi kõik), mille puhul võetakse vastu nullhüpotees H_0.

Kuidas tõlgendada saadavaid tulemusi?

Kõigepealt paneme tähele, et tulemused sõltuvad α väärtusest. Muutes α väiksemaks, raskendame me H_1 vastuvõtmist ning võib juhtuda, et olukorrast 1) või 2) satume olukorda 3).

Muutes aga α suuremaks, hõlbustame me H_1 vastuvõtmist ning tulemuseks võib olla olukord 1).

Väga oluline on ka valimi mahu n mõju.

Kui valimi maht n on suhteliselt väike, on tüüpiline olukord 3) - vastu võetakse nullhüpoteesid.

Suurte ja ülisuurte valimite korral aga on tüüpiline olukord 1) - empiiriline jaotus ei ole ühegi teoreetilise jaotusega kooskõlas. Eriti viimane asjaolu tekitab uurijates vahetevahel segadust.

Kirjeldatud tulemustest aru saamiseks paneme tähele järgmist: tõestada saab üksnes jaotuste sobimatust, mitte aga sobivust.

Tulemus 1) tähendab seda, et ükski antud teoreetilistest jaotustest ei sobi vaadeldava empiirilise jaotuse kirjeldamiseks. Teisisõnu, nendest teoreetilistest jaotustest sellise juhusliku valimi saamine, nagu meil on olemas, on äärmiselt vähe tõenäone (kui mitte võimatu).

Tulemus 2) tähendab seda, et kõigi antud teoreetiliste jaotuste seast ainus, mis vaadeldava empiirilise jaotuse kirjeldamiseks võiks sobida, on P_t. Muidugi on võimalik, et ka see sobib halvasti (näiteks kui olulisuse tõenäosus on kaunis väike), mingil juhul ei saa tulemust tõlgendada kui tõestust, et valim tõepoolest pärineb teoreetilisest jaotusest P_t.

Tulemus 3) ei ole vastuoluline, vaid lihtsalt näitab mitut teoreetilist jaotust, mille puhul olemasoleva valimi saamine on võimalik, s.t mitte liialt vähe tõenäone. Võrreldes omavahel olulisuse tõenäosusi, võib valida välja ka selle jaotuse, mille puhul antud valimi saamise tõenäosus on suurim. Ometi ei saa seda tulemust üle tähtsustada - ka see pole range tõestus, et tunnuse jaotus üldkogumis on nimelt selline.

5.5.4. Mitteparameetriline statistika. Suurema osa statistikaprotseduuride korral on tarvis eeldada, et üldkogumi jaotus on teada, enamasti nõutakse, et üldkogum oleks normaaljaotusega.

Hädavajalik on selline eeldus enamiku hüpoteeside kontrollimise ja vahemikhindamise meetodite puhul.

On aga selge, et kaugeltki alati ei saa eeldada, et tegemist on normaaljaotusega. Et leida universaalne aparatuur ka niisuguste otsustamisülesannete lahendamiseks, kus üldkogumi jaotus ei ole teada (või on põhjust arvata, et jaotus erineb normaalsest), on välja töötatud metoodika, mida nimetatakse mitteparameetriliseks (ka jaotusvabaks).

Võrreldes nn standardsete (parameetriliste) meetoditega on mitteparameetrilised mõnevõrra väiksema võimsusega, s.t nõuavad samade järelduste tegemiseks mõnevõrra (näit 5-10 %) rohkem vaatlusi. Seevastu on mitteparameetrilised hinnangud enamasti stabiilsemad ja ei lase end juhuslike või jämedate vigade toimest mõjutada. Mitteparameetrilised meetodid on üldiselt eelistatavamad ka järjestustunnuste töötlemisel.

Väga sageli tuginevad mitteparameetrilised meetodid järkstatistikutele, s.t variatsioonrea liikmetele või nende järjekorranumbritele ehk astakutele.

Kuigi suurele osale standardsetest statistikaprotseduuridest on välja töötatud ka mitteparameetrilised analoogid, on need üldiselt vähem tuntud ja harvem programmidena realiseeritud. Seetõttu kasutatakse tihti standardseid meetodeid nendelgi puhkudel, kus korrektsem oleks töötada mittestandardse meetodi abil. SG süsteemis on siiski terve rida mitteparameetrilisi meetodeid realiseeritud ning kõigi meetodite kirjeldamisel esitame ka vastavad mitteparameetrilised analoogid.

Märgime, et jaotuste võrdlemise meetodid kuuluvad üldiselt mitteparameetrilisse statistikasse.

6. ÜHE TUNNUSE ANALÜÜS

§ 6.1. Üksiktunnuse analüüs ja SG süsteemi võimalused

6.1.1. Üksiktunnuse analüüsimise eesmärk. Statistilise materjali analüüsimine algab tavapäraselt üksiktunnuste kirjeldamisest ja nende statistilise olemuse selgitamisest.

Üksiktunnuste analüüsi ulatus ja sügavus sõltub tervikprobleemist - mõnikord on tarvis kõik tunnused ühekaupa väga detailselt läbi analüüsida, teinekord piisab suurema osa tunnuste kohta ainult üpris üldjoonelisest kirjeldusest. Üksiktunnuste analüüsimisel peetakse tavaliselt silmas järgmisi eesmärke:

1) esmane tutvumine materjaliga;

2) materjali korrektsuse kontrollimine jämedate vigade osas (mõõtmisel, kodeerimisel, arvutisse sisestamisel tekkinud ebatäpsused).

Üksiktunnuste analüüsimise käigus saadud informatsioon on enamasti tarvilik nende tunnuste edasise töötlemise strateegia ja taktika väljatöötamiseks.

Mõnikord aga annab üksiktunnuste töötlus olulisi tulemusi püstitatud tööülesande seisukohalt, seega pole päris õige käsitleda üksiktunnuste analüüsi ainult tegelikku töötlust sissejuhatava tööetapina.

6.1.2. Üksiktunnuste analüüsi metoodika. Üksiktunnuste analüüsimine võib toimuda erinevatel tasemetel niihästi analüüsi põhjalikkuse ja sügavuse kui ka näitlikustamise poolest. Oluline on märkida, et iga tunnuse analüüsimisel lubatavate meetodite valik sõltub oluliselt selle tunnuse tüübist ja tema suhtes tehtud eeldustest.

Üldiselt võiks üksiktunnuse analüüsi meetodid jaotada järgmiselt:

- tunnuse empiirilise jaotuse kirjeldamine;
- tunnuse jaotusparameetrite ja mitmesuguste karakteristikute hindamine (punkthindamise mõttes);
- tunnuse jaotusparameetrite vahemikhindamine ning statistiliste hüpoteeside kontrollimine jaotusparameetrite kohta;
- tunnuse empiirilise jaotuse võrdlemine teoreetiliste jaotustega, statistiliste hüpoteeside kontrollimine tunnuse (teoreetilise) jaotuse kohta.

Kuivõrd igasugused statistilised mudelid nõuavad vähemalt kahe tunnuse käsitlemist (funktsioon- ja argumenttunnused), siis ühe tunnuse analüüsi meetodite hulgas mudeli parameetrite hindamisega seotud protseduure ei ole.

Mõned statistikaprotseduurid (mis küll põhimõtteliselt kuuluvad ülalnimetatud loetelusse) on välja töötatud "kahtlaste" väärtuste ülesotsimiseks (tuginedes seejuures teatavatele üldistele või konkreetsetele eeldustele uuritava tunnuse kohta) ja nendele tähelepanu juhtimiseks. Niisuguseid protseduure on sobiv kasutada töötluse algetapil kontrolli maks andmestiku korrektsust nn jämedate vigade suhtes.

6.1.3. SG-süsteemi võimalused üksiktunnuse analüüsiks. SG-süsteem sisaldab ligi 20 protseduuri üksiktunnuse analüüsimiseks. Mõned nendest lahendavad isegi mitu eelmises punktis loetletud ülesannet ning pakuvad väljundina mitmesuguseid tabeleid ja ka graafikuid. Mõned seevastu on aga üsna väikese mahuga ning on kasutatavad kaunis kitsastel eeldustel.

Üldinformatsiooni SG protseduuride kohta üksiktunnuste analüüsiks pakub tabel 6.1.

Selgituseks tabeli kohta märgime järgmist:

- lahtrisse on tehtud märk + siis, kui vastavas veerus märgitud asjaolu peab antud reas märgitud protseduuri kohta paika, (+) aga sel juhul, kui ta peab paika tingimisi, osaliselt või kitsendustega;
- tunnuste tüübid vastavad I osa § 1.2 definitsioonidele.

Tabel 6.1

	Jrk nr	Protseduuri nimi	Lubatav tunnuse tüüp					SG-muutuja tüüp			Normaalsus[1]	Mitteparameetriline	Protseduur		Väljund		
			Arv-		Järjestus-		Nominaal-	N, I	C				Hindam.	Hüpot. kontroll	Graafik	Tekst, arvud	Tabel
			Pidev	Diskr.		Binaar-											
ES	1.	Barcharts		+	+	(+)	+	+	+				+		+		
IA	2.	Box-and-Whisker Plot	+	+	+			+				+	+		+		
HA	3.	Critical Values	+	+				+					+			+	
HA	4.	Distribution Fitting	+	+	+			+				+	+	+	+	+	+
F3	5.	Frequency Histogram		+	+	(+)	+	+	+				+		+		
F2	6.	Frequency Tabulation	+	+	(+)		+	+	+				+		+	+	+
GA	7.	Hanging Histobars	+	+	(+)			+			+		+	+	+	+	
KG	8.	Kolmogorov-Smirnov One-Sample Test	+					+				+		+		+	
G3	9.	Normal Probability Plot	+	+				+			+			+	+		
GA	10.	One-Sample Analysis	+	+	(+)			+			+			+		+	
FS	11.	Percentiles	+	+	(+)			+					+			+	
EG	12.	Piecharts		+	+	(+)	+	+	+				+		+		
I7	13.	Stem-and-Leaf Display	+	+	(+)			+					+			+	
FA	14.	Summary Statistics	+	+	(+)	(+)		+					+				+
I6	15.	Suspended Rootogram	+	+				+			+			+	+		
H3	16.	Tail Area Probabilities	+	+	(+)			+					+	+		+	
RA	17.	Tests for Binary Sequences				+		+	+			+		+			
R3	18.	Tests for Location	+	+	+	+		+				+		+			
RA	19.	Tests for Randomness	+	+	+	+		+				+		+			
F4	20.	Weighted Averages	+	+		(+)		+					+			+	

1) Normaaljaotus on eelduseks vahemikhindamisel või hüpoteeside kontrollimisel (nullhüpoteesile vastaval juhul).

Sellesse tabelisse ei ole paigutatud aegridade töötlemiseks ette nähtud protseduure, kuigi ka neid võib käsitleda ühe (ajast sõltuva) tunnuse töötlusprotseduuridena. Selle põhjuseks on asjaolu, et sisuliselt on siin tegemist siiski kahe tunnuse analüüsiga, liiatigi erimetoodika alusel, mida me käesolevas õppevahendi osas ei käsitle.

6.1.4. Ühe tunnuse analüüsimisel kasutatav sümboolika. Kuivõrd käesolevas paragrahvis vaadeldakse korraga ainult üht tunnust, siis tähistame selle alati tähisega X.

Tunnuse mõõdetud väärtused - valimi - tähistame vektoriga $(x_1, x_2, \ldots, x_n)$, kus n on valimi maht. Siin eeldame, et väärtused on nende mõõtmise (arvutisse sisestamise) järjekorras.

Valimi järjestamisel (kui tegemist on arv- või järjestustunnusega, siis tohib seda teha) saadakse variatsioonrida; seda tähistame $x'_1, \ldots, x'_n$. Siin x'_1 on valimi vähim ja x'_n suurim väärtus,

$$x'_1 = \min_{1 \le i \le n} x_i; \quad x'_n = \max_{1 \le i \le n} x_i.$$

Võib muidugi juhtuda, et tunnusel on erinevaid väärtusi vähem kui n, nimelt k. Tähistame need sümbolitega $a_1, a_2, \ldots, a_k$ ($a_i < a_{i+1}$, i=1, ..., k-1). Sel juhul on variatsioonrida kordustega, s.t väärtus a_1 esineb variatsioonreas n_1 korda, väärtus a_2 - n_2 korda jne, kusjuures $n = \sum_{i=1}^{k} n_i$.

§ 6.2. Tunnuse karakteristikud

Tunnuse jaotust iseloomustavad karakteristikud (ka jaotusparameetrid) võime üldjoontes jagada järgmistesse rühmadesse:

- momentide abil arvutatavad karakteristikud, mis on lubatavad üksnes arvtunnuste korral (tüüp A, vt. [7], lk 18-24); teatava reservatsiooniga kasutatakse neid siiski ka järjestustunnuste puhul (tüüp B);

- järkstatistikute abil arvutatavad karakteristikud, mis on lubatavad arv- ja järjestustunnuste korral (tüübid A ja B);

- sageduste abil arvutatavad karakteristikud, mis on lubatavad kõigi tunnuste korral.

Järgnevas kirjeldame kõiki põhilisi karakteristikuid rühmade kaupa. Kuivõrd tegemist on valimi põhjal arvutatud suurustega, on need kõik valimikarakteristikud ehk empiirilised karakteristikud. Lihtsuse mõttes me edaspidi seda ei rõhuta, küll aga märgime iga mõiste esmamainimisel tema korrektse nimetuse.

6.2.1. Momendid ja neist tuletatud karakteristikud. Juhusliku suuruse h-järku moment μ_h (meie käsitluses on h naturaalarv) on defineeritud kui selle juhusliku suuruse h-nda astme keskväärtus:

$$\mu_h = E(X^h).$$

Momentide kõrval kasutatakse ka tunnuse tsentraalseid momente $\bar{\mu}_h$,

$$\bar{\mu}_h = E(X - EX)^h.$$

Valimmoment m_h arvutatakse järgmisest valemist:

$$m_h = \frac{1}{n} \sum_{i=1}^{n} x_i^h.$$

Momentide abil defineeritakse terve rida olulisi tunnust iseloomustavaid arvkarakteristikuid, mida järgnevalt tutvustamegi.

6.2.2. Keskväärtus. Keskväärtus EX ehk esimene moment μ_1 on tunnuse oluline paiknemise karakteristik, ta iseloomustab tunnuse väärtuste paiknemist arvsirgel. Keskväärtuse jaoks kasutatakse ka nimetusi matemaatiline ootus, ooteväärtus (mõnikord ka keskmine).

Keskväärtuse hinnanguks valimi põhjal on keskmine (ka aritmeetiline keskmine). Täpne on kasutada termineid valimkeskmine, empiiriline keskmine, kuid edaspidi me lihtsuse mõttes seda ei tee.

Keskmise sümboliks on $\bar{x}$ ja arvutusvaleмiks

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

s.o esimest järku moment m_1. Diskreetse tunnuse puhul, millel on k väärtust, arvutatakse keskmine valemist

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{k} a_j n_j,$$

kus a_j on tunnuse väärtus, n_j - vastav sagedus.

Keskmine $\bar{x}$ on nihketa ja efektiivseks hinnanguks teoreetilisele keskväärtusele μ_1, sõltumata lähtetunnuse X teoreetilisest jaotusest. Hinnangu $\bar{x}$ puuduseks on aga tema tundlikkus jämedate vigade suhtes: tõepoolest, ülemäära suur või väike tunnuse väärtus rikub $\bar{x}$ üsna palju, eriti siis, kui valimi maht pole eriti suur.

Keskmine $\bar{x}$ ei iseloomusta eriti hästi ka selliseid tunnuseid, mille väärtused on ühes suunas n-ö "välja venitatud", nende puhul on $\bar{x}$ interpretatsioon tülikam.

6.2.3. Dispersioon. Dispersioon DX on defineeritud kui tunnuse teist järku tsentraalne moment:

$$DX = \bar{\mu}_2 = E(X - EX)^2.$$

Dispersioon on tunnuse hajuvuse karakteristikuks. Kuivõrd dispersioon on alati mittenegatiivne, siis kasutatakse tema tähisena ka sümbolit σ^2 (sigma-ruut).

Dispersiooni nihketa hinnanguks valimi põhjal on valimdispersioon (edaspidi dispersioon) s^2, mis arvutatakse vastavalt eeskirjale

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2. \qquad (6.1)$$

Mõnikord kasutatakse ka sellist dispersiooni hinnangut, mis erineb valemiga (6.1) esitatust selle poolest, et teguriks on $\frac{1}{n-1}$ asemel $\frac{1}{n}$. Üldiselt on selline hinnang nihkega ja alahindab tunnuse hajuvust. Suurte valimite korral ($n > 100$) võib kasutada ka nihkega hinnangut.

6.2.4. Standardhälve. Standardhälve σ on defineeritud kui ruutjuur dispersioonist. Standardhälve on samuti kui dispersioongi hajuvuse karakteristik. Tema eeliseks on see, et ta on mõõdetav samades mõõtühikutes mis tunnus X, sellepärast kasutatakse teda praktikas märksa sagedamini kui dispersiooni.

Standardhälbe hinnanguks on s, s.o ruutjuur valemiga 6.1 arvutatud dispersiooni hinnangust; s ei ole küll nihketa hinnang teoreetilisele standardhälbele σ, kuid tema nihe on küllalt väike.

6.2.5. Standardviga on kokkuleppeline nimetus keskväärtuse hinnangu $\bar{x}$ standardhälbele, mis arvutatakse vastavalt alljärgnevale valemile:

$$\sqrt{D(\bar{x})} = \frac{\sigma}{\sqrt{n}} .$$

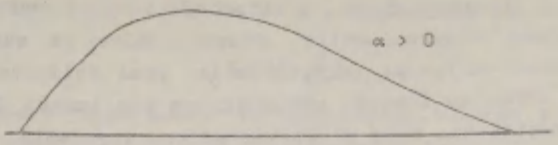
Valimi standardviga tähistatakse tavaliselt tähega m ning tema arvutamiseks kasutatakse standardhälbe hinnangut s:

$$m = \frac{s}{\sqrt{n}} .$$

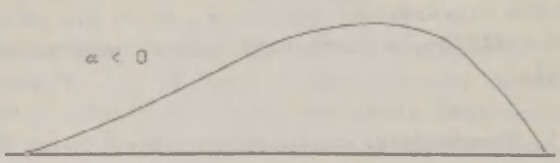
6.2.6. Asümmeetria kordaja iseloomustab tunnuse sümmeetrilisust; asümmeetria kordaja α arvutatakse tunnuse kolmandat järku tsentraalse momendi kaudu vastavalt järgnevale valemile:

$$\alpha = \frac{\bar{\mu}_3}{\sigma^3} .$$

Sümmeetrilise jaotuse korral on alati $\alpha = 0$; vastupidine väide ei ole küll alati õige, kuid üldiselt on jaotused, mille puhul $\alpha = 0$, sümmeetrilistele küllaltki lähedased. Asümmeetria kordaja nullist erinev väärtus iseloomustab jaotuse vasak- või parempoolset asümmeetriat (vt jooniseid 6.1 ja 6.2).



Jn 6.1. Parempoolse asümmeetriaga tunnus



Jn 6.2. Vasakpoolse asümmeetriaga tunnus

Asümmeetriakordaja nihketa hinnanguks on

$$a = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^{3/2}} \cdot \frac{n\sqrt{n-1}}{n-2}.$$

Selle hinnangu standardhälbe hinnanguks s(a) on

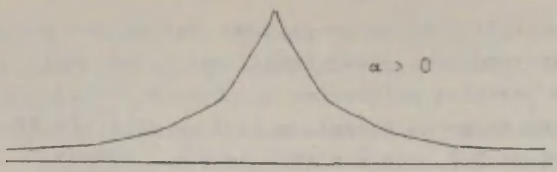
$$s(a) = \sqrt{\frac{6}{n}},$$

suurust $\frac{a}{s(a)}$ nimetatakse standardiseeritud asümmeetria kordajaks.

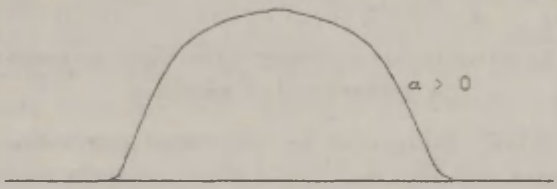
6.2.7. Ekstsess. Ekstsess (ka järsakuse kordaja) ε iseloomustab tunnuse jaotuse kuju ning avaldub neljanda tsentaalse momendi kaudu:

$$\varepsilon = \frac{\bar{\mu}_4}{\sigma^4} - 3.$$

Normaaljaotuse puhul on ekstsess võrdne nulliga. Ekstsess on positiivne siis, kui jaotusel on "rasked sabad" (ja enamasti ka terav tipp), vt jn 6.3. Negatiivse ekstsessiga jaotus on suhteliselt lame, tihti tõkestatud väärtustega ("sabad" puuduvad või on "kerged"), vt jn 6.4.



Jn 6.3. Positiivse ekstsessiga jaotus



Jn 8.4. Negatiivse ekstsessiga jaotus

Ekstsessi nihketa hinnanguks on

$$e = \left[\frac{\sum_{i=1}^{n} (x_i - \bar{x})^4 \cdot n(n+1)}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^2 (n-1)} - 3 \right] \frac{(n-1)^2}{(n-2)(n-3)} .$$

Selle hinnangu standardhälbe asümptootiline hinnang on s(e),

$$s(e) = \sqrt{\frac{n}{24}} ,$$

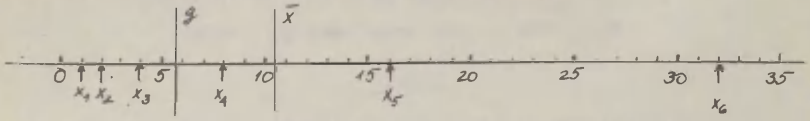
suurust $\frac{e}{s(e)}$ nimetatakse standardiseeritud ekstsessiks.

6.2.8. Geomeetriline keskmine ja kaalutud keskmine. Tinglikult loeme käesolevasse arvkarakteristikute rühma kuuluvaks veel kaks keskmisest tuletatud karakteristikut, mis on samuti kui momendidki arvutatavad üksnes arvtunnuste korral. Tuleb märkida, et need mõlemad on otse valimi jaoks defineeritud karakteristikud ning nende vahekorda üldkogumi analoogidega me käesolevas õppevahendis ei käsitle.

Esimene neist on geomeetriline keskmine g, mis on defineeritud valemiga

$$g = \sqrt[n]{x_1 \cdot \ldots \cdot x_n} .$$

Geomeetriline keskmine on määratud üksnes positiivsete väärtustega tunnuste jaoks. Samuti kui keskminegi on ka geomeetriline keskmine paiknemise karakteristikuks, mis on hästi kasutatav tunnuste korral, millel on märkimisväärne asümmeetria (vt jn 6.5, kus $\bar{x} = 10{,}5$ ja $g = 5{,}65685$).



Jn 6.5. Paarempoolse asümmeetriaga tunnuse keskmine ja geomeetriline keskmine

Mõningates ülesannetes ei ole vaatlused/valimi üksikobjektid samaväärsed, vaid on tarvis omistada erinevatele vaatlustele erinev osatähtsus - kaal. Sel juhul kasutatakse nn kaalutud keskmist $x(w)$,

$$x(w) = \frac{1}{w} \sum_{i=1}^{n} w_i x_i,$$

kus

$$w = \sum_{i=1}^{n} w_i.$$

6.2.9. Järkstatistikutel põhinevad jaotuse karakteristikud. Kui varemnimetatud arvkarakteristikuid on tarvis arvutada, st nende leidmiseks tuleb teha aritmeetilisi tehteid, siis järgnevalt vaatleme niisuguseid näitajaid, mille leidmiseks ei ole tarvis aritmeetilisi tehteid teha, vaid piisab valimi elementide järjestamisest ja loendamisest.

Juhusliku suuruse p-kvantiil on defineeritud kui väärtus ξ_p, mis rahuldab tingimust $P(X < \xi_p) = p$ (vt [7], lk 112). Teoreetilise p-kvantiili ξ_p hinnanguks sobib niisugune valimi väärtus, millest on np elementi väiksemad. Enamasti pole seda võimalik täpselt saavutada ning seetõttu loetakse empiiriliseks p-kvantiiliks niisugune valimi element, mis kõige paremini rahuldab kaht tingimust:

- temast on np valimi elementi väiksemad;
- temast on n(1-p) valimi elementi suuremad.

Niisugust tingimust täidab tavaliselt variatsioonrea element järjekorranumbriga [np] või [np]+1, kus [a] tähistab arvu a täisosa.

Mõningatel juhtudel kasutatakse ka interpolatsioonireegleid, arvutades p-kvantiilide jaoks väärtusi, mis ei lange variatsioonrea väärtustega kokku, vaid paiknevad nende vahel.

Kvantiilidest tähtsaimad on järgmised.

1. 0,5-kvantiil ehk mediaan, tähis med, variatsioonrea keskel paiknev element (paarisarvulise valimi korral arvutatakse tihti kahe keskmise elemendi poolsumma). Mediaan on tunnuse paiknemise karakteristik samuti kui keskminegi.

2. 0,25-kvantiil ja 0,75-kvantiil ehk alumine ja ülemine kvartiil. Kvartiilide vahet kasutatakse (eriti arvtunnuste puhul) tunnuse hajuvuse karakteristikuna samuti kui standardhälvet.

3. 0,95-kvantiil ehk 95 %-punkt.

Järkstatistikuid võib kasutada kõigi arvuliste tunnuste, aga ka järjestustunnuste korral. Nimetatud karakteristikute eeliseks on see, et nad on suhteliselt vähe tundlikud üksikute jämedate vigade suhtes.

Järkstatistikute hulka kuuluvad ka valimi minimaalne ja maksimaalne element, mille tähiseks on lihtsalt min ja max. Minimaalse ja maksimaalse elemendi vahet, nn variatsiooniulatust e haaret kasutatakse vahel ka tunnuse hajuvuse karakteristikuna. Siinjuures tuleb aga meeles pidada, et variatsiooniulatus sõltub valimi mahust, seetõttu ei saa isegi sama tunnuse erinahulisi valimeid variatsiooniulatuse järgi võrrelda.

Kui üldkogumi jaotuseks on normaaljaotus, siis on võimalik (vastavate tabelite abil) variatsiooniulatuse põhjal hinnata standardhälvet, vt [1].

6.2.10. Mood. Mood on see tunnuse väärtus, millel on suurim sagedus.

Moodi saab samuti käsitleda paiknemise karakteristikuna, kusjuures tema kasutamine ei ole tunnuse tüübiga piiratud. Märkigem, et ükski eelpool nimetatud karakteristikutest

ei ole lubatud nominaaltunnuse analüüsimisel, seega on mood ainus universaalne karakteristik. Siiski, pideva tunnuse korral kaotab mood tihti mõtte - suure mõõtmistäpsuse korral võivad kõik valimi elemendid olla erinevate väärtustega, seega mood puudub.

Mõnikord võib mitmel erineval tunnuse väärtusel olla võrdne, ülejäänutest suurem sagedus. Siis loetakse kõiki neid väärtusi moodideks, jaotust nimetatakse kas bimodaalseks (kahe moodi korral) või multimodaalseks (mitme moodi korral). Statistikasüsteemid leiavad tavaliselt siiski ühe moodi, selleks on kokkuleppeliselt näiteks väiksein mood.

§ 6.3. Tunnuse karakteristikute leidmine SG-süsteemi abil

6.3.1. Karakteristikute arvutamine protseduuriga *Summary Statistics*. Protseduur *Summary Statistics* on esimene allmenüüst *F. Descriptive Methods*.

See protseduur on ainulaadne selles mõttes, et ta võimaldab korraga ja üksteisest sõltumatult analüüsida mitut tunnust. Ta on rakendatav arvtunnustele (tüübid N, I).

Protseduuri *Summary Statistics* tellimuspaneel on esitatud joonisel 6.6. Paneeli täitmisel tuleb kõigepealt trükkida (üksteise alla) tunnuste nimed. Seejärel on võimalik töödeldavate tunnuste jaoks tellida 17 karakteristikut, mis on tähistatud tähtedega *A...Q*. Vaikimisi leiab programm nende kõigi väärtused. Arvutatavate statistikute hulga vähendamiseks tuleb mittesoovitavad jadast *Statistics* tühikuklahvi või kustutamisklahvi Del abil kustutada.

Statistikute tähendused on järgmised:

(a) Average - keskmine (vt p 6.2.2);

(b) Median - mediaan (vt p 6.2.9);

(c) Mode - mood (vt p 6.2.10);

(d) Geometric mean - geomeetriline keskmine (vt p 6.2.8);

(e) Variance - dispersioon (vt p 6.2.3);

(f) Std. deviation - standardhälve (vt p 6.2.4);

(g) Std. error - standardviga (vt p 6.2.5);

(h) Minimum - miinimum (vt p 6.2.9);

(i) Maximum - maksimum (vt p 6.2.9);

(j) Range - haare e variatsiooniulatus (vt p 6.2.9);

(k) Lower quartile - alumine kvartiil (vt p 6.2.9);

(l) Upper quartile - ülemine kvartiil (vt p 6.2.9);

(m) Interquar. range - kvartiilide vahe (vt p 6.2.9);

(n) Skewness - asümmeetriakordaja (vt p 6.2.6);

(o) Std. skewness - standardiseeritud asümmeetriakordaja (vt p 6.2.6);

(p) Kurtosis - ekstsess e ekstsessikordaja (vt p 6.2.7).

(q) Std. kurtosis - standardiseeritud ekstsess (vt p 6.2.7).

Summary Statistics

Data vectors: vanus
kaal
sugu

Statistics: ABCDEFGHIJKLMNOPQ

(a) Average	(f) Std. deviation	(k) Lower quartile	(p) Kurtosis
(b) Median	(g) Std. error	(l) Upper quartile	(q) Std. kurtc
(c) Mode	(h) Minimum	(m) Interquar. range	
(d) Geometric mean	(i) Maximum	(n) Skewness	
(e) Variance	(j) Range	(o) Std. skewness	

Jn 6.6

Teisi muudetavaid parameetreid tellimuspaneel ei sisalda, samuti puuduvad valitavad režiimid ja täiendav informatsioon (*options*).

Protseduur käivitatakse tavalisel viisil klahviga **F6**. Joonisel 6.6 esitatud tellimuse tulemus on esitatud joonisel 6.7. Näeme, et lisaks tellitud statistikutele on lisatud veel ka tunnuse valimi maht (*Sample size*).

Vaadates esimest arvudeveergu tabelis (tunnuse vanus arvkarakteristikud), näeme, et uuritavas kogumis on 24-53-aastaseid inimesi, neist pooled on kuni 28-aastased ja kõige arvukam on 27-aastaste rühm; standardhälbe suuruseks on 7,35; poolte vastanute vanus on 26,5 ja 35,5 aasta vahel. Standardiseeritud asümmeetriakordaja väärtus 3,96, mis on kriitilisest väärtusest 2 märksa suurem, näitab, et tunnuse jaotus on oluliselt ebasümmeetriline. Kuivõrd asümmeetria on positiivne, siis on vanuse jaotus nimelt suurte väärtuste suunas "välja venitatud". Samuti tuleb standardiseeritud ekstsessikordaja põhjal järeldada, et jaotuskõvera kuju erineb oluliselt normaaljaotusest, ja seda nimelt "raskete sabade" poolest. Niisugusele järeldusele jõuame seetõttu, et standardiseeritud ekstsess on absoluutväärtuselt suurem kui 2 ja positiivne.

Variable:	vanus	kaal	sugu
Sample size	28	15	15
Average	31.1786	65.3333	1.6
Median	28	65	2
Mode	27	65	2
Geometric mean	30.4895	64.8137	1.51572
Variance	54.078	72.2381	0.257143
Standard deviation	7.35378	8.4993	0.507093
Standard error	1.38973	2.19451	0.130931
Minimum	24	50	1
Maximum	53	80	2
Range	29	30	1
Lower quartile	26.5	60	1
Upper quartile	35.5	72	2
Interquartile range	9	12	1
Skewness	1.8339	0.102437	-0.455083
Standardized skewness	3.96168	0.161968	-0.719549
Kurtosis	3.44639	-0.432962	-2.09402
Standardized kurtosis	3.72253	-0.342287	-1.65547

Jn 6.7

Tabeli teises veerus paiknevad tunnuse kaal, kolmandas tunnuse sugu karakteristikud. Kuivõrd tunnus sugu on kaheväärtuseline (seega järjestustunnus, vt [7], lk 18 jj), tohib tema jaoks mõningad arvkarakteristikud leida, kuid nendes sisalduv informatsioon on küllaltki tagasihoidlik.

Keskmine annab ühtaegu edasi tõenäosusunktsiooni: 1,6 - 1 = 0,6 on väärtuse 2 suhteline sagedus; kuivõrd

$n = 15$, siis $n_2 = 9$ ja $n_1 = 6$. Dispersioon, standardhälve, asümmeetria ja ekstsess mingit lisainfot ei anna, pealegi pole nende kasutamine korrektne.

Ülejäänud karakteristikud kinnitavad teadaolevat: väärtused paiknevad 1 (min) ja 2 (max) vahel, väärtus 2 on mood, kuid ühtlasi ka mediaan. See tõsiasi, et kvartiilid ei lange ühte, näitab, et kumbki väärtus ei oma sagedust üle 75 %.

Märkigem, et protseduuri *Summary Statistics* on väga sobiv kasutada tunnuse väärtuste õigsuse (jämedate vigade puudumise) kontrollimisel. Selleks piisab, kui vaadata läbi tunnuse miinimum ja maksimum.

6.3.2. Karpdiagramm (*Box-and-Whisker Plot*) järkstatistikute illustreerimiseks. Protseduur *Box-and Whisker Plot* on esimene allmenüüs *I. Exploratory Data Analysis*.

See protseduur arvutab ja esitab graafiliselt tunnuse järgmised karakteristikud:

miinimum,
alumine kvartiil,
mediaan,
ülemine kvartiil,
maksimum.

Mõningatel juhtudel esitatakse graafiliselt veel ka miinimumile järgnevaid ja/või maksimumile eelnevaid punkte.

Protseduuri tellimuspaneelile tuleb trükkida (või valida võtme F7 abil) uuritava tunnuse nimi. Tunnus peab olema arvuline (tüübid N ja I). Mingeid valitavaid parameetreid sellel protseduuril ei ole, samuti puudub lisainformatsioon (*options*).

Protseduuri käivitamise tulemusena ilmub ekraanile graafik (vt jn 6.8).

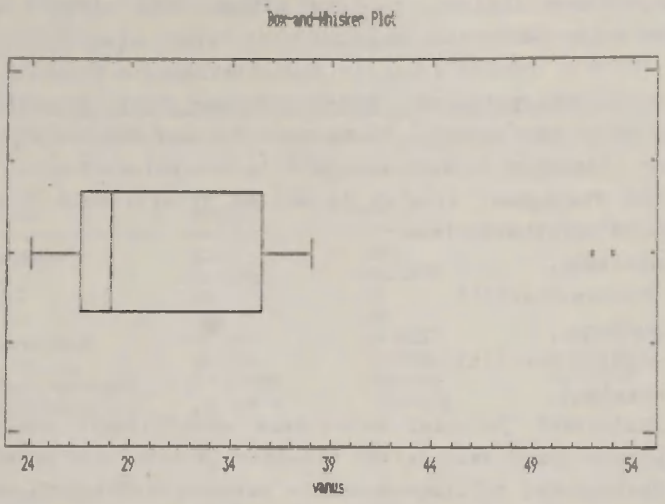
Joonisel näeme tunnuse vanus põhjal konstrueeritud karpi, mille vasakpoolseks servaks on alumine ja parempoolseks ülemine kvartiil, seega katab karp parajasti poole valimist. Karpi lõikav vertikaallõik kujutab mediaani. Karbist ulatuvad välja "vuntsid", mille otspunktideks on vastavalt valimi

vähim ja suurim element, mis rahuldavad järgmist tingimust:

$$|x_i - \text{med}| < 1{,}5(q_u - q_a) \qquad (6.2)$$

(q_u ja q_a tähistavad vastavalt ülemist ja alumist kvartiili).

Mediaanist kaugemal kui 1,5 kvartiilide vahet paiknevad valimi elemendid esitatakse eraldiseisvate punktidena. Seega võivad nii miinimum kui maksimum olla kujutatud kas "vuntsi" otspunktina (kui on täidetud tingimus 6.2) või äärmise eraldi asuva punktina.



Jn 6.6.

Vaadeldes joonist 6.8, näeme, et miinimum, kvartiilid ja mediaan on tõepoolest kooskõlas joonisel 6.7 esitatud andmetega. Et aga maksimumi (53 aastat) erinevus mediaanist (28 aastat) on märksa suurem poolteisekordsest kvartiilide vahest (14,5 aastat), siis on see joonisel märgitud diskreetse punktina. Meie jaoks uus on teave selle kohta, et ka maksimumile järgnev punkt erineb ülejäänutest tunduvalt ning alles kolmas punkt rahuldab "vuntside tingimust".

Karpdiagrammi on käepärane kasutada esmatutvumiseks tunnusega, samuti jämedate vigade avastamiseks. On ju diskreetselt paiknevad punktid n-ö potentsiaalsed jämedad vead;

kuid nad võivad olla ka andmestiku "seaduslikud", kuigi ülejäänutest erinevad, punktid, nagu käesolevalgi juhul.

6.3.3. Keskväärtuse ja dispersiooni usalduspiirid ja hüpoteeside kontrollimine nende kohta (*One Sample Analysis*). Protseduur *One Sample Analysis* on esimene allmenüüs *G. Estimation and Testing*.

See protseduur arvutab arvtunnuse keskmise, dispersiooni, standardhälbe, mediaani ning keskväärtuse ja dispersiooni usalduspiirid vastavalt tellija poolt soovitud usaldusnivoole.

Peale selle võimaldab nimetatud protseduur kontrollida keskväärtuse kohta ühe- ja kahepoolseid hüpoteese tellija poolt soovitud olulisuse nivool.

Protseduuri kasutamisel on oluliseks sisuliseks kitsenduseks nõue, et analüüsitav tunnus peab olema arvuline, kusjuures statistiliseks eelduseks on ühtlasi see, et tunnuse jaotus üldkogumis ei tohiks liialt palju erineda normaaljaotusest.

Olgu märgitud, et nimetatud protseduuri kasutamisel normaaljaotusest erinevate jaotuste korral ei teki tavaliselt olulisi vigu siis, kui tunnuse "sabad" on normaaljaotusega võrreldes "kergemad", seevastu "raskete sabade" (kui ekstsessikordaja on oluliselt positiivne) ja väikeste valimite korral on ekslike järelduste oht olemas.

Protseduuri tellimiseks trükitakse tunnuse nimi ja käivitatakse võtmega F6. Tulemusena saadakse paneel, mis on kujutatud joonisel 6.9.

Sellel on keskväärtuse ja dispersiooni usalduspiirid arvutatud 95 % (vaikimisi määratud) usaldusnivool. Paneeli alumises servas on kontrollitud hüpoteesipaari

$$H_0: \mu = 0,$$
$$H_1: \mu \neq 0,$$

kusjuures olulisuse nivoo $\alpha = 0,05$.

Saadud resultaatpaneelil on rida muudetavaid parameetreid (aktiivseid välju), mida teisendades on võimalik saada sama tunnuse kohta uusi tulemusi.

One-Sample Analysis Results

		kasv
Sample Statistics:	Number of Obs.	15
	Average	172.533
	Variance	44.8381
	Std. Deviation	6.69613
	Median	172
Confidence Interval for Mean: Sample 1	95 Percent	
	168.824 176.242	14 D.F.
Confidence Interval for Variance: Sample 1	0 Percent	
Hypothesis Test for H0: Mean = 0 vs Alt: NE at Alpha = 0.05	Computed t statistic = 99.7918	
	Sig. Level = 0	
	so reject H0.	

Jn 6.9.

Muuta saab usaldusnivood (II ja III blokis).

Hüpoteeside kontrollimisel on võimalikud järgmised muudatused:

1) null-hüpoteesis võib võrduse $\mu = 0$ asendada võrdusega $\mu = \mu_0$, kus μ_0 on mingi suvaline arv;

2) hüpoteesi H_1 võib valida ka ühepoolsena, kas

H_1: $\mu > \mu_1$ (siis tuleb NE asemele trükkida GT)

või

H_1: $\mu < \mu_1$ (NE asemele trükkida LT);

3) muuta saab olulisuse nivood α.

Pärast parameetrite muutmist tuleb protseduur uuesti võtme F6 abil käivitada.

Vaatleme veel hüpoteesi kontrollimise kohta väljastatavat teavet.

1. Arvutatud t-statistiku (*Computed t statistic*) väärtus t võimaldab, arvestades soovitud olulisuse nivood ja leitud vabadusastmete arvu (*D.F. - degrees of freedom*), tabelite abil teha järeldusi hüpoteesi kehtivuse kohta:

kui $|t| > t_\alpha(f)$, siis võetakse vastu H_1, kus $t_\alpha(f)$ on tabelis antud t-jaotuse kriitiline väärtus.

Tegelikult pole tabelit kasutada vaja, sest vajalik teave väljastatakse ka mugavamal kujul järgmistes ridades.

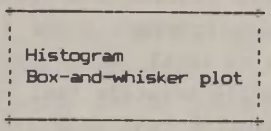
2. Olulisuse tõenäosus (*Sig. Level*) p. Kui $p \leq \alpha$, siis on H_1 tõestatud, kui aga $p > \alpha$, siis võetakse vastu H_0 (kuid seda ei loeta tõestatuks).

3. Hüpoteesi kontrollimise tulemus tekstina:

a) *so reject* H_0 (st H_0 kummutatakse ja H_1 võetakse vastu);

b) *so do not reject* H_0 (st H_0 ei kummutata, H_1 ei saa vastu võtta);

Protseduuril on kaks täiendavat võimalust (*Options*). Võtmele F5 vajutamisel ilmub aken (vt jn 6.10) võimaluste valikuks. Esimene neist on histogramm (seda kirjeldame punktis 6.4.2), teine - karpdiagramm (vt p 6.3.2).



Jn 6.10

6.3.4. Kaalutud keskmine (*Weighted Averages*). Protseduur *Weighted Averages* on neljas protseduur allmenüüs *F. Descriptive Methods*. See protseduur arvutab tunnuse kaalutud keskmise vastavalt antud kaaludele, võimaldades seega näiteks kaalutud valimi rakendamist (vt [7], lk 27-29).

Erinevalt eelmistest protseduuridest vajab kaalutud keskmiste arvutamine lisaks tunnuse väärtustele veel täiendavaid andmeid, nimelt kaalusid. Kaalud peaksid olema vormistatud samuti tunnusena, mis on sama pikkusega kui töödeldav tunnuski. Kui töödeldavad tunnused on ühendatud objekt-tunnus-maatriksisse, siis on võimalik kõigi tunnuste kaalutud keskmised (samade kaalude alusel) korraga arvutada.

Kaalutud keskmisi tohib arvutada arvtunnuste jaoks (tüübid N ja I).

Kaalutud keskmiste tellimuspaneel on esitatud joonisel 6.11.

Siin on kasutatud SG võimalust esitada tellimuses vajalik tunnus väärtuste loeteluna (n = 5).

Weighted Averages

Data vector or matrix: 1.5 2 2.5 3 3.5

Weight vector: 5 4 3 2 1

Jn 6.11

Arvutustulemus 2,1667 erineb ilmselt tavalisest keskmisest (mis on 2,5) väiksemate väärtuste suurema osatähtsuse tõttu.

Protseduuril ei ole täiendava teabe režiime (*Options*).

6.3.5. Kvantiilid (*Percentiles*). Protseduur *Percentiles* on viies protseduur allmenüüs *F. Descriptive Methods*. See protseduur võimaldab arvutada antud tunnuse empiirilise p-kvantiili väärtust vastavalt antud tõenäosusele p; ühe arvutussammuga on võimalik arvutada kuni 15 kvantiili.

Samuti on võimalik teha pöördtehet: vastavalt tunnuse väärtusele x arvutada empiiriline tõenäosus P(**X** < x), st hinnata tõenäosust, et vastava tunnuse väärtus juhuslikul objektil on väiksem kui x.

Tunnus **X** peab olema arvuline (tüüp N või I).

Tellimuse vormistamisel trükitakse kõigepealt tunnuse nimi, seejärel täidetakse tellimuspaneel (vt jn 6.12), trükkides veergu *Percentages* soovitavad tõenäosuse p väärtused (protsentides). Käivitamise järel ilmuvad väärtused ka parempoolsesse tulpa (*Percentiles*), vt jn 6.13.

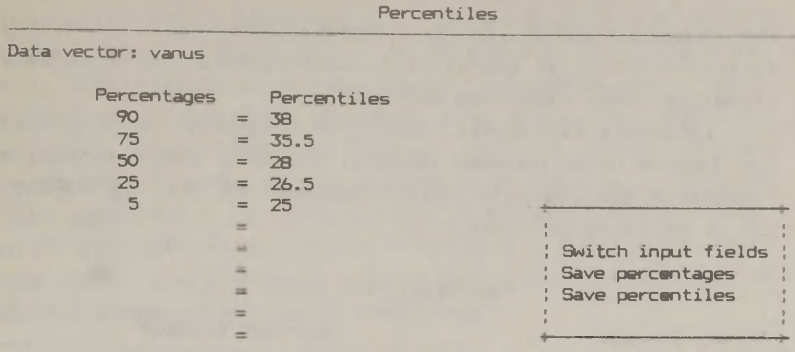
Percentiles

Data vector: vanus

Percentages		Percentiles
90	=	
75	=	
50	=	
25	=	
5	=	

Jn 6.12

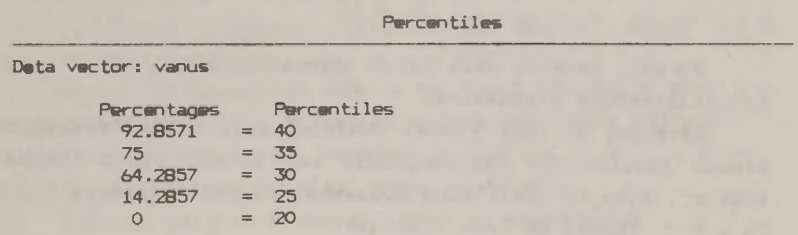
Võtme F5 abil saame akna täiendavate võimaluste loeteluga. Neist viimased kaks võimaldavad salvestada tabelis



Jn 6.13

olevaid veerge (NB! kasulik on salvestada kindlasti mõlemad, üksikult on nad väheinformatiivsed).

Esimene lisaprotseduur *Switch input fields* (lülita ümber sisendväli) aga vahetab protseduuris argumendi ja funktsiooni, seega suudab aktiivseks tabeli teise veeru, millesse võib nüüd trükkida soovitavaid tunnuse väärtusi (kvantiile). Käivitamisel arvutatakse vastavad empiirilise jaotusfunktsiooni väärtused (tõenäosused), vt jn 6.14.



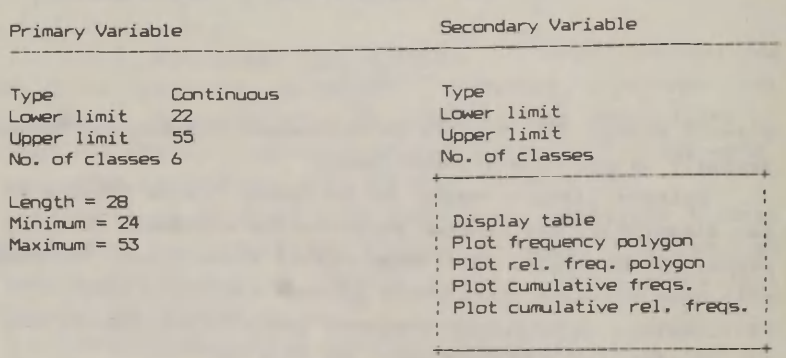
Jn 6.14

§ 6.4. Tunnuse jaotust iseloomustavate tabelite ja graafikute leidmine SG-süsteemi abil

6.4.1. Sagedustabel (*Frequency Tabulation*). Protseduur *Frequency Tabulation* paikneb allmenüüs *F. Descriptive Methods* teisel kohal. Selle protseduuri abil on võimalik arvutada

ühe tunnuse sagedus- ja jaotustabelit (vastavalt etteantud klassipiiridele) ja samuti nelja nurdjoonekujulist graafikut illustreerimaks nimetatud tabeleid.

Tellimuse esitamisel trükitakse kõigepealt tunnuse nimi (on lubatud kõik tunnuse tüübid). Seejärel ilmub ekraanile standardne jaotustabeli tellimuspaneel (vt [7], lk 59-60), mis on esitatud joonisel 6.15.



Jn 6.15

Paneeli vasakus osas tuleb täpsustada soovitava jaotustabeli/graafiku parameetrid.

Märgime, et tüüp (*Type*) tähistab siin tabeli formeerimiseks kasutatavat tunnusetüüpi; see erineb pisut tunnuse tüübist, mida on käsitletud õppevahendis [7] punktides 1.2.3 ja 2.6.1. Tüübil on kolm väärtust:

1) *Continuous* - sel juhul on tellijal võimalik vabalt valida klassipiirid ja klasside arv;

2) *Discrete* - sel juhul määratakse tunnuse väärtusteks= =väärtusklassideks täisarvud;

3) *List* - sel juhul vastab igale erinevale väärtusele üks klass.

Sümboltunnuste (tüüp C) puhul on kasutatav ainult viimane võimalus, kusjuures süsteem järjestab väärtused alfabeetiliselt.

Teised tabeli parameetrid - alumine piir (*Lower limit*) ja ülemine piir (*Upper limit*) tuleb määrata tellijale sobivalt, kasutades tunnuse kohta antud informatsiooni - minimaalset ja maksimaalset väärtust, ning arvestades tõsiasja, et igasse klassi loetakse ülemine otspunkt sisse, alumine aga välja, st i-nda klassi struktuur on $(a_i, a_{i+1}]$. Tellija poolt määratav on ka klasside arv. Olgu märgitud, et programmiliselt määrab SG-süsteem tavaliselt ülearu suure klasside arvu; rusikareeglina võiks soovitada klasside arv k määrata valimi mahu n järgi eeskirjaga

$$k \approx \sqrt{n} .$$

Klasside otspunkte ja arvu sobivalt valides on võimalik saada ka "ümmargusi" klassipiire.

Tähelepanelik tuleb olla täisarvuliste väärtustega (või muidu diskreetselt paiknevate) tunnuste puhul; siin tuleks klassipiirid määrata nii, et igas klassis oleks võrdne arv tunnuse väärtusi. Näiteks klassipiirid 0; 2,5; 5; 7,5; 10 on täisarvulise tunnuse analüüsimisel halvad seetõttu, et nii tekkinud klassides paiknevad vastavalt tunnuse väärtused 1, 2; 3, 4, 5; 6, 7; 8, 9, 10, seega osas klassidest on 2, osas 3 tunnuse väärtust. Märgime, et selles mõttes on ka joonisel 6.15 esitatud tellimus täisarvulise tunnuse vanus jaoks halb - osas klassides on 5, osas 6 väärtust.

Pärast parameetrite määramist ilmub ekraanile aken (vt jn 6.15 paremal), mis võimaldab valida soovitava väljundi.

Vaatleme kõigepealt sagedus-jaotustabelit (vt jn 6.16), mis saadakse valiku *Display table* tulemusena.

Tabelis paikneb 8 veergu, need on vastavalt

Class - klassi number (i),

Lower Limit - klassi alumine piir (a_i),

Upper Limit - klassi ülemine piir (a_{i+1}),

Midpoint - klassi keskpunkt $(a_i+a_{i+1})/2$,

Frequency - sagedus - klassi kuuluvate objektide arv n_i

Relative Frequency - suhteline sagedus n_i/n (n on valimi maht),

Cumulative Frequency - summeeritud (kumuleeritud) sagedus $\sum_{j=1}^{i} n_j$,

Cum. Rel. Frequency - summeeritud (kumuleeritud) suhteline sagedus (ka empiiriline jaotusfunktsioon) $\frac{1}{n}\sum_{j=1}^{k} n_j$.

Frequency Tabulation

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below		22.00		0	.0000	0	.000
1	22.00	27.50	24.75	12	.4286	12	.429
2	27.50	33.00	30.25	6	.2143	18	.643
3	33.00	38.50	35.75	8	.2857	26	.929
4	38.50	44.00	41.25	0	.0000	26	.929
5	44.00	49.50	46.75	0	.0000	26	.929
6	49.50	55.00	52.25	2	.0714	28	1.000
above	55.00			0	.0000	28	1.000

Mean = 31.1786 Standard Deviation = 7.35378 Median =28

Jn 6.16

Ridade arv tabelis on kahe võrra suurem tellitud klasside arvust.

Esimene rida vastab klassile "*at or below* a_1", sellesse klassi arvatakse alumisel piiril a_1 või sellest allpool paiknevad tunnuse väärtused.

Järgnevad tellimuse põhjal moodustatud (võrdse pikkusega) klassid. Viimane rida "*above* a_{k+1}" koondab informatsiooni viimase klassi ülemist piiri ületavate objektide kohta.

Kriipsu alla on trükitud ka olulisemad arvparameetrid - keskmine, standardhälve ja mediaan.

Peale tabeli nägemist tekib sageli vajadus muuta tabuleerimisparameetreid, st vältida niihästi liiga tihedaid kui ka liiga välja venitatud tabeleid. Joonisel 6.17 ongi sama tabel muudetud parameetritega, klassi pikkus on 2 aastat, klasse on 8 ja 2 objekti (mis paiknevad ülejäänutest väga kaugel) on tabelist välja jäänud. Ühtlasi on parandatud ka viga - kõigisse tabeli lahtreisse satub täpselt 2 täisarvulist väärtust.

Sageduspolügooni ehk sagedushulknurga saame valiku *Plot frequency polygon* tulemusena (vt jn 6.15). Polügooni, mis vastab tabelile jooniselt 6.16, kujutab joonis 6.18. Siin on

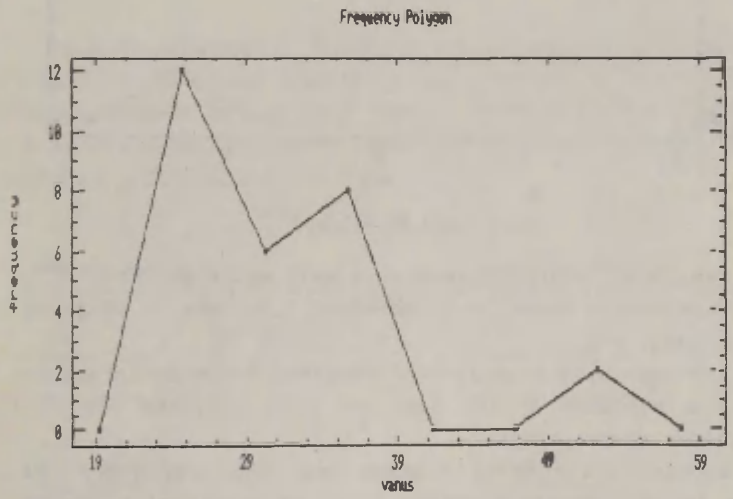
x-teljel uuritava tunnuse väärtused, y-teljel vastavad sagedused (vt jn 6.16, tabeli viies veerg).

Frequency Tabulation

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below		23.00		0	.0000	0	.000
1	23.00	25.00	24.00	4	.1429	4	.143
2	25.00	27.00	26.00	8	.2857	12	.429
3	27.00	29.00	28.00	5	.1786	17	.607
4	29.00	31.00	30.00	1	.0357	18	.643
5	31.00	33.00	32.00	0	.0000	18	.643
6	33.00	35.00	34.00	3	.1071	21	.750
7	35.00	37.00	36.00	4	.1429	25	.893
8	37.00	39.00	38.00	1	.0357	26	.929
above	39.00			2	.0714	28	1.000

Mean = 31.1786 Standard Deviation = 7.35378 Median = 28

Jn 6.17

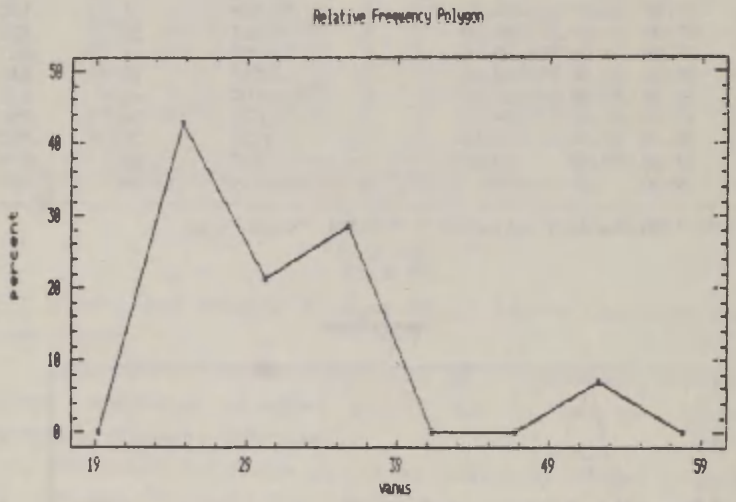


Jn 6.18

Polügooni tippudeks on punktid, mille x-koordinaadiks on vastava klassi keskpunkt (neljas veerg) ja y-koordinaadiks sagedus (viies veerg). Murdjoone algus- ja lõpp-punkti

jaoks on arvutatud kunstlikud "keskpunktid" nullinda (esimesele eelneva) ja (k+1)-se (k-ndale järgneva) klassi jaoks, lugedes need klassid ülejäänutega ühepikkuseks.

Suhteliste sageduste polügooni saame valiku *Plot rel. freq. polygon* tulemusena. Joonisel 6.16 esitatud tabelile vastavat suhtelise sageduse polügooni kujutab joonis 6.19.

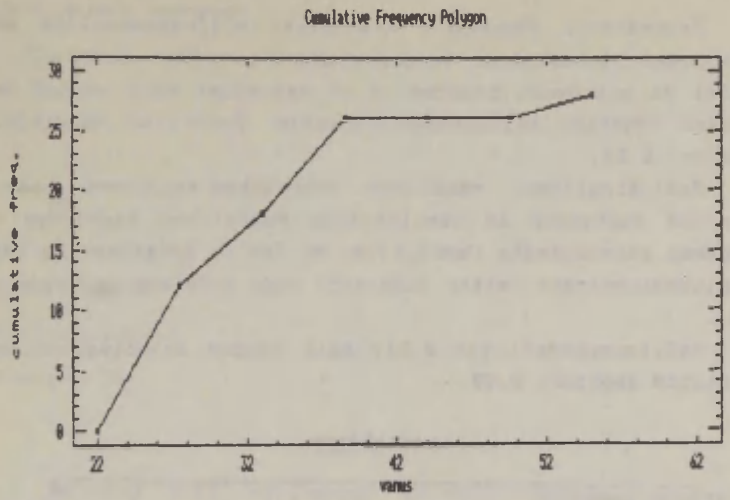


Jn 6.19

See joonis erineb eelmisest ainult selle poolest, et y-teljel paiknevad suhtelised sagedused n_i/n, mis on väljendatud protsentides.

Joonisel 6.20 on kujutatud kumuleeritud sageduste graafik, mis saadakse aknast joonisel 6.15 neljanda rea *Plot cumulative freqs.* valikul.

Joonis, mille saame viimase rea *Plot cumulative rel. freqs.* valikul aknast joonisel 6.15, on joonisega 6.20 sarnane, ainsa erinevusega, et y-teljel on kujutatud suhtelised sagedused (protsentides).



Jn 6.20

6.4.2. Sageduste tulpdiagramm (*Frequency Histogram*). Protseduur *Frequency Histogram* on allmenüüs *F. Descriptive Methods* kolmas. See on ette nähtud teise tüüpilise sagedustabelit illustreeriva graafikute pere - tulpdiagrammi e histogrammi - valmistamiseks.

Tabulation Input Panel

Primary Variable		Secondary Variable
Type	Continuous	Type
Lower limit	22	Lower limit
Upper limit	55	Upper limit
No. of classes	11	No. of classes
Length = 28		Length =
Minimum = 24		Minimum =
Maximum = 53		Maximum =

Top Title: sageduste tulpdiagramm
(2 lines)
X-axis title: vanus
Y-axis title: sagedused

Cumulative: No Relative: No

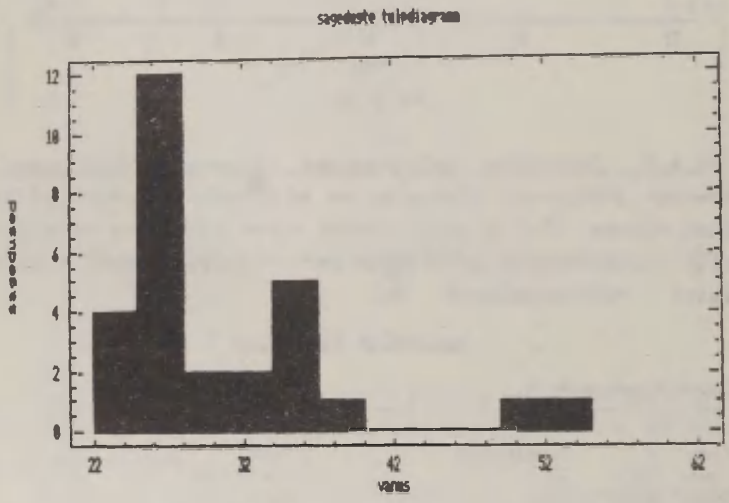
Jn 6.21

7

Protseduuri *Frequency Histogram* tellimuspaneeliks on standardne jaotustabeli tellimuspaneel (vt [7], lk 59-60), millel on erinevalt joonisel 6.15 esitatust ette nähtud ka joonise tekstide kujundamise võimalus. Paneel on kujutatud joonisel 6.21.

Neli graafikut - sagedused, suhtelised sagedused, kumuleeritud sagedused ja kumuleeritud suhtelised sagedused - saadakse parameetrite *Cumulative: No/Yes* ja *Relative: No/Yes* kombinatsioonidena (mitte akna abil nagu eelmises protseduuris).

Tellimuspaneeli (jn 6.21) abil saadud tulpdiagramm on kujutatud joonisel 6.22.



Jn 6.22

6.4.3. Horisontaalne tulpdiagramm (*Barcharts*). Protseduur *Barcharts* on allmenüüs *E. Plotting Functions* viies.

See protseduur on mõeldud eeskätt nominaaltunnuste ja väikese väärtuste arvuga diskreetsete tunnuste kirjeldamiseks; põhimõtteliselt on tema abil võimalik ka kaht tunnust üheaegselt töödelda, kuid seda varianti me siin veel ei käsitle. Protseduuri *Barcharts* tellimuspaneeli kirjeldab joonis 6.23.

Data vectors: synnikoht

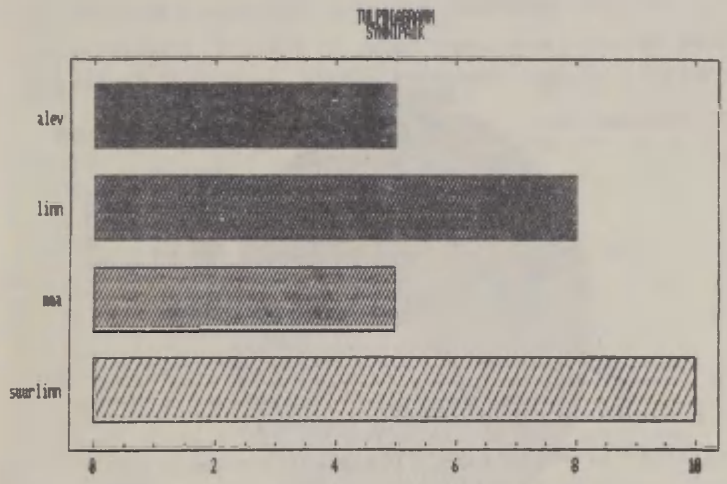
Primary Labels:

Secondary Labels:

Tabulate: Yes

Jn 6.23

Kuivõrd meie tellimuses esitatud tunnuse **synnikoht** väärtusteks on nimetused alev, linn, **maa** ja **suurlinn**, ei ole tarvis spetsiaalseid märgendeid (*labels*) sisestada ning seetõttu jääb teine rida tellimuspaneelil täitmata. Tellimuse täitmise tulemust kujutab joonis 6.24.



Jn 6.24

6.4.4. Sektordiagramm (*Piecharts*). Protseduur *Piecharts* on allmenüüs *E. Plotting Functions* kuues. See on ette nähtud nominaalse või diskreetse tunnuse jaotuse illustreerimiseks nn sektordiagrammi abil.

Tellimuspaneeli (vt jn 6.25) täitmine on sarnane eelmise protseduuri omaga. Detailsemaks graafiku kujundamiseks ilmub veel teinegi paneel, mis lubab määrata graafiku suuruse, paiknemise, pinnakujunduse ja ka tekstid. Selle paneeli abil on võimalik eemaldada sektordiagrammist üht või mitut sektorit tellija poolt valitud kaugusele, et juhtida sellele erilist tähelepanu. Tulemusena saadud graafik on kujutatud joonisel 6.26.

Piecharts

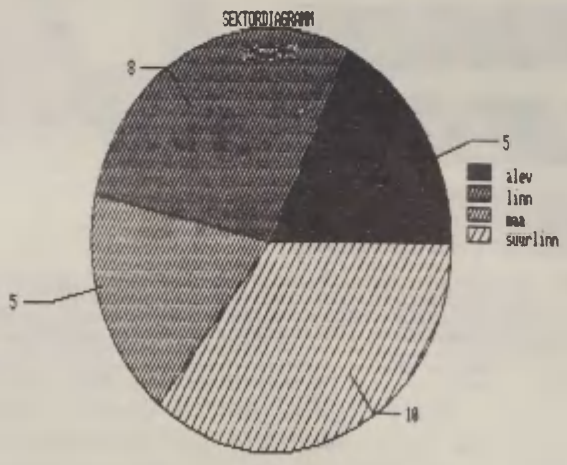
Data vector: synnikoht

Labels:

Tabulate: Yes

Percentages: No

Jn 6.25



Jn 6.26

6.4.5. Tüvidiagramm (*Stem-and-Leaf Display*). Protseduur *Stem-and-Leaf Display* on allmenüüs *I. Exploratory Data Analysis* viimane, s.o seitsmes.

See protseduur on ette nähtud arvulise tunnuse väärtuste kompaktseks esitamiseks nn tüvidiagrammi kujul.

Tellimuse esitamiseks piisab tunnuse nime trükkimisest (vt jn 6.27), protseduuril puuduvad täiendavad parameetrid ja lisarežiimid.

```
                    Stem-and-Leaf Display
-------------------------------------------------------------

Data vector: vanus
```

Jn 6.27

Tunnuse **vanus** tüvidigrammi esitab joonis 6.28. Näeme, et siin vastab iga rida kahele aastakäigule, kusjuures kümnelised on märgitud n-ö "tüvel", ühelised aga okstel "lehtedena". Tüvel on lisaks kümnelistele märgitud ka sümbol, mis tähistab üheliste järjekorranumbrit: * - 0 või 1, T - 2 või 3, F - 4 või 5, S - 6 või 7, o - 8 või 9. Iga konkreetse andmestiku korral valitakse "klassid" üldiselt erinevalt. Ülejäänutest tugevasti erinevad väärtused (klassid) paigutatakse eraldi. Vasakul on sagedused, mis enne mediaani (arv sulgudes) liidetakse alates algusest, pärast - alates lõpust.

```
Stem-and-leaf display for vanus: unit = 1    1|2  represents 12

   4     2F|4555
  12     2S|66677777
 (5)     2o|88889
  11     3*|0
  10     3T|
  10     3F|445
   7     3S|6666
   3     3o|8

         HI|52,53
```

Jn 6.28

Näeme, et selline tüvidiagramm kopeerib teatavas mõttes joonistel 6.18 ja 6.22 esitatud graafikuid, kuid kasutab mõneti lihtsamaid vahendeid ja esitab informatsiooni täielikumalt (algandmete tasemel). Mitte eriti suure n korral on ka tüvidiagramm väga sobiv esmatutvuseks tunnusega.

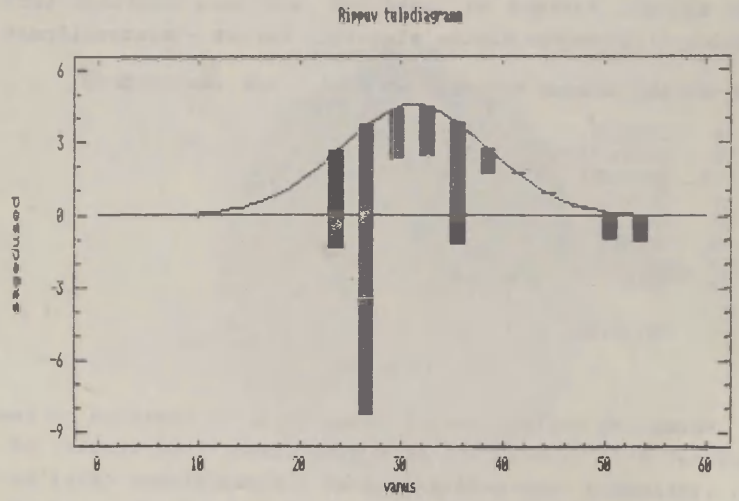
§ 6.5. Jaotuste sobitamine. Hüpoteeside kontrollimine jaotuste kohta

Üheks oluliseks probleemiks andmetöötluses on sellise teoreetilise jaotuse leidmine, mis võiks olla üldkogumi jaotuseks. Selle ülesande lahendamine ongi jaotuste sobitamine (*distributions fitting*). Ülesande lahendamist võime vaadelda kaheetapilisena:

- hüpoteeside genereerimine (visuaalse võrdlemise teel),
- hüpoteeside kontrollimine testide abil.

SG-süsteem sisaldab rea võimalusi nii ühe kui teise etapi teostamiseks. Vaatleme neid järgnevalt.

6.5.1. Rippuv tulpdiagramm (*Hanging Histobars*) empiirilise jaotuse võrdlemiseks normaaljaotusega. Protseduur *Hanging Histobars* on allmenüüs *G. Estimation and Testing* neljas. See protseduur on ette nähtud arvtunnuse empiirilise jaotuse võrdlemiseks normaaljaotusega, kusjuures võrreldav normaaljaotus identifitseeritakse tunnuse keskmise ja standardhälbe kaudu.



Jn 6.29

Protseduuri tellimisel trükitakse kõigepealt tunnuse nimi, seejärel tuleb täita standardne jaotustabeli tellimuspaneel (vt nt jn 6.21). Protseduuri ainsaks tulemuseks on graafik, vt jn 6.28. Puuduvad ka lisavõimalused.

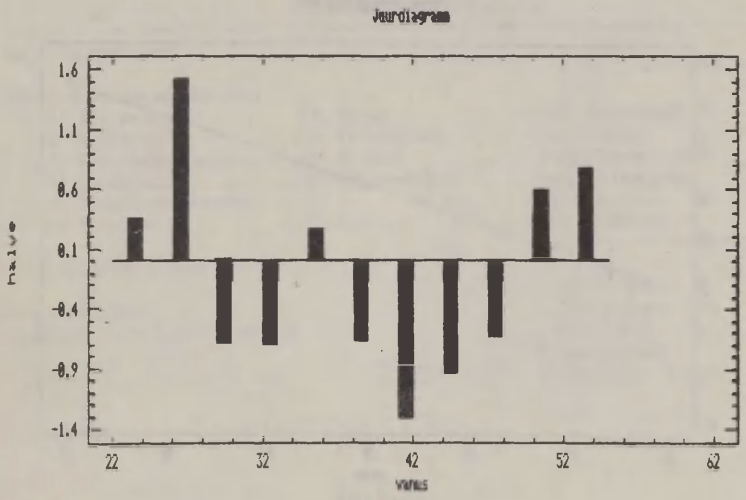
Selle protseduuri omapäraks on empiirilist jaotust kujutavate tulpade iseäralik rippuv asetus.

Võrdlus normaaljaotuse tihedusfunktsiooniga, mis on esitatud pideva kõverana, viib igatahes mõttele, et toodud näites on tegemist normaaljaotusest märgatavalt erineva jaotusega.

6.5.2. Juurdiagramm (*Suspended Rootogram*) empiirilise jaotuse võrdlemiseks normaaljaotusega. Protseduur *Suspended Rootogram* on allmenüüs *I. Exploratory Data Analysis* kuues, ta on samuti nagu eelminegi protseduur ette nähtud arvtunnuse empiirilise jaotuse võrdlemiseks normaaljaotusega.

Tellimine toimub samal viisil nagu eelmisegi protseduuri korral: tunnuse nime trükkimise järel ilmub ekraanile standardne tellimuspaneel (vt jn 6.21).

Pärast tellimuspaneeli täitmist ja protseduuri käivitamist saame graafiku (vt jn 6.30), millel on täpselt samuti



Jn 6.30

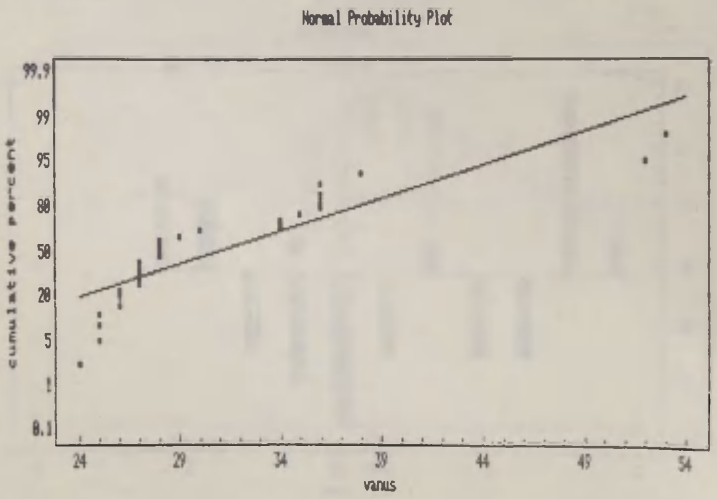
11 tulpa nagu ka joonisel 6.29 (kus küll kolm tulpa on pikkusega 0). Tähelepanelikul vaatlemisel näeme ka, et kõigile neile tulpadele, mis joonisel 6.29 on liiga pikad, vastavad joonisel 6.30 ülespoole, liiga lühikestele aga allapoole suunatud tulbad. Tulpade pikkusteks on suurused

$$\sqrt{n_i} - \sqrt{p_i n}$$

(vt [5]), kus n_i on i-nda väärtuse (väärtusklassi) empiiriline sagedus, $p_i n$ aga teoreetiline sagedus, kusjuures p_i on vastava väärtusklassi tõenäosus normaaljaotuse eeldusel.

Mingit täiendavat ega valitavat lisainformatsiooni see protseduur ei paku.

6.4.3. Empiirilise jaotuse võrdlemine normaaljaotusega nn tõenäosuspaberi abil (*Normal Probability Plot*). Protseduur *Normal Probability Plot* on allmenüüs *G. Estimation and Testing* kolmas. See sisaldab n-ö klassikalist meetodit empiirilise jaotuse võrdlemiseks normaaljaotusega. Võrdluse aluseks on empiirilise jaotusfunktsiooni graafik (lähend sellele on kujutatud joonisel 6.20), mida võrreldakse normaaljaotuse jaotuse jaotusfunktsiooniga (vt [7], jn 3.6).



Jn 6.31

Meetodi iseärasuseks on mittelineaarse skaala kasutamine graafiku y-teljel, mis teisendab normaaljaotuse jaotusfunktsiooni sirgeks ning muudab seega empiirilise jaotusfunktsiooni hälbed hästi jälgitavateks.

Protseduuri tellimiseks on tarvis trükkida võrreldava arvtunnuse nimi. Tulemuseks on graafik, vt jn 6.31.

Empiirilise jaotusfunktsiooni saamiseks tuleks üksikpunktid murdjoonega ühendada, kuid niigi on nähtavad küllaltki suured hälbed.

Lisainformatsiooni see protseduur ei paku.

6.4.4. Hüpoteeside kontrollimine jaotuse sobivuse kohta (*Distribution Fitting*). Protseduur *Distribution Fitting* on esimene allmenüüs *H. Distribution Functions*. See annab võimaluse kontrollida hüpoteesipaari

$$H_0: P_x = P_{\vartheta_0},$$
$$H_1: P_x \neq P_{\vartheta_0},$$

kus P_x on antud valimile vastava üldkogumi jaotus, P_{ϑ_0} aga mingi antud jaotus parameetrilisest perest $\mathcal{P}_\vartheta$.

Võrreldav tunnus peab olema arvuline.

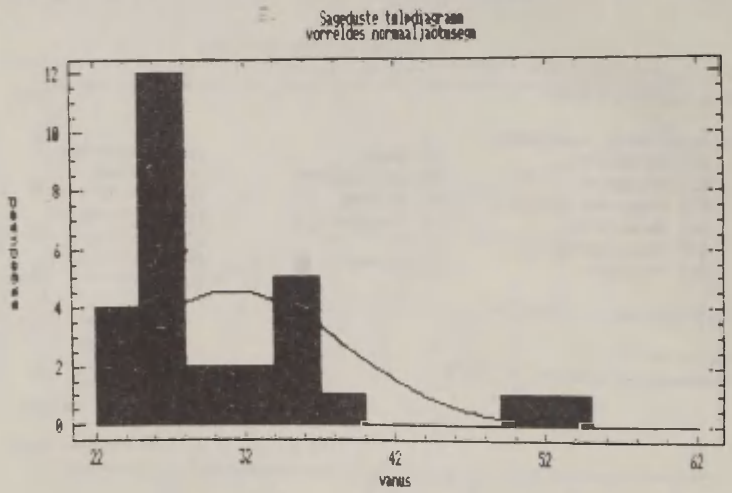
```
                        Distribution Fitting
------------------------------------------------------------------------

Data vector: vanus

Distributions available:
   (1) Bernoulli          (7) Beta            (13) Lognormal
   (2) Binomial           (8) Chi-square      (14) Normal
   (3) Discrete uniform   (9) Erlang          (15) Student's t
   (4) Geometric          (10) Exponential    (16) Triangular
   (5) Negative binomial  (11) F              (17) Uniform
   (6) Poisson            (12) Gamma          (18) Weibull
                                              +-----------------+
Distribution number: 14                       |                 |
                                              | Histogram       |
Mean: 31.1786                                 | Chi-square test |
Standard deviation: 7.35378                   | K-S test        |
                                              | Tail areas      |
                                              | Critical values |
                                              |                 |
                                              +-----------------+
```

Jn 6.32

Protseduuri *Distribution Fitting* tellimiseks kasutatav paneel on sarnane õppevahendi esimeses osas peatükis 3 kirjeldatud jaotusseaduste paneeliga, erinedes vaid selle poolest, et tellimus algab uuritava tunnuse nime sisestamisega (vt jn 6.32). Seejärel valitakse eeldatava teoreetilise jaotuse (jaotuste pere $\mathcal{P}_\vartheta$) number. Käivitamise järel arvutatakse valimi põhjal sobivaimad parameetri väärtused ϑ_0. Näeme, et joonisel 6.32, kus sobitatavaks jaotuseks on valitud normaaljaotus (nr. 14), on identifitseerivateks parameetriteks keskmine (*mean*) ja standardhälve (*standard deviation*), mille sobivaimad väärtused on trükitud paneelile.

Olgu märgitud, et sellel sammul võib arvutatud parameetrid asendada ka mingist muust arutlusest leitud väärtustega. Mõnikord on isegi võimalik saada mõne teise parameetri väärtuse korral paremat lähendit (selle põhjuseks on lähenduskriteeriumide erinevus).

Protseduuri väljundite valik toimub akna abil (vt jn 6.32 paremal all nurgas). Kahe esimese - histogrammi ja χ^2-testi - saamiseks tuleb täita veel tabeli kujundamise paneel (vt nt jn 6.21).



Jn 6.33

Histogrammi graafikule kantakse ka sobitatava pideva jaotuse tihedusfunktsiooni graafik või diskreetse jaotuse tõenäosusfunktsiooni tulpdiagramm. Joonisel 6.33 näeme, vastavalt tehtud tellimusele, normaaljaotuse tihedusfunktsiooni graafikut.

Uus võrreldes eelmiste protseduuridega on hüpoteeside kontrollimise võimalus χ^2-testi abil, mis tekib, kui valida joonisel 6.32 kujutatud aknast teine rida *Chi-square test*. Vastavaid arvutusi kujutab jn 6.34.

Chisquare Test

	Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chisquare
	at or below	25.000	4	5.6	.463
	25.000	31.000	14	8.1	4.263
	31.000	37.000	7	8.3	.195
above	37.000		3	6.0	1.500

Chisquare = 6.42099 with 1 d.f. Sig. level = 0.0112779

Jn 6.34

Tabelis on viis veergu, neist kaks esimest tähistavad klasside piire. Klassipiiride valikul ei jälgita tellija soovi siis, kui sel juhul jääksid klassid liialt tühjaks või kui klasse tekiks liialt vähe.

Käesoleval juhul on klasside arv 4, mis ongi minimaalne võimalik, nagu selgub.

Järgmises veerus on empiirilised sagedused n_i, seejärel teoreetilised sagedused np_i, kus n on valimi maht ja

$$p_i = F(a_{i+1}) - F(a_i),$$

F(x) on antud parameetritega normaaljaotuse jaotusfunktsioon ning kokkuleppeliselt $F(a_1) = 0$, $F(a_{k+1}) = 1$.

Viimases reas on nn hii-ruudud, mis arvutatakse valemist

$$\chi^2(i) = \frac{(n_i - np_i)^2}{np_i},$$

joone all aga on viimase veeru summa $\sum_{i=1}^{k} \chi^2(i)$. Vabadusastmete arv f arvutatakse seosest

$$f = k - 1 - r,$$

kus r on hinnatavate parameetrite arv. Seega käesoleval juhul f = 1.

Viimane tabeli all olev arv on olulisuse tõenäosus p, mis leitakse χ^2-jaotuse alusel (vastavalt vabadusastmete arvule f). Kui $p \leq \alpha$, siis võetakse vastu hüpotees H_1, seega võime antud juhul olulisuse nivooga 0,05 kinnitada, et tunnuse vanus jaotus erineb normaaljaotusest.

Meenutame, et seda näitasid (visuaalselt) ka kõik eelnevad protseduurid. Sama järeldus tulenes ka asümmeetria kordaja ja ekstsessi väärtustest (vt p 6.3.1).

Jätkame jaotuste sobitamise protseduuriga tutvumist ja valime aknast kolmanda rea - *K-S test* (Kolmogorov-Smirnovi test jaotuste võrdlemiseks), vt jn 6.35. Märgime, et Kolmogorov-Smirnovi testi kasutamine on õigustatud ainult pideva teoreetilise jaotuse P_3 korral.

```
Estimated KOLMOGOROV statistic DPLUS = 0.238647
Estimated KOLMOGOROV statistic DMINUS = 0.164685
Estimated overall statistic DN = 0.238647
Approximate significance level = 0.0823932
```

Jn 6.35

Kolmogorov-Smirnovi statistikuks on teoreetilise ja empiirilise jaotusfunktsiooni suurim erinevus $\Delta = DN$

Näeme, et meie näite korral on selle statistiku põhjal arvutatud olulisuse tõenäosuseks 0,082, seega ei õnnestu H_1 kummutada.

Kokkuvõttes see aga meie juba tehtud järeldust ei muuda, sest kui jaotus mingi kriteeriumi järgi erineb normaaljaotusest, siis on erinevus sellega juba kindlaks tehtud ning võimalikud vastupidised tulemused ei saa seda enam tühistada.

Järgmised kaks valikut aknast (*Tail areas* ja *Critical values*) on kirjeldatud raamatu [7] paragrahvis 3.4.

§ 6.6. Mitteparameetrilised testid tunnuse (jada) mõningate omaduste kohta

Ka eelmised jaotuste sobitamise testid (χ^2- ja Kolmogorov-Smirnovi test) kuuluvad mitteparameetriliste testide hulka. Käesolevas paragrahvis lisame neile veel mõned SG-süsteemis realiseeritud mitteparameetrilised testid ühe tunnuse/jada mõningate omaduste kontrollimiseks.

6.6.1. Härgitest ja Wald-Wolfowitzi test binaarse (kahe-elemendilise) jada kohta (*Tests for Binary Sequences*). Protseduur *Tests for Binary Sequences* on allmenüüs *R. Nonparametric Methods* esimene.

Selle protseduuri abil on võimalik

1) kahe väärtusega tunnuse puhul kontrollida hüpoteesipaari

$$H_0: p_1 = p_2 \ (= 0{,}5),$$
$$H_1: p_1 \neq p_2,$$

kus p_1 ja p_2 on vastavalt kummagi väärtuse tõenäosused (üldkogumis);

2) kontrollida, kas kahe-elemendilises jadas elementide järjestus on juhuslik.

Protseduuri tellimus algab tunnuse nime trükkimisega. Käivitamise järel väljastatakse tulemused kolme blokina (vt jn 6.36).

Tests for Binary Sequences

Data: sugu

Element types: 1 2

Number of elements of first type = 6
Number of elements of second type = 9
Expected number = 7.5
Large sample test statistic Z = 0.516398
Probability of equaling or exceeding Z = 0.605574

Number of runs = 8
Expected number = 8.2
Large sample test statistic Z = 0.168005
Two-tailed probability of equaling or exceeding Z = 0.866574

Jn 6.36

Esimeses blokis fikseeritakse tunnuse/jada väärtused (tekst: *Element types*).

Teine blokk annab märgitesti (vt [10]) tulemused. Siin esitatakse kõigepealt kummagi väärtuse sagedused n_1 ja n_2 (*Number of elemnts of first/second type*) ja oodatav ühine sagedus n/2 (*Expected number*).

Seejärel esitatakse teststatistiku Z väärtus (Z on asümptootiliselt normaaljaotusega),

$$Z = \frac{\left[\left|n_1 - \frac{n}{2}\right|\right]}{\sqrt{\frac{n}{4}}},$$

ning sellele vastav olulisuse tõenäosus p (mis on saadud normaaljaotuse tabelist). Z-statistik on tuletatud lihtsalt binoomjaotusest parameetritega n ja 0,5.

Kui $p \leq \alpha$, siis võetakse vastu sisukas hüpotees $p_1 \neq p_2$, st üldkogumis ei ole tunnuse väärtuste esinemistõenäosus võrdne.

Käesoleval juhul nullhüpoteesi kummutada pole võimalik, seega jääme oletuse juurde, et meeste (1) ja naiste (2) arvud üldkogumis on võrdsed.

Kolmas blokk on pühendatud Wald-Wolfowitzi jadade (*runs*) testi tulemuste esitamisele.

Kaht elementi sisaldav jada koosneb alati teatavast hulgast konstantsetest osajadadest e seeriatest (*runs*), olgu nende arv t. Sel juhul kui jada on juhuslik, saab (lähtudes kummagi elemendi esinemissagedustest n_1 ja n_2) leida ootuspärase osajadade arvu τ_0. Kui tegelikkuses on seeriate arv väiksem, pole jada juhuslik - elemente on valitud süstemaatiliselt. Ka siis, kui seeriate arv on liiga suur, pole jada juhuslik - sellisel juhul on tegemist mittejuhusliku võnkumisega.

Kontrollitav hüpoteesipaar on käesoleval juhul järgmine:

H_0: jada on juhuslikus järjestuses, $\tau = \tau_0$,

H_1: jada ei ole juhuslikus järjestuses, $\tau \neq \tau_0$.

Ootuspärane seeriate arv vaadeldava valimi jaoks τ_o arvutatakse valemist

$$\tau_o = \frac{2n_1n_2}{n} + 1$$

(vt [1], lk 289) ning selle normeerimisel saadakse asümptootiliselt normaaljaotusega statistik Z:

$$Z = \frac{t - \tau_o + 1/2}{s},$$

kus

$$s = \frac{1}{n}\sqrt{\frac{2n_1n_2(2n_1n_2-n)}{n-1}}$$

Käesoleva näite puhul on jada (vt [7], lk 30) järgmine:

2 2 1 2 1 1 2 2 2 2 2 2 1 1,

seega seeriate arv 8 on maksimaalselt lähedane oodatavale (8,2, vt jn 6.36), mistõttu ka teststatistiku Z väärtus on nullilähedane ja olulisuse tõenäosus suur. Vastu võetakse nullhüpotees: jada on juhuslikus järjestuses.

6.6.2. Juhuslikkuse testid (*Tests for Randomness*). Protseduur *Tests for Randomness* paikneb teisel kohal allmenüüs *R. Nonparametric Methods*.

Ka see protseduur on ette nähtud arv- või järjestustunnuse väärtuste jada juhuslikkuse kontrollimiseks. Protseduur koosneb kahest testist, kusjuures mõlemad kasutavad teststatistikuna tunnuse väärtuste jada kasvavate ja kahanevate osajadade arvu.

Tellimus koosneb ainult sisestatava tunnuse nimest. Tulemused esitatakse trükisena (vt jn 6.37), mis koosneb kahest osast.

Esimene test loendab kokku kõik sellised osajadad, mis paiknevad tervikuna kas allpool või ülevalpool tunnuse mediaani, ning arvutab nende arvu põhjal välja asümptootiliselt normaaljaotusega teststatistiku Z, mille väärtuse järgi leitakse ka olulisuse tõenäosus.

See test on tundlik andmetes oleva trendi suhtes.

Tests for Randomness

Data: sugu

Median = 2 based on 15 observations.
Number of runs above and below median = 8
Expected number = 8.2
Large sample test statistic Z = 0.168005
Two-tailed probability of equaling or exceeding Z = 0.866574

Number of runs up and down = 7
Expected number = 5
Large sample test statistic Z = 1.43019
Two-tailed probability of equaling or exceeding Z = 0.152661

NOTE: 7 adjacent values ignored.

Jn 6.37

Teine test kasutab asümptootiliselt normaaljaotusega statistiku Z arvutamiseks kasvavate ja kahanevate osajadade koguarvu, kusjuures konstantsetest seeriatest jäetakse alles ainult üks element. Seda testi soovitatakse kasutada sel juhul, kui võib oletada andmete mõjustatust pikaajaliste tsüklite poolt.

Trükisel on I blokis antud

- mediaani väärtus ja valimi maht,
- ülal- ja allpool mediaani paiknevate osajadade arv,
- selle arvu teoreetiline väärtus,
- esimese teststatistiku Z väärtus,
- olulisuse tõenäosus,

II blokis

- kasvavate ja kahanevate osajadade arv,
- selle arvu teoreetiline väärtus,
- teise teststatistiku Z väärtus,
- olulisuse tõenäosus.

Lõpus on ära märgitud (II testis) välja jäetud vaatluste arv.

6.6.3. Paiknemise test (*Tests for Location*). Protseduur *Tests for Location* on kolmas allmenüüs *R. Nonparametric Methods*.

See test on ette nähtud kontrollimaks hüpoteese mediaani kohta.

Testil on kaks varianti:

1) loendatakse mediaanist all- ja ülalpool paiknevad väärtused ja rakendatakse märgitesti;

2) Wilcoxoni test, mis võtab aluseks vaatlustulemuste ja oletatava mediaani vahed ning järjestab viimased.

Tuleb märkida, et variant 2) ei ole rakendatav järjestustunnuse puhul, sest kasutab väärtuste vahesid, need aga ei oma järjestustunnuste puhul üldiselt mõtet.

Tellimus koosneb käesoleval juhul tunnuse nimest ja oletatavast (etteantud) väärtusest med_o, mille alusel kontrollitakse hüpoteesipaari

$$H_o: med = med_o,$$
$$H_1: med \neq med_o,$$

kus med on üldkogumi mediaan.

Lisaks sellele tuleb tellimuses märkida meetod (valida kas *signs* või *ranks*).

Protseduuri tulemused on esitatud joonisel 6.38, nende tõlgendus on sarnane eelnevatele näidetele.

Tests for Location

Data: sugu

Hypothesized median: 1.5

Test based on: Signs

Sample median = 2
Number of values above hypothesized median = 9
Number of values below hypothesized median = 6
Expected number = 7.5
Large sample test statistic Z = 0.516398
Two-tailed probability of equaling or exceeding Z = 0.605574

NOTE: 15 observations. 0 values equal to hypothesized median ignored.

Jn 6.38

6.6.4. Empiirilise ja teoreetilise jaotuse võrdlemine χ^2-statistiku abil (*Chi-Square Goodness-of-Fit Statistic*). χ^2-jaotuse abil võrdlemise protseduur on kolmas allmenüüs *P. Categorical Data Analysis*.

Selle protseduuri abil saab antud empiirilist jaotust võrrelda mingi teoreetilise jaotusega. Loomulikult on sellel

protseduuril mõtet üksnes siis, kui vastav teoreetiline jaotus ei kuulu loetelusse, mis on antud protseduuris *Distribution Fitting*.

Selle protseduuri lähteandmeteks on

- empiirilised sagedused (sagedustabel),
- teoreetilised sagedused.

Lisaks sellele antakse ka minimaalne lubatav sagedus (mingisse klassi kuuluvate punktide arv).

Protseduuri käivitamise tulemusena saadakse tulemuste tabel, mis langeb ühte χ^2-testi tabeliga, mida kirjeldasime punktis 6.4.4.

6.6.5. Kolmogorov-Smirnovi test tunnuse jaotuse võrdlemiseks antud jaotusega (*Kolmogorov-Smirnov One-Sample Test*). Protseduur *Kolmogorov-Smirnov One-Sample Test* on allmenüüs *R. Nonparametric Methods* kuues. Selle protseduuri eesmärgiks on võrrelda tunnuse jaotust mingi ette antud jaotusega.

Loomulikult on seda protseduuri mõtet rakendada vaid siis, kui etteantud jaotus ei kuulu protseduuris *Distribution Fitting* loetletud 18 teoreetilise tüüpjaotuse hulka.

Protseduuri tellimisel on tarvis sisestada

- kontrollitava tunnuse nimi,
- antud jaotuse jaotusfunktsiooni väärtused vastavalt tunnuse väärtustele (neid on võimalik esitada ka SG-avaldise abil).

Nõutav on, et jaotusfunktsiooni väärtuste hulk võrduks kontrollitava tunnuse pikkusega.

Protseduuri käigus kontrollitakse hüpoteesipaari

$$H_0: P_x = P,$$
$$H_1: P_x \neq P,$$

kus P_x on uuritava tunnuse jaotus, P aga etteantud jaotus.

Protseduuri tulemusena väljastatakse

- Kolmogorov-Smirnovi teststatistikud (antud jaotusfunktsiooni ja tunnuse empiirilise jaotusfunktsiooni vahe ekstremaalsed väärtused;
- olulisuse tõenäosus p.

Kui $p \leq \alpha$, siis loetakse hüpotees H_1 tõestatuks.

Täiendavalt on võimalik saada ka graafik, millel on kujutatud mõlemad jaotusfunktsioonid.

Näitena võrdleme tunnuse kasv jaotust jaotusfunktsiooniga, mille sisestame üksikväärtuste kaudu (vt jn 6.39), kasutades käsurežiimi (vt [7], p 4.1.7; paneme tähele, et selle funktsiooni F(i) väärtused sõltuvad ainult indeksist i, mitte aga jaotusfunktsiooni argumendist x.

```
: EMPJAOT GETS .00 .01 .02 .03 .04 .05 .06 .08 .1 .15 .16 .19 .2 .3 .36 .37
.4 .45 .5 .51 .55 .59 .6 .65 .7 .8 .9 .99
:
```

Jn 6.39

Joonisel 6.40 on kujutatud tellimus (kolm esimest rida) ja samuti ka tulemus - teststatistikud

$$DPLUS = \max_{1 \le i \le 28} (F_x(i) - F(i)),$$

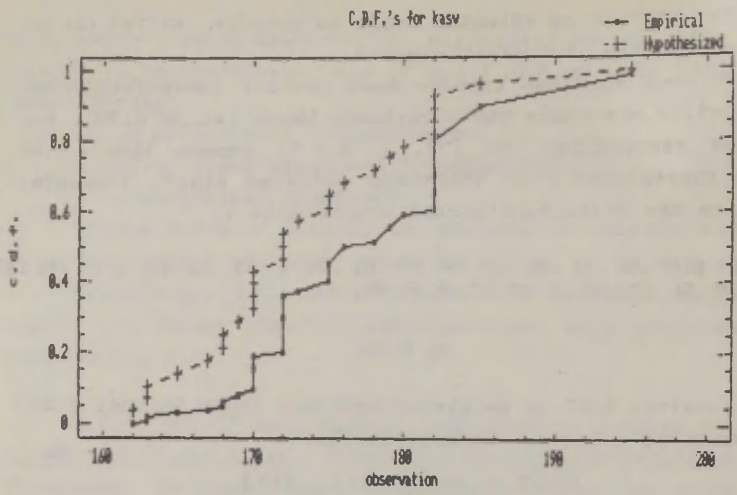
$$DMINUS = \max_{1 \le i \le 28} (F(i) - F_x(i)),$$

$$DN = \max (DPLUS, DMINUS)$$

ja olulisuse tõenäosus p. Selle väärtusest järeldub, et olulisuse nivool 0,05 saame lugeda tõestatuks sisuka hüpoteesi: tunnuse kasv jaotus erineb oluliselt antud jaotusest. Viimast väidet illustreerib ka graafik joonisel 6.41.

```
                    Kolmogorov-Smirnov One-Sample Test
-------------------------------------------------------------------

Data: kasv

Empirical c.d.f.: EMPJAOT


Estimated KOLMOGOROV statistic DPLUS = 0.264286
Estimated KOLMOGOROV statistic DMINUS = 0.0257143
Estimated overall statistic DN = 0.264286
Approximate significance level = 0.0400235
```

Jn 6.40


C.D.F.'s for kasv
Empirical
Hypothesized
c.d.f.
1
0.8
0.6
0.4
0.2
0
160
170
180
190
200
observation

Jn 6.41

7. KAHE TUNNUSE ANALÜÜS

§ 7.1. Kahe tunnuse analüüsi eesmärk, põhimõisted ja SG võimalused

Mitme (sh ka kahe) tunnuse analüüsimise juures on oluline tunnustevaheliste seoste kirjeldamine ja hindamine, kusjuures lõppeesmärgiks on enamasti statistilise mudeli konstrueerimine, et ühe tunnuse käitumist teiste abil kirjeldada.

Mõningatel juhtudel sisaldavad niisugused mudelid ainult kaht tunnust - sel juhul viib kahe tunnuse analüüs uurimise lõppeesmärgini. Tavalisem on aga olukord, kus uuritavaid tunnuseid on märksa rohkem, kahe tunnuse analüüs aga moodustab uurimistöös olulise vaheetapi, mille käigus tehakse kindlaks potentsiaalsete argumentide mõju oletatavale funktsioontunnusele, samuti argumentide omavaheline koosmõju, jäetakse kõrvale mõjutud tunnused jne.

7.1.1. Kahe tunnuse ühisjaotus. Kahe tunnuse (olgu nad edaspidi **X** ja **Y**) ühist käitumist kirjeldab ühisjaotus, mida tähistame sümboliga P_{XY}. Kui tunnused on pidevad, siis esitab ühisjaotust tihedusfunktsioon $f_{XY}(x,y)$. Kui mõlemad tunnused on diskreetsed, siis kirjeldab ühisjaotust tõenäosusfunktsioon $\{p_{ij};\ i=1, \ldots, k;\ j=1, \ldots, h\}$, kus $p_{ij} = P(\mathbf{X}=x_i, \mathbf{Y}=y_j)$, x_i ja y_j on tunnuste **X** ja **Y** väärtused ning k ja h vastavate väärtuste arvud. Tõenäosusfunktsioon kirjeldab ka mitteарvuliste tunnuste ühisjaotust, sel juhul on tunnuste väärtused $x_1, \ldots, x_k$ ja $y_1, \ldots, y_h$ (või ka ainult üks neist jadadest) mitteарvulised.

Kui üks tunnus (olgu see näiteks **X**) on diskreetne, teine (**Y**) aga pidev, siis kirjeldab ühisjaotust tunnuse **Y** tinglike tihedusfunktsioonide hulk $\{f_{Y/i}(y):\ i=1, \ldots, k\}$.

7.1.2. Kahemõõtmeline valimjaotus. Valimi põhjal kahe tunnuse ühisjaotuse kohta saadavat teavet annab kõige täielikumalt edasi kahe tunnuse empiiriline ühisjaotus.

Hästi kasutatav on kahe tunnuse empiiriline ühisjaotus siis, kui

- tunnused X ja Y on diskreetsed ja
- tunnuste väärtuste arvud k ja h on küllalt väikesed, võrreldes valimi mahuga n, nii et on täidetud tingimus

$$kh < n.$$

Sel korral on empiiriline jaotus esitatav jaotustabelina, milles antakse hinnang tõenäosusfunktsiooni p_{ij} üksiktõenäosustele,

$$p_{ij} = \frac{n_{ij}}{n}, \tag{7.1}$$

kus n_{ij} on valimi nende objektide arv, mille puhul tunnuse X väärtus on x_i ja Y väärtus y_j.

Pidevate tunnuste ühisjaotuste kirjeldamiseks kasutatakse mitmesuguseid graafilisi vahendeid; loomulikult on võimalik ka pidevaid tunnuseid diskreetsetega sobivalt lähendada.

7.1.3. Tunnustevaheline statistiline sõltuvus ja selle kirjeldamine. Kahe tunnuse ühisjaotus sisaldab eneses üldiselt rohkem informatsiooni kui nende tunnuste jaotused eraldi võetuna. On aga ka mõningaid erandeid.

1° Kui tunnused on statistiliselt sõltumatud, siis kehtib seos

$$P_{XY} = P_X P_Y. \tag{7.2}$$

Seega on sõltumatute jaotuste korral ühisjaotus lähtetunnuste jaotustega täielikult määratud. Ühe tunnuse konkreetse väärtuse teadmine ei lisa mingit informatsiooni teise tunnuse käitumise kohta.

2° Kui ühe tunnuse (Y) iga väärtus on teise tunnuse (X) vastava väärtuse järgi täpselt arvutatav,

$$y = \phi(x), \tag{7.3}$$

siis me ütleme, et tunnus **Y** sõltub täielikult (funktsionaalselt) tunnusest **X**, ning sel juhul on üldiselt niihästi tunnuse **Y** jaotus P_Y kui ka ühisjaotus P_{XY} tunnuse **X** jaotuse P_X järgi täpselt arvutatavad.

3° Kui tunnused ei ole statistiliselt sõltumatud, st seos (7.2) ei kehti, siis nimetatakse tunnuseid statistiliselt sõltuvateks.

Statistiline sõltuvus on kõige üldisem tunnustevahelise sõltuvuse vorm, see omab mõtet kõikvõimalike tunnusetüüpide korral. Statistilise sõltuvuse erijuhtudeks, mida järgnevas lühidalt vaatleme, on

- monotoonne sõltuvus,
- regressioonsõltuvus,
- korrelatiivne sõltuvus.

Funktsionaalset sõltuvust võib käsitleda statistilise sõltuvuse teatava piirjuhuna.

7.1.4. Monotoonne sõltuvus. Olgu **X** ja **Y** järjestus- või arvtunnused, ning olgu (x_i, y_i) ja (x_j, y_j) kaks suvalist üldkogumi objekti.

Kui asjaolust

$$x_i < x_j \tag{7.4}$$

järeldub, et

$$P(y_i < y_j) > P(y_i > y_j), \tag{7.5}$$

siis öeldakse, et tegemist on monotoonselt kasvava sõltuvusega; kui aga võrratusest (7.4) järeldub

$$P(y_i > y_j) > P(y_i < y_j), \tag{7.5'}$$

siis on tunnuspaari **X**, **Y** vahel monotoonselt kahanev sõltuvus.

Kui tunnuste vahel on kas monotoonselt kasvav või monotoonselt kahanev sõltuvus, siis öeldakse, et tunnused on monotoonselt sõltuvad.

Tunnused, mis ei ole monotoonselt sõltuvad, on monotoonselt sõltumatud, kuid sellest, et tunnuste vahel puudub monotoonne sõltuvus, ei järeldu statistiline sõltumatus.

Küll aga järeldub statistilisest sõltumatusest, st võrdusest (7.2) igasugune, sh ka monotoonne sõltumatus.

7.1.5. Regressioonsõltuvus. Olgu tunnused X ja Y arvtunnused.

Igale tunnuse X väärtusele x vastab üldiselt tunnuse Y tinglik jaotus $P_{Y/x}$, mis võib olla üldiselt pidev või diskreetne nagu tunnuse Y enda jaotuski. Tähistame tingliku jaotuse $P_{Y/x}$ järgi arvutatud keskväärtuse sümboliga EY/x; üldiselt sõltub tinglik keskväärtus tingimust määravast tunnuse X väärtusest x. Tinglikku keskväärtust nimetatakse ka tunnuse Y regressioonifunktsiooniks tunnuse X järgi.

Kui

$$EY/x = \text{const} \tag{7.6}$$

kõigi X väärtuste korral, siis tunnus Y ei sõltu tunnusest X regressiooni mõttes. Kui aga tingimus (7.6) ei ole täidetud, siis on tunnus Y regressioonsõltuvuses tunnusest X.

Regressioonsõltuvus on statistilise sõltuvuse alaliik; statistilise sõltuvuse puudumisest järeldub ka regressioonsõltuvuse puudumine, kuid vastupidine seos üldiselt ei kehti.

7.1.6. Korrelatiivne sõltuvus. Korrelatiivne sõltuvus on regressioonsõltuvuse erijuht, mis on defineeritud kahe arvtunnuse vahel. Korrelatiivne sõltuvus leiab aset siis, kui tinglik keskväärtus EY/x on (vähimruutude mõttes) lähendatav argumendi X lineaarse funktsiooniga,

$$EY/x \approx a + bX, \tag{7.7}$$

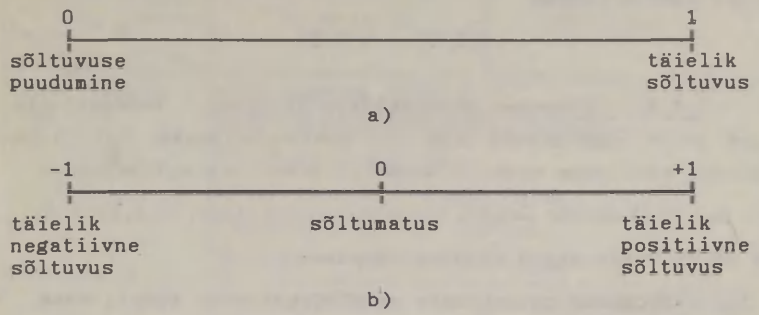
kusjuures nõutav on, et regressioonikordaja b erineks nullist,

$$b \neq 0. \tag{7.8}$$

Korrelatiivne sõltuvus on ühtaegu nii regressioonsõltuvus (sest täidetud pole tingimus (7.6)) kui ka monotoonne sõltuvus (kui $b > 0$, siis monotoonselt kasvav, kui $b < 0$, siis monotoonselt kahanev). Samuti järeldub korrelatiivse sõltuvuse olemasolust statistiline sõltuvus tunnuste vahel.

7.1.7. Sõltuvuse tugevuse mõõtmine. Tunnustevahelise sõltuvuse tugevust mõõdavad mitmesugused seosekordajad (tähistame neid üldiselt sümboliga $\tau(X,Y)$), mille konkreetne väärtus iseloomustab vastava seose tugevust: mida suurem on

seosekordaja väärtus (absoluutväärtus), seda tugevam on sõltuvus. Sõltuvuse tugevuse mõõtmiseks kasutatakse üht kahest järgmisest skaalatüübist (vt jn 7.1).



Jn 7.1. Statistilise sõltuvuse skaalad

Skaala a) vastab üldisele statistilisele ja regressioonsõltuvusele, skaala b) - monotoonsele ja korrelatiivsele sõltuvusele.

Täielik sõltuvus tähendab üldjuhul funktsionaalset sõltuvust, mille korral tunnuse Y väärtused on tunnuse X väärtustega üheselt määratud.

Valimi põhjal arvutatakse seosekordajale valim- (e empiiriline) väärtus, mida üldjuhul tähistame sümboliga $t(X,Y)$.

7.1.8. Sõltuvuse suund. Seni käsitlesime tunnust X argument- (sõltumatu) tunnusena, tunnust Y funktsioon- e sõltuva tunnusena.

Enamasti on võimalik (vähemalt formaalselt) ka nimetatud tunnuste rollid vastupidiseks muuta. Üldjuhul muutub siis ka seose tugevus.

Kui ühel uuritavatest tunnustest (olgu see X) on palju väärtusi, teisel (Y) - vähe väärtusi, siis üldjuhul saab tunnuse X järgi tunnust Y prognoosida märksa edukamalt kui vastupidi. Sel juhul öeldakse ka, et tunnuse Y sõltuvus tunnusest X on tugevam kui tunnuse X sõltuvus tunnusest Y.

Tunnuse Y sõltuvust tunnusest X mõõtvat seosekordajat tähistame sümboliga $\tau(Y/X)$, vastassuunalist sõltuvust mõõtvat seosekordajat vastavalt sümboliga $\tau(X/Y)$.

Mõned seosekordajad on defineeritud ka sümmeetriliste-na, nende tähistuseks kasutame üldsümbolit $\tau(X,Y)$.

Märgime, et korrelatiivne sõltuvus ei olene suunast, alati kehtib võrdus

$$\tau(X/Y) = \tau(Y/X).$$

7.1.9. Sõltuvuse statistiline olulisus. Tunnustevahelist seost nimetatakse statistiliselt oluliseks, kui (kokku lepitud olulisuse nivool) õnnestub kummutada nullhüpotees

H_0: üldkogumis puudub tunnuste vahel seos, $\tau(X,Y) = 0$

ja seega võtta vastu sisukas hüpotees

H_1: üldkogumis on tunnuste vahel (uuritavat tüüpi) seos, $\tau(X,Y) \neq 0$.

Konkreetse valimi puhul iseloomustab nullhüpoteesi tõepärasust olulisuse tõenäosus p, mille väärtus annab aluse otsuse langetamiseks uuritava sõltuvuse olulisuse kohta.

Kui seosekordaja on seotud teatava statistilise mudeliga (eriti on see nii regressioon- ja korrelatiivse sõltuvuse korral), siis järeldub vastava seose olulisusest antud tüüpi mudeli olulisus. Seose mitteolulisusest tuleneb, et vastava mudeli kasutamine uuritava nähtuse kirjeldamiseks, st ühe tunnuse prognoosimiseks teise järgi, ei ole korrektne.

7.1.10. Kahe tunnuse analüüsi põhilised meetodid süsteemis SG. Kahe tunnuse analüüsimisel rakendatavad meetodid olenevad sellest, millisesse tüüpi kuuluvad analüüsitavad tunnused.

Tabelis 7.1 ongi ära toodud kahe tunnuse analüüsimisel kasutatavad statistikameetodid ja 20 neid realiseerivat SG-protseduuri; mõned neist (*Crosstabulation, Contingency Tables, Log-linear Analysis*) on kasutatavad ka rohkem kui kahe tunnuse korral, protseduuri *Barcharts* aga saab kasutada ka ühe tunnuse analüüsimisel.

Kommenteerimist vajab tabeli päis - klassid "Diskreetne" ja "Arvuline" ei välista teineteist. Siit järeldubki, et diskreetsete väärtustega arvtunnus mahub mõlemasse lahtrisse

ning seetõttu on niisuguse tunnuse töötlusvõimalused kõige avaramad.

T a b e l 7.1

Funktsioon-tunnus / Argu-menttunnus	Diskreetne	Arvuline
Kahe väärtusega	KAHE JAOTUSE VÕRDLUS KAHE JAOTUSE PARAMEETRITE VÕRDLUS *Two-Sample Analysis* *COmparison of Poisson Rates* *Comparison of Two Samples* *Kolmogorov-Smirnov Two-sample Test*	
Diskreetne	SAGEDUSTABELID SEOSEKORDAJAD TULPDIAGRAMMID *Barcharts* *Three-Dimensional Histogram* *Median Polish of Two-Way Table* *Chi-Square Goodness-of-Fit Statistic* *Crosstabulation* *Contingency Tables* *Log-Linear Analysis* *Rank Correlation Coefficients*	TINGLIKUD PARAMEETRID ÜHEFAKTORILINE DISPERSIOONANALÜÜS *Codebook Procedure* *Multiple Box-and-Whisker Plot* *Notched Box-and-Whisker Plot* *One-Way Analysis of variance* *Kruskal-Wallis One-Way Analysis by Ranks*
Arvuline (pidev)		KORRELATSIOONIVÄLI REGRESSIOONANALÜÜS *X-Y Line and Scatterplots* *Simple Regression* *Interactive Outlier Rejection*

Enamiku diskreetsete tunnuste töötlemisel ei nõuta nende arvulisust või järjestatust, kuid on ka mõningaid erandeid, millele me viitame vastava meetodi kirjeldamisel.

Olgu veel märgitud, et tabeli 7.1 kõikides lahtrites paiknevate protseduuride hulgas on niihästi kirjeldavaid (sh graafilisi), hindavaid, hüpoteese genereerivaid kui ka hüpoteeside kontrollimiseks ette nähtuid.

§ 7.2. Kahe tunnuse ühisjaotust ja seosekordajaid analüüsivad protseduurid süsteemis SG

Käesolevas paragrahvis vaatleme me nelja kõige põhilisemat protseduuri kahe diskreetse tunnuse ühisjaotuse analüüsimiseks, kirjeldamiseks ja illustreerimiseks. Täiendavaid võimalusi diskreetse tunnuspaari jaotuse süvitsi analüüsimiseks käsitleme veel paragrahvis 7.6.

7.2.1. Kahe tunnuse ühisjaotus (*Crosstabulation*). Protseduur *Crosstabulation* on esimene allmenüüs *P. Categorical Data Analysis*.

See protseduur on ette nähtud kahe (või enama) diskreetse tunnuse ühisjaotuse moodustamiseks.

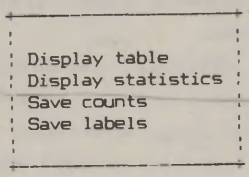
Tellimuse esitamisel tuleb tellimuspaneelil märkida (vähemalt) kahe tunnuse nimed vastavalt ridadele *Data vector A* ja *Data vector B* ning reale *Percentages* valida üks kolmest võimalusest *Columnwise*, *Rowwise*, *Tablewise*. Viimane valik määrab jaotuse tüübi (esimese tunnuse tinglik jaotus horisontaalselt paikneva, teisena nimetatud tunnuse suhtes, teise tunnuse tinglik jaotus vertikaalselt paikneva, esimesena nimetatud tunnuse suhtes, tunnuste tingimatu ühisjaotus). Täidetud tellimuspaneel on kujutatud joonisel 7.2.

Märgime, et joonisel 7.2 on tunnused valitud võtme F7 abil üldloetelust, seetõttu on tunnuse nime ees andmestiku nimi. Üldiselt pole andmestiku nime trükkimine tarvilik (vt [7] § 2.6 ja p 2.5.1), see ainult halvendab trükiste loetavust, vrdl jn 7.4 ja 7.6.

Peale tellimuspaneeli täitmist ja protseduuri käivitamist võtmega F6 ilmub ekraanile aken (jn 7.3), mis võimaldab valida väljundi. Jaotuste saamiseks on tarvis teha esimene valik - *Display table*.

```
Data vector A: ASPIRANT.synnipaik
Labels:
Data vector B: ASPIRANT.eriala
Labels:
Data vector C:
Labels:
Data vector D:
Labels:
Data vector E:
Labels:
Data vector F:
Labels:
Data vector G:
Labels:
Data vector H:
Labels:
Data vector I:
Labels:

Percentages: Tablewise
```

Jn 7.2



Jn 7.3.

Statistilises andmetöötluses kasutatakse kahe tunnuse ühiskäitumise kirjeldamiseks nelja erinevat tabelit - üht sagedus- ja kolme jaotustabelit. SG-protseduur väljastab standardselt ühes ruudustikus kaks tabelit, neist esimene (ülemised numbrid) on alati sagedustabel, teine aga jaotustabel, mille tüüp valitakse tellimuse esitamisel (*Percentages*).

Kirjeldame neid järjest, käsitledes kõigepealt sagedustabelit. See on tabel, milles on k rida ja h veergu, kus k ja h on vastavalt esimese ja teise tunnuse väärtuste arvud, vt jn 7.4.

Tabel sisaldab sagedusi n_{ij}, st valimi nende objektide arve, millel esimese tunnuse (veerutunnuse) väärtuseks on x_i

Crosstabulation of ASPIRANT.synnipai

ASPIRANT.e / ASPIRANT.sy	fil	mat	med	Row Total
alev	1	1	3	5
	3.6	3.6	10.7	17.9
linn	2	5	1	8
	7.1	17.9	3.6	28.6
maa	2	2	1	5
	7.1	7.1	3.6	17.9
suurlinn	5	0	5	10
	17.9	.0	17.9	35.7
Column Total	10	8	10	28
	35.7	28.6	35.7	100.0

Jn 7.4

(mis paikneb i-ndas reas) ja teise tunnuse (reatunnuse) väärtuseks on y_j (mis paikneb j-ndas veerus).

Sageduste summeerimisel saame üksiktunnuste sagedused ehk nn marginaalsagedused:

1) teise tunnuse väärtuste sagedused $n_{.1}, \ldots, n_{.h}$,

$$n_{.j} = \sum_{i=1}^{k} n_{ij}, \quad j = 1, \ldots, h, \tag{7.9}$$

mis paiknevad tabeli all (k+1)-ses reas sõnade *Column Total* järel, vt jn 7.4;

2) esimese tunnuse väärtuste sagedused $n_{1.}, \ldots, n_{k.}$,

$$n_{i.} = \sum_{j=1}^{h} n_{ij}, \quad i = 1, \ldots, k,$$

mis paiknevad tabeli viimases, (h+1)-ses veerus *Row Total*;

3) valimi maht n,

$$n = \sum_{i=1}^{k} \sum_{j=1}^{h} n_{ij} = \sum_{i=1}^{k} n_{i.} = \sum_{j=1}^{h} n_{.j},$$

mis paikneb tabeli paremposlses alumises lahtris. Olgu märgitud, et see on nn kahe tunnuse ühisvalimi maht, millesse on loetud ainult need objektid, millel/kellel on mõõdetud mõlemad uuritavad tunnused.

Sagedustabeli põhjal saab leida kolm erinevat jaotustabelit, mida kirjeldame järgmises punktis.

7.2.2. Kahe tunnuse jaotustabelid. Arvutades kogu tabeli jaoks suhtelised sagedused

$$p_{ij} = \frac{n_{ij}}{n} ,$$

saame kahe tunnuse ühisjaotuse. Ühisjaotus avaldatakse tavaliselt protsentides ning selle puhul on õige võrdus

$$\sum_{i=1}^{k} \sum_{j=1}^{h} p_{ij} = 100\ \% .$$

Protsentides arvutatakse ka marginaaljaotused. Ühisjaotuse tabel esitatakse siis, kui tellimuses on real *Percentages* märgitud sõna *Tablewise*, vt jn 7.2 ja 7.4.

Arvutades iga veeru jaoks suhtelised sagedused

$$\frac{n_{ij}}{n_{.j}} ,$$

saame tabeli igasse veergu esimese tunnuse tingliku jaotuse tingimusel $y = y_j$, st esimese tunnuse jaotuse nende objektide jaoks, millel teise tunnuse väärtus on y_j. Sel juhul on iga veeru suhteliste sageduste summa 100 %. See tabel saadakse, valides tellimuspaneelil sõna *Columnwise*, vt jn 7.5.

Crosstabulation of ASPIRANT.synnipai

ASPIRANT.e / ASPIRANT.sy	fil	mat	med	Row Total
alev	1	1	3	5
	10.0	12.5	30.0	17.9
linn	2	5	1	8
	20.0	62.5	10.0	28.6
maa	2	2	1	5
	20.0	25.0	10.0	17.9
suurlinn	5	0	5	10
	50.0	.0	50.0	35.7
Column Total	10	8	10	28
	35.7	28.6	35.7	100.0

Jn 7.5

Joonisel 7.5 esitatud tabelist saame filoloogia-, matemaatika ja meditsiiniaspirantide jaotused sünnipaiga järgi: näeme, et filoloogide ja meedikute hulgas on suurlinnast pärinejaid 50 %, matemaatikute hulgas aga suurlinlased puuduvad hoopis, seevastu linlasi on 62,5 %.

Märgime ühtlasi, et tabelite moodustamisel järjestatakse nominaaltunnuse väärtused tähestiku järjekorras, arvulise või arvuliseks kodeeritud tunnuse väärtused - kasvavalt.

Arvutades iga rea jaoks suhtelised sagedused

$$\frac{n_{i\,j}}{n_{i\,.}} ,$$

saame tabeli igasse ritta teise tunnuse tingliku jaotuse tingimusel $X = x_i$, st teise tunnuse tingliku jaotuse nende objektide jaoks, millel esimese tunnuse väärtus on x_i. Sel korral on iga rea suhteliste sageduste summa 100 %. See tabel saadakse, valides *Percentages: Rowwise*, vt jn 7.6.

Crosstabulation of synnipaik by eria

eriala synnipaik	fil	mat	med	Row Total
alev	1 20.0	1 20.0	3 60.0	5 17.9
linn	2 25.0	5 62.5	1 12.5	8 28.6
maa	2 40.0	2 40.0	1 20.0	5 17.9
suurlinn	5 50.0	0 .0	5 50.0	10 35.7
Column Total	10 35.7	8 28.6	10 35.7	28 100.0

Jn 7.6

Sellest tabelist näeme, et alevist pärinevatest aspirantidest on 60 % meditsiinihuvilised, linnadest pärinevatest aspirantidest aga 62,5 % spetsialiseerunud matemaatikale.

Suurte k ja h väärtuste korral ei mahu sagedus- ja jaotustabel korraga ekraanile, vaid esitatakse mitme ekraanitäiena.

Ekraani alumises parempoolses nurgas on märgitud, mitmel ekraanitäiel tabel paikneb ja milline neist asub parasjagu ekraanil; siinjuures tuleb olla tähelepanelik, et mitte kaotada (unustada vaatamast) osa tabelist.

7.2.3. χ^2-statistik tunnustevahelise statistilise sõltuvuse kontrollimiseks. Protseduur *Crosstabulation* võimaldab arvutada ka terve seeria seosekordajaid. Selleks tuleb aknast (vt jn 7.3) valida rida *Display statistics*. Selle tulemusena ilmub ekraanile tabel *Summary Statistics for Contingency Tables*, vt jn 7.7.

Summary Statistics for Contingency Tables (Page 1)

Chi-square	D.F.	Significance
10.5875	6	0.101993

WARNING: Expected values in 12 cells < 5 and 6 cells < 2.

Statistic	Symmetric	With rows dependent	With columns dependent
Lambda	0.27778	0.27778	0.27778
Uncertainty Coeff.	0.18941	0.17192	0.21085
Somer's D	-0.10642	-0.11154	-0.10175
Eta		0.34857	0.23717

Jn 7.7

Esmalt esitatakse informatsioon χ^2-statistiku kohta. See statistik arvutatakse avaldisest

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{h} \frac{(n_{ij} - n_{i.}n_{.j}/n)^2}{(n_{i.}n_{.j})/n} . \qquad (7.10)$$

Ta on asümptootiliselt χ^2-jaotusega (vt [7], p 3.2.4), kusjuures vabadusastmete arvuks on $f = (k-1)(h-1)$.

χ^2-statistik sobib hüpoteeside kontrollimiseks statistilise sõltuvuse olemasolu kohta:

$$H_0: P_{XY} = P_X P_Y,$$

$$H_1: P_{XY} \neq P_X P_Y,$$

kuid testi rakendamine on lubatav ainult sel juhul, kui kõigi lahtrite korral on täidetud tingimus

$$n_{i.}n_{.j}/n \geq 2, \quad i = 1, \ldots, k, \quad j = 1, \ldots, h, \qquad (7.11)$$

ning igati korrektne siis, kui

$$n_{i.}n_{.j}/n \geq 5. \qquad (7.12)$$

Trükise (jn 7.7) esimeses blokis väljastataksegi χ^2-statistiku väärtus (päis: *Chi-square*), tema vabadusastmete arv (*D.F.*) ja olulisuse tõenäosus (*Significance*), kuid alati, kui tingimus (7.12) ei ole täidetud, trükitakse lisaks ka hoiatus (*Warning*), milles märgitakse, mitmes tabeli lahtris (*cells*) suurus $n_{i.}n_{.j}/n$, nn teoreetiline sagedus (*Expected values*) ei rahulda tingimust (7.11) ja/või (7.12).

7.2.4. Seosekordajad. Sama trükise järgmises blokis on antud rea seosekordajate väärtused tabelina, mille esimene veerg (*Statistic*) tähistab seosekordaja nimetust, teine (*Symmetric*) - vastava seosekordaja sümmeetrilist versiooni (kui see eksisteerib), kolmas (*With rows dependent*) - seosekordaja niisugust versiooni, kus sõltuvaks tunnuseks Y loetakse horisontaalselt (ridades) paiknev tunnus. Neljas veerg (*With columns dependent*) sisaldab seosekordaja sellist versiooni, kus sõltuv on veerutunnus

Käsitletavad seosekordajad on järgmised.

1° Esimeses reas on nn Guttmani λ kordaja, mis mõõdab üldist statistilist seost tunnuste vahel.

2° Teises reas on nn määramatuse (*uncertainty*) kordaja, mis samuti mõõdab üldist statistilist seost tunnuste vahel.

Märgime, et nende seosekordajate olulisuse kontrollimiseks SG-süsteem üldiselt vahendeid ei paku. Kuid kuivõrd χ^2-statistik mõõdab üldse statistilise sõltuvuse olemasolu, saab järeldusi nende seosekordajate olulisuse kohta teha χ^2 statistiku olulisustõenäosuse põhjal (kui $p \leq \alpha$, siis seos on oluline). Muidugi kehtib see siis, kui χ^2 statistiku kasutamise eeldused on täidetud, st on täidetud vähemalt tingimus (7.11) (pole trükitud hoiatust selle mittetäidetuse kohta).

3° Kolmandas reas on trükitud Somersi D-kordaja, mis mõõdab tunnustevahelise monotoonse seose olemasolu. Siin tuleb tähele panna niihästi seosekordaja suurust kui ka märki (positiivse seosekordaja puhul üldreeglina ühe tunnuse suurenedes suureneb ka teine ning ühe tunnuse vähenedes väheneb ka teine, negatiivse seosekordaja puhul ühe tunnuse suurenedes üldiselt teine tunnus väheneb).

Selle seosekordaja olulisust kontrollida ei saa.

4° Viimases reas on regressioonisuhted $\eta(X/Y)$ ja $\eta(Y/X)$, mis mõõdavad regressiooniseose tugevust. Ka nende kohta ei ole olulisuse tõenäosusi välja trükitud. Tegelikult on siiski üsna lihtne kontrollida hüpoteesipaari

H_0: tunnus Y ei sõltu tunnusest X regressiooni mõttes, st
$EY/x_i = \text{const}, \quad i = 1, \ldots, k,$

H_1: tunnus Y sõltub tunnusest X regressiooni mõttes, st
$EY/x_i \neq \text{const}, \quad i = 1, \ldots, k$

F-jaotuse tabeli abil. Selleks arvutatakse välja järgmine statistik

$$F = \frac{\eta^2(n-k)}{(1-\eta^2)(k-1)}$$

ning võrreldakse seda F-tabeli kriitilise väärtusega vabadusastmete arvude k-1 ja n-k korral (k on tunnuse X väärtuste arv) soovitaval olulisuse nivool α (näiteks 0,05). Vastava kriitilise väärtuse F_α võib leida, kasutades **SG**-protseduuri *Critical Values* (vt [7], p 3.4.2) ja võttes parameetriks (tekstis *Area at or below* ... võrdusmärgile järgnev arv) suuruse $1-\alpha$.

Kui $F \geq F_\alpha$, siis on regressioonsõltuvus statistiliselt oluline.

7.2.5. Täiendavad seosekordajad. Osa seosekordajaid on paigutatud väljundi 2. leheküljele, mille saamiseks tuleb pärast 1. leheküljе analüüsi lõpetamist vastavalt ekraanil olevale õpetusele vajutada klahvi **ENTER**. Ekraanile ilmub sama pealkirjaga tabel, küsjuures erinevuseks on vaid see, et leheküljе number (*Page*) on nüüd 2. Tabelis on nüüd kolm veergu - kordaja nimetus (*Statistic*), tema väärtus

(*Value*) ja olulisustõenäosus (*Significance*). Kahjuks ei paku programm aga kaugeltki mitte kõikide seosekordajate jaoks olulisuse tõenäosusi.

Esitatavad seosekordajad on järgmised (vt jn 7.8).

Summary Statistics for Contingency Tables (Page 2)

Statistic	Value	Significance
Contingency Coeff.	0.52381	
Cramer's V	0.43481	
Conditional Gamma	-0.14286	
Pearson R	-0.11225	
Kendall's Tau B	-0.10653	0.52324
Kendall's Tau C	-0.11097	

Jn 7.8

1° Kontingentsuse kordaja (*Contingency Coeff.*), mis mõõdab statistilise seose tugevust; tema väärtused on üldjuhul suuremad kui teistel kordajatel ja võivad omandada ka ühest suuremaid väärtusi. See aga ei tähenda, nagu mõõdaks see kordaja mingeid seose erilisi omadusi.

2° Crameri seosekordaja V, mis samuti mõõdab üldise statistilise seose tugevust.

Mõlema nimetatud seosekordaja olulisus tuleneb vahetult χ^2 olulisusest.

3° Goodman-Kruskali γ (*Conditional Gamma*), mis mõõdab monotoonse seose tugevust; tema jaoks puudub olulisuse tõenäosus.

4° Lineaarne korrelatsioonikordaja r (*Pearson R*), mis mõõdab korrelatiivse seose tugevust. Korrelatiivse sõltuvuse olulisuse kontrollimiseks võib kasutada suurust

$$t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

mis on t-jaotusega vabadusastmete arvuga n-2 (vt [7], p 3.2.5); kui $t_{n-2} \geq t_{\alpha}$ (kriitiline väärtus), siis on tunnustevaheline korrelatiivne sõltuvus statistiliselt oluline.

5° Kendalli τ_B mõõdab monotoonset sõltuvust, selle jaoks on arvutatud ka olulisuse tõenäosus.

6° Kendalli τ_C mõõdab samuti monotoonset sõltuvust.

Programm väljastab monotoonse, regressioon- ja korrelatiivse seose kordajad ka sümboltunnuste puhul, samuti arvudeks kodeeritud nominaaltunnuste korral. Seetõttu peab tellija ise kõrvale jätma sisutud seosekordajad ning kasutama üksnes neid, mille rakendamine vaadeldavate tunnusetüüpide korral on lubatud (vt § 7.1).

Olgu märgitud, et juhul kui tabel on "tühi", st lahtrite arv on ligikaudu võrdne objektide arvuga või sellest suuremgi, on tabel ise ja samuti statistilist seost mõõtvad näitajad (χ^2, λ, määramatuse kordaja, kontingentsuse kordaja, Crameri V) vähe sisukad ja nende kasutamine mõttetu. Küll aga on sobiv kasutada monotoonse seose kordajaid - Somersi D, ν, Kendalli kordajaid ja samuti ka lineaarset korrelatsioonikordajat.

Joonistel 7.7 ja ja 7.8 esitatud tabelitest järeldub, et sünnipaiga ja eriala vahelised seosed on nõrgad ja juhuslikud. Kõigi seosekordajate väärtused on madalad (peale kontingentsuse kordaja, mille väärtused on üldiselt suuremad, on kõik teised kordajad absoluutväärtuselt alla 0,5, enamasti 0,1-0,3 piires, ning ühtlasi on nad statistiliselt mitteolulised .

Kui mõlemal tunnusel on kaks väärtust (nn binaarsed või dihhotoomilised tunnused, mille jaotuseks on Bernoulli jaotus, vt [7], p 3.3.2), siis lisatakse tabelisse veel seitsmes rida, mis esitab nn Fisheri täpse testi kontrollimaks sõltuvuse olemasolu. Fisheri täpne test kontrollib hüpoteesipaari

$$H_0: p_{ij} = p_{i.} \cdot p_{.j},$$
$$H_1: p_{ij} \neq p_{i.} \cdot p_{.j}, \quad i,j = 1,2,$$

kasutades polünomiaaljaotuste täpseid väärtusi (mitte normaalset lähendit, nagu paljud teised testid). Väljastatakse olulisuse tõenäosus.

Olgu märgitud, et ka kahe dihhotoomilise tunnuse omavahelist seost võib käsitleda suunatuna: kui jaotus on selline, et mõlema tunnuse esimesed väärtused ja teised väärtused esinevad üheaegselt suhteliselt suurema tõenäosusega, siis on seos positiivne, vastasel korral negatiivne

Seda arvestades saab ka Fisheri täpse testiga kontrollitava sisuka hüpoteesi jaotada kaheks:

H_1': tunnused ei ole sõltumatud, nende vahel on positiivne seos,

H_1'': tunnused ei ole sõltumatud, nende vahel on negatiivne seos.

Hüpoteesi H_1' või H_1'' tohib kontrollida siis, kui on olemas teatav eelinformatsioon, kuid seda kasutada on üldiselt soodus: ühepoolne test on võimsam, st tema puhul on olulisuse tõenäosus väiksem ja sisuka hüpoteesi vastuvõtmine "kergem". Võib esineda olukordi, mil ühepoolne sisukas hüpotees on teatava olulisuse nivoo korral vastu võetav, kuid kahepoolset (mille puhul seose suund ei ole ette teada) sama olulisuse nivoo korral vastu võtta ei saa.

Joonisel 7.9 on esitatud Fisheri test tunnuste **sugu** ja **abielus** seose kontrollimiseks. Testist järeldub, et seos puudub.

```
Fisher's Exact Test                    0.70145 (one-tail)
                                       1.00000 (two-tail)
```

Jn 7.9

Ülejäänud kaks rida aknas joonisel 7.3 on selleks, et salvestada väljund. Kolmas rida *Save counts* salvestab sagedustabeli, neljas rida *Save labels* sagedustabelis paiknevad tunnuse väärtuste nimetused.

7.2.6. Seosekordajate arvutamine tabeli põhjal (*Contingency Tables*). Protseduur *Contingency Tables* (teine protseduur allmenüüst *P. Categorical Data Analysis*) on ette nähtud seosekordajate arvutamiseks olemasoleva sagedustabeli põhjal.

Protseduuri lähteinformatsiooniks on sagedustabel, mis võib olla ka näiteks eelmise protseduuri tulemusena salvestatud (vt jn 7.10).

Arvutatavad statistikud langevad põhiliselt ühte punktides 7.2.3 ja 7.2.4 kirjeldatutega, ka on sama nende paigutus. Puudub ainult regressioonisuhe η (*Eta*) esimese leheküljе lõpust, sest selle arvutamiseks ei piisa ainult sagedustabeli teadmisest, vaid on tarvis teada ka tunnuse väärtusi.

```
: TABLE
1 1 3
2 5 1
2 2 1
5 0 5
```

Jn 7.10

7.2.7. Kahe tunnuse ühisjaotuse graafiline illustratsioon (*Three-Dimensional Histogram*). Protseduur *Three-Dimensional Histogram* on seitsmes (viimane) allmenüüs *F. Descriptive Methods*.

See protseduur võimaldab esitada arvtunnuste sagedus- ja jaotustabeleid graafiliselt "ruumilise" tulpdiagrammi kujul.

Käivitamisel ilmub ekraanile tellimuspaneel (vt jn 7.11), millele tuleb kanda analüüsitavate tunnuste nimed.

```
                    Three-Dimensional Histogram
-------------------------------------------------------------------

Sample 1: tulu

Sample 2: pere
```

Jn 7.11

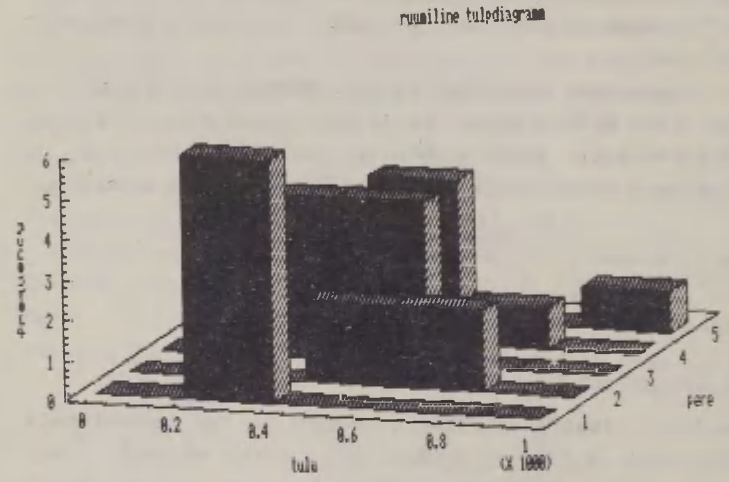
Seejärel ilmub ekraanile varasemast tuttav jaotustabeli tellimuspaneel, millel sedapuhku tuleb täita niihästi vasak kui ka parem pool (vastavalt esimese ja teise tunnuse jaoks),

```
                    Tabulation Input Panel

Primary Variable                        Secondary Variable
-------------------------------------------------------------------

Type            Continuous              Type            Discrete
Lower limit     0                       Lower limit     1
Upper limit     1000                    Upper limit     5
No. of classes  5                       No. of classes  5

Length = 27                             Length = 27
Minimum = 200                           Minimum = 1
Maximum = 900                           Maximum = 5

     Top Title:    ruumiline tulpdiagramm
     (7 lines)
     X-axis title: tulu
     Y-axis title: pere

     Cumulative: No                     Relative: No
```

Jn 7.12

vt 7.12. Tabeli täitmisprintsiibid sarnanevad varem kirjeldatutega. Oluline on siin see, et klasside arv (*No. of classes*) ei tohiks olla liiga suur, sest siis kipuvad tagumised tulbad esimeste taha varju jääma (see oht on niikuinii olemas).

Protseduuri käivitamise tulemusena saame ekraanile ruumilise tulpdiagrammi graafiku (vt jn 7.13).



Jn 7.13

Olgu märgitud, et tulpdiagrammi ruumilist konfiguratsiooni saab muuta, kasutades võimalust *Graphics Options*, vt [7] jn 2.12, kus paremal allnurgas on vaatenurka (*Viewpoint*) määravad parameetrid. Selle graafiku vormistamisprotseduurini saab jõuda võtme **F5** abil graafilisest režiimist, valides aknast ([7], jn 2.18) teise rea.

7.2.8. Horisontaalne tulpdiagramm (*Barcharts*) kahe tunnuse jaoks. Teine võimalus ühisjaotust graafiliselt esitada on kasutada tulpdiagrammi, kus ühe tunnuse sagedustele vastavad tulbad on teise tunnuse järgi grupeeritud. Seda võimalust realiseerib juba tuttav protseduur *Barcharts* (vt p 6.4.3).

Käesoleval juhul tuleb tellimuse vormistamisel märkida kahe tunnuse nimed (vt jn 7.14).

Data vectors: synnipaik
eriala

Jn 7.14

Barchart Options

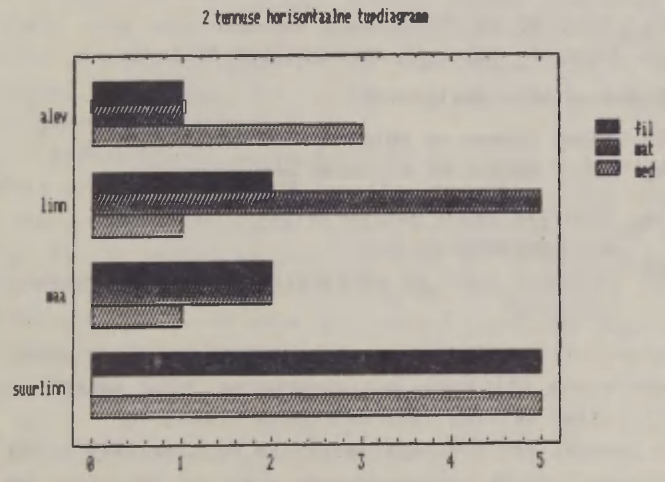
Primary Classification				Secondary Classification			
Class	Label	Color	Fill	Class	Label	Color	Fill
1	alev			1	fil	4	1
2	linn			2	mat	5	2
3	maa			3	med	6	3
4	suurlinn						

	Horiz.	Vert.
Plot Origin:	.000	.000
Plot Size:	.750	1.000
Legend at:	1.150	.770

Top Title: 2 tunnuse horisontaalne tupdiagramm (2 lines)
Axis Title:

Format: Clustered
Baseline: 0
Direction: Horizontal
Grid: No

Jn. 7.15



Jn 7.16

Protseduuri käivitamise järel ilmub tulpdiagrammi tellimuspaneel (vt jn 7.15), mis võimaldab anda tunnuste väärtustele või väärtusklassidele (*Class*) nimetused (*Label*), kui neid ei ole (käesolevas näites kasutatakse nominaaltunnuseid, millel nimetused on juba olemas). Valida on võimalik graafiku värve ning tulpade orientatsiooni ja täitemustreid, joonise suurust, pealkirju jm.

Pärast tellimuspaneeli täitmisele järgnevat uut käivitamist ilmub ekraanile graafik (vt jn 7.16).

§ 7.3. Kahe jaotuse/jaotusparameetrite võrdlemine

Ühe erivaldkonna kahe tunnuse analüüsist moodustab juhtum, mil üks tunnustest on dihhotoomiline (kahe väärtusega). Sel juhul on analüüsimeetodite põhiliseks sisuks kahe jaotuse või vastavate jaotusparameetrite võrdlemine.

Märgime, et sel juhul on sisuliselt ekvivalentsed järgnised hüpoteesipaarid:

a) H_0: kahes rühmas on jaotused võrdsed,
   H_1: kahes rühmas on jaotused erinevad,

b) H_0: tunnused on sõltumatud,
   H_1: tunnuste vahel on statistiline sõltuvus,

samuti järgnised hüpoteesipaarid:

a) H_0: kahes rühmas on võrdsed keskväärtused,
   H_1: kahes rühmas on erinevad keskväärtused,

b) H_0: tunnuste vahel puudub korrelatiivne ja regressioonsõltuvus,
   H_1: tunnuste vahel on korrelatiivne ja regressioonsõltuvus.

Märgime, et korrelatiivse sõltuvuse olemasolust järeldub ka monotoonse sõltuvuse eksisteerimine, ning annugi tuleneb siit üldise statistilise sõltuvuse olemasolu.

Kahe jaotuse või jaotusparameetrite võrdlemiseks on SG-süsteemis ette nähtud 4 protseduuri, mida me järgnevas käsitlemegi.

7.3.1. Keskmiste ja dispersioonide võrdlemine protseduuriga *Two-Sample Analysis*. Protseduur *Two-Sample Analysis* on teine allmenüüs *G. Estimation and Testing*.

Selle protseduuri abil arvutatakse keskmiste ja dispersioonide jaoks punkthinnangud ning on võimalik kontrollida mitmesuguseid hüpoteese võrdlemaks keskväärtusi ja dispersioone eeldusel, et võrreldavad valimid on sõltumatud.

Protseduuri *Two-Sample Analysis* tellimuse vorm on kahjuks niisugune, et võimaldab kergesti esitada ka ekslikke tellimusi, mille puhul sõltumatuse eeldus ei ole täidetud ja seetõttu on tulemusedki mittekorrektsed.

Tellimus algab kahe võrreldava andmekogumi (tunnuse) nimetusega (*Sample 1* ja *Sample 2*), vt jn 7.17.

Two-Sample Analysis

Sample 1: tulu SELECT sugu EQ 1

Sample 2: tulu SELECT sugu EQ 2

Jn 7.17

Vormiliselt võib siin kasutada kahe erineva tunnuse nime (ei ole nõutud, et need oleksid võrdse pikkusega). Sisuliselt pole kahe tunnuse mõõtmistulemuste võrdlemine õigustatud juhul, kui tegemist on tunnustega, mis on mõõdetud samadel objektidel, sest metoodika, mida rakendatakse, eeldab tunnuste sõltumatust.

Andmekogumite kirjeldamiseks tellimuse esitamisel sobib hästi operatsioon **SELECT** (vt [7], lk 159), mille abil mingi tunnuse mõõtmistulemuste hulgast moodustatakse kaks osavalimit. Nii on tehtud ka joonisel 7.17 esitatud tellimuses, et võrrelda mees- ja naisaspirantide sissetulekut. Tunnuste nimetuste asemel on siia kirjutatud tunnuseid **sugu** ja **tulu** sisaldavad SG-avaldised

tulu SELECT sugu EQ 1

ja

tulu SELECT sugu EQ 2.

Protseduuri tulemuste trükis (vt jn 7.18) koosneb, samuti kui talle lähedase protseduuri *One-Sample Analysis* (vt p 6.3.3) oma, neljast blokist.

Two-Sample Analysis Results

		Sample 1	Sample 2	Pooled
Sample Statistics:	Number of Obs.	12	15	27
	Average	531.25	417.333	467.963
	Variance	38736.9	23292.4	30088
	Std. Deviation	196.817	152.618	173.459
	Median	500	400	450

Difference between Means = 113.917

		95 Percent	
Conf. Interval For Diff. in Means:			
(Equal Vars.)	Sample 1 - Sample 2	-24.4767 252.31	25 D.F.
(Unequal Vars.)	Sample 1 - Sample 2	-30.1619 257.995	20.4 D.F.

Ratio of Variances = 1.66307

Conf. Interval for Ratio of Variances: 0 Percent
Sample 1 → Sample 2

Hypothesis Test for H0: Diff = 0 Computed t statistic = 1.69568
vs Alt: NE Sig. Level = 0.102366
at Alpha = 0.05 so do not reject H0.

Jn 7.18

Esimeses blokis on vaatluste arv (*Number of Obs.*), keskmine (*Average*), dispersioon (*Variance*), standardhälve (*Std. Deviation*) ja mediaan (*Median*) esimese rühma (*Sample 1*), teise rühma (*Sample 2*) ja nende summa (*Pooled*) jaoks.

Teises blokis on arvutatud keskmiste erinevus (I rühma keskmine - II rühma keskmine) ning leitud selle erinevuse 95 % usalduspiirkond kahel erineval eeldusel:

1) rühmade dispersioonid on võrdsed (*Equal Vars.*),

2) rühmade dispersioonid ei ole võrdsed (*Unequal Vars.*).

Mõlema juhu jaoks arvutatakse ka vabaduastmete arv t-testi kasutamiseks, kusjuures juhul 2) on see ligikaudne.

Olgu märgitud, et juba selle bloki põhjal võib teha järelduse: kui keskmiste vahe usalduspiirkond sisaldab 0-punkti, tuleb vastu võtta nullhüpotees, mis väidab keskmiste võrdsust.

Kolmandas blokis on arvutatud dispersioonide suhe ning selle suhte usalduspiirid. Selle bloki tulemusi on kasulik vaadata enne teise bloki põhjal järelduste tegemist: kui

usalduspiirkond ei sisalda arvu 1, on dispersioonid oluliselt erinevad ning otsustamisel tuleb arvestada juhtu 2).

Neljas blokk on hüpoteeside kontrollimiseks keskväärtuste kohta.

Vaikimisi kontrollitakse olulisuse nivool 0,05 hüpoteesipaari

$$H_0: \mu_1 - \mu_2 = 0,$$
$$H_1: \mu_1 - \mu_2 \neq 0,$$

kusjuures väljastatakse järgmised tulemused:

1° Arvutatakse välja t-statistik *Computed t statistic*, mille põhjal on võimalik teha järeldus kas tabelite või protseduuri **TAILS** (vt [7], p 3.4.1) abil.

2° Leitakse sellele t-statistiku väärtusele vastav olulisuse tõenäosus p (*Sig. level*). Selle näitaja abil saab teha otsustuse ilma tabeleid või lisaprotseduure kasutamata:

kui $p \leq \alpha$, siis võetakse vastu H_1, muidu H_0.

3° Sõnastatakse hüpoteesi kontrollimise resultaat, väljastades kas lause

so do not reject H_0

(H_0 ei saa kummutada, H_0 võetakse vastu) või

reject H_0

(H_0 kummutatakse, H_1 on sellega tõestatud).

Paneme tähele, et tulemuste tabelis on mõned parameetrid muudetavad. Nendeks on blokkides 2 ja 3

- usaldusnivoo (*Conf. Interval for ...*), mis vaikimisi on 95 %,

ja blokis 4

- vahe väärtus (vaikimisi 0),
- hüpoteesi H_1 määrav võrdsus- või võrratussuhe,
- olulisuse nivoo α väärtus *Alpha*.

Pärast esimeste tulemuste kättesaamist on neid parameetreid võimalik muuta. Joonisel 7.19 on hüpotees H_1 muudetud ühepoolseks, seega kontrollitakse nüüd hüpoteesipaari

$$H_0: \mu_1 \leq \mu_2,$$
$$H_1: \mu_1 > \mu_2,$$

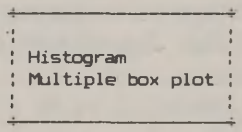
st me soovime tõestada, et meesaspirantide sissetulek on

kõrgem. Tulemus on aga jällegi negatiivne, sest p = = 0,0511831 > 0,05. Kasutades mõnevõrra suuremat olulisuse nivood (nt α = 0,10) või ka suurendades valimi mahtu, õnnestuks nimetatud hüpotees tõenäoliselt ka tõestada.

```
Hypothesis Test for H0: Diff = 0        Computed t statistic = 1.69568
                    vs Alt: GT          Sig. Level = 0.0511831
                at Alpha = 0.05         so do not reject H0.
```

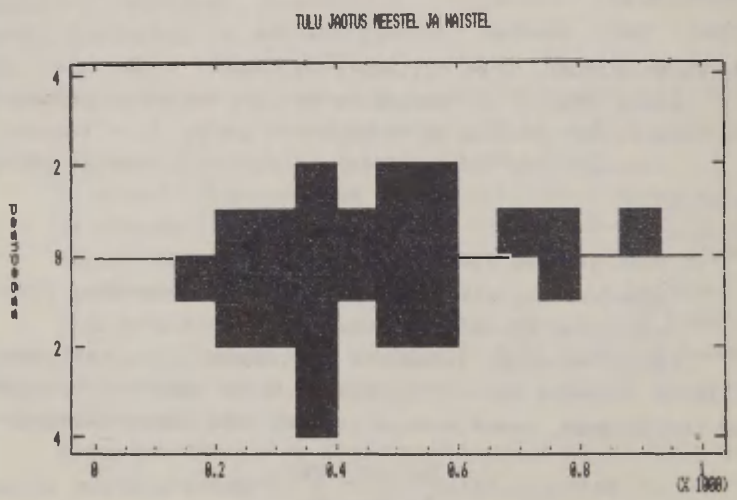
Jn 7.19

7.3.2. Täiendavad graafilised võimalused protseduuri *Two-Sample Analysis* juures. Täiendavate võimaluste loetelu akna (vt jn 7.20) saame tellida võtme F5 abil.



Jn 7.20

Üks võimalusi on kahesuunaline tulpdiagramm (*Histogram*). Selle väliku tulemusena saame ekraanile jaotustabeli

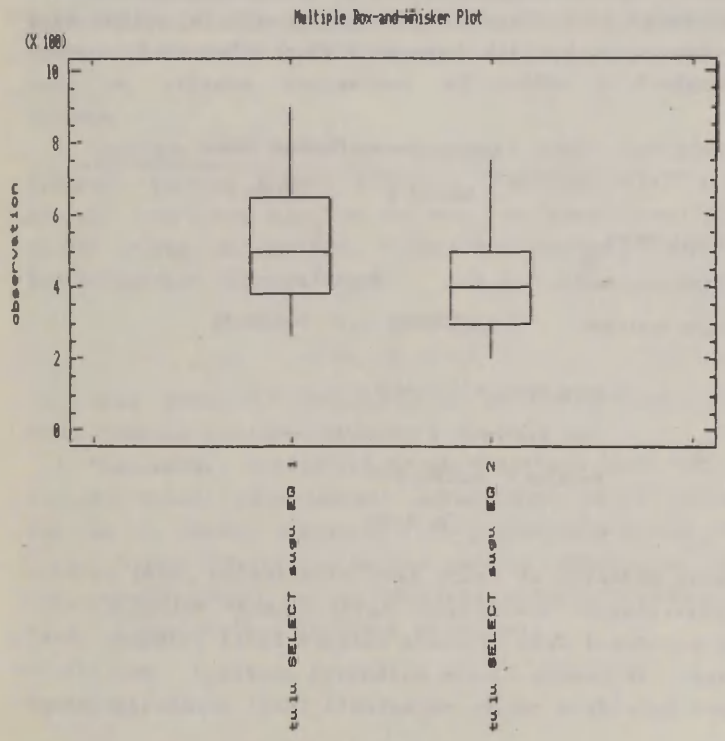


Jn 7.21

tellimuspaneeli, millest käesoleval juhul tuleb täita ainult vasak pool ja soovi korral tekstide kujundamise blokk.

Protseduuri käivitamise tulemusena moodustatakse kaks histogrammi, kasutades ühist klassijaotuse skaalat (vt jn 7.21). Neist esimesele andmekogumile vastab joonise ülemisel poolel paiknev, ülespidi suunatud tulpadega histogramm ja teisele joonise alumisel poolel paiknev, allapidi suunatud tulpadega histogramm.

Teine täiendav võimalus on karpdiagrammide moodustamine (vt p 6.3.2). Karpdiagrammid joonistatakse vertikaalselt (uuritava tunnuse väärtuste skaala on vertikaalteljel), neid on kõrvuti kaks tükki, kusjuures esimene vastab I, teine - II andmekogumile, vt jn 7.22.



Jn 7.22

7.3.3. Poissoni jaotuse parameetrite võrdlemine (*Comparison of Poisson Rates*). Poissoni jaotuse parameetrite võrdlemine on viimane (viies) protseduur allmenüüs *G. Estimation and Testing*.

Selle protseduuri sisuks on tegelikult üksiksündmuse sageduste võrdlemine kahe sõltumatu katseseeria korral, kusjuures eeldatakse, et sagedused on väikesed (sel juhul on jaotus lähendatav Poissoni jaotusega).

Protseduuri käivitamise järel ilmub ekraanile tellimuspaneel, kus esimeses blokis on 3 rida ja 3 veergu - esimene on nimetuste veerg, teine ja kolmas veerg vastavad I ja II valimile või katseseeriale (*Sample 1, Sample 2*).

Märgime, et tühjad lahtrid paneelil on täidetud vaikimisi väärtustega, mis tellimuse seisukohalt on sisutud ja mis tuleb asendada konkreetsete andmetega. Need on esimeses reas sündmuse esinemiste arv (*Event Count*) ja teises reas katse-- seeria pikkus ehk katsete üldarv (*Interval length*), vt jn 7.23.

Comparison of Poisson Rates

	Sample 1	Sample 2
Event count	15	7
Interval length	65	111
Rate estimate	0.230769	0.0630631

Rate ratio = 3.65934

Test statistic z = 2.8162

P-value = 4.85967E-3

Jn 7.23

Nende andmete, st nelja arvu sisestamise järel protseduur käivitatakse. Arvutatakse katsetulemuse esinemise suhtelised sagedused *Rate estimate* esimese bloki viimases reas, kuid samuti ka teises blokis paiknevad suurused - suhteliste sageduste suhe *Rate ratio*, asümptootiliselt normaaljaotusega

z-statistiku väärtus *Test statistic z* ja olulisuse tõenäosus p *P-value*. Viimase järgi on võimalik standardsel viisil kontrollida hüpoteesipaari

$$H_0: \lambda_1 = \lambda_2,$$
$$H_1: \lambda_1 \neq \lambda_2,$$

kus λ_1 ja λ_2 on sündmuse esinemise tõenäosused (Poissoni jaotuse parameetrite väärtused) kummaski katseseerias. Hüpotees H_1 võetakse vastu siis, kui

$$p \leq \alpha,$$

kus α on ette antud olulisuse nivoo (nt $\alpha = 0{,}05$).

Märgime, et korrektse tulemuse saamiseks peavad katseseeriad olema küllalt pikad.

7.3.4. Kahe empiirilise jaotuse võrdlemine Kolmogorov-Smirnovi testi abil (*Kolmogorov-Smirnov Two-Sample Test*). Kolmogorov-Smirnovi test kahe empiirilise jaotuse võrdlemiseks on viimane protseduur allmenüüs *R. Nonparametric Methods*.

Märkigem, et Kolmogorov-Smirnovi testi kasutamine on üldiselt lubatud üksnes pidevate (teoreetiliste) jaotuste korral, kusjuures oluline on see, et kasutatavad valimid peavad olema sõltumatud. Kolmogorov-Smirnovi testi abil kontrollitakse hüpoteesipaari

$$H_0: P_1 = P_2,$$
$$H_1: P_1 \neq P_2,$$

kusjuures kontrollstatistikuks DN on empiiriliste jaotusfunktsioonide maksimaalne vahe.

Protseduuri tellimisel tuleb sisestada kahe võrreldava tunnuse nimed; käivitamisel arvutatakse välja statistiku väärtus ja samuti ligikaudne olulisuse tõenäosuse väärtus (vt jn 7.24). Käesolevas näites (kus on sisestatud lihtsalt tunnuste väärtused) on nullhüpotees ülimalt tõenäone, pole mingit alust oletada jaotuste erinevust.

Kolmogorov-Smirnov Two-Sample Test

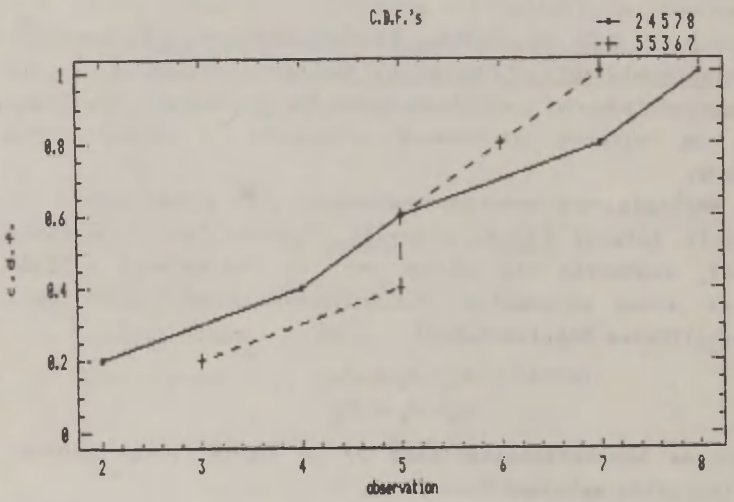
Sample 1: 2 4 5 7 8

Sample 2: 5 5 3 6 7

Estimated overall statistic DN = 0.4
Approximate significance level = 0.999983

Jn 7.24

Protseduuri on võimalik ka illustreerida; vastaval graafikul kujutatakse mõlemad empiirilised jaotusfunktsioonid (vt jn 7.25, siin kirjeldab I tunnust pidev, II tunnust punktiirjoon).



Jn 7.25

7.3.5. Mitteparameetriline test kahe tunnuse mediaanide võrdlemiseks (*Comparison of Two Samples*). *Comparison of Two Samples* on neljas protseduur allmenüüs *R. Nonparametric Methods*.

Selle protseduuri abil saab kontrollida hüpoteesi kahe arv- või järjestustunnuse mediaanide ühtelangemise kohta, st kontrollitakse hüpoteesipaari

$$H_0: med_1 = med_2,$$
$$H_1: med_1 \neq med_2.$$

Enamasti võrreldakse selle protseduuri abil sisuliselt sama tunnust kahes erinevas üldkogumis.

Protseduur koosneb kolmest alamprotseduurist, mida järgnevas kirjeldamegi.

Tellimuspaneelile tuleb märkida kahe tunnuse/andmekogumi nimed (*Sample 1* ja *Sample 2*), seejärel tuleb valida meetod, kasutades selleks akent *Test based on* (vt jn 7.26).

Comparison of Two Samples

Sample 1: kaal

Sample 2: idealkaal

Test based on: Signs

Jn 7.26

Siin on võimalikud järgmised valikud, mis on vahetatavad tühikuklahvi abil:

1° *Signs* - märgitest, kasutatav kahe tunnuse jaoks, mis on mõõdetud samadel objektidel (nn *paired samples*);

2° *Ranks* - Wilcoxoni märgistatud astakute test (*signed ranks*) samadel objektidel mõõdetud tunnuste võrdlemiseks;

3° *Pairs* - Mann-Whitney test kahe sõltumatu kogumi võrdlemiseks; selle testi puhul pole nõutav tunnuste võrdne pikkus.

Peale valiku tegemist käivitatakse protseduur võtmega F6. Tulemused ilmuvad samale tellimuspaneelile. Nende hulgas on kindlasti kontrollstatistiku Z väärtus (tegemist on asümptootiliselt normaaljaotusega statistikuga, seetõttu trükitakse *Large sample test statistic*) ning olulisuse tõenäosus p (*Two-tailed probability of equaling or exceeding Z*). Hüpotees H_1 võetakse vastu parajasti siis, kui $p \leq \alpha$.

```
: idealkaal GETS kasv - 105
: idealkaal
57 58 58 60 62 63 63 64 65 65 65 65 67 67 67 68 70 70 71 73 74 75 77 77 77
     77 80 90
:
```

Jn 7.27

Käesolevas näites (vt jn 7.28) on tunnust **kaal** võrreldud uue tunnusega **ideaalkaal** (vt jn 7.27), mis on arvutatud mõõdetud tunnuse **kasv** järgi (vt [7], p 4.3.4).

Comparison of Two Samples

Sample 1: kaal

Sample 2: idealkaal

Test based on: Signs

Number of positive differences = 10
Number of negative differences = 16
Expected number = 13
Large sample test statistic Z = 0.980581
Two-tailed probability of equaling or exceeding Z = 0.326798

NOTE: 28 total pairs. 2 tied pairs ignored.

Jn 7.28

Nagu näha jooniselt 7.28, võib lugeda **kaalu** ja **ideaalkaalu** mediaanid võrdseiks.

Kuivõrd märgitest kasutab ainult väärtuste positiivseid ja negatiivseid vahesid, on teststatistiku konstrueerimisel jäänud välja 2 objekti, kelle puhul **kaal** **ideaalkaaluga** ühte langeb. Seda kinnitab märkus joonise 7.28 allservas (märgime, et nullvahet tähistab siin sõna *tie*).

§ 7.4. Tinglike jaotuste ja keskmiste analüüs

Käesolevas paragrahvis käsitleme kahe tunnuse analüüsi juhul, kui üks tunnustest on diskreetne (arvuline või nominaalne), teine arvuline (kas pidev või diskreetne). Sellise tunnusepaari jaoks on süsteemis SG viis protseduuri, mida vaatlema järgnevas asumegi.

7.4.1. Tinglike jaotusparameetrite arvutamine (*Codebook*). *Codebook Procedure* on allmenüüs *F. Descriptive Methods* kuues. See protseduur on ette nähtud arvtunnuse tinglike jaotusparameetrite arvutamiseks, kusjuures tingimus on määratud diskreetse tunnusega.

Protseduuri *Codebook* tellimuspaneel on esitatud joonisel 7.29. Siin tuleb täita järgmised lahtrid: uuritava tunnuse nimi - *Data,* rühmitava tunnuse nimi - *Level codes.* Kui on soov omistada rühmitava tunnuse tasemetele (uued) koodnimetused, võib seda teha, kandes nimetusi sisaldava (sümbol-) tunnuse nime lahtrisse *Labels.* Kui sellist soovi pole (näiteks kui tunnuse väärtustel juba on sobivad nimetused), võib vastava lahtri tühjaks jätta.

Codebook

Data: tulu

Level codes: eriala

Labels:

Statistics: ABCDEFGHIJKLMNOPQ Confidence level: 95.00Point symbols: No

(a) Average	(f) Std. deviation	(k) Lower quartile	(p) Kurtosis
(b) Median	(g) Std. error	(l) Upper quartile	(q) Std. kurtosis
(c) Mode	(h) Minimum	(m) Interquar. range	
(d) Geometric mean	(i) Maximum	(n) Skewness	
(e) Variance	(j) Range	(o) Std. skewness	

Jn 7.29

Edasine paneeli täitmine sarnaneb protseduuri *Summary Statistics* (vt p 6.3.1) tellimuspaneeli täitmisele: nimelt on esitatud statistikute loetelu, mida tähistavad tähed A... ...Q. Vaikimisi arvutatakse kõik statistikud. Valiku tegemiseks kustutatakse sümboljadast *Statistics* need tähed, millele vastavaid statistikuid ei soovita arvutada.

Lahtrid *Confidence level* (usaldusnivoo, vaikimisi 95 %) ja *Point symbols* (punktide sümbolid, vaikimisi *No,* st ei kasutata) on vajalikud jooniste kujundamiseks.

Joonisel 30 on esitatud protseduuri *Codebook* täielik (vaikimisi) trükis, mille põhjal on võimalik visuaalselt võrrelda erinevate erialade aspirantide sissetulekuid. Märgime, et selline trükis on saadud klahvi **F5** (trükirežiim) abil, mille tulemusena neljarealine tabel on kujundatud ühele leheküljele.

Täiendavate võimaluste loetelu esitab võtme **F5** abil välja kutsutav aken, vt jn 7.31.

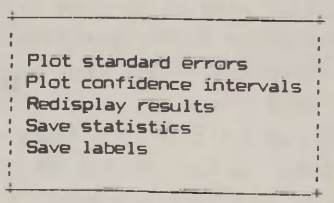
Level	Sample size	Average	Median	Mode	Geometric mean
fil	10	491.000	425.000	350.000	435.900
mat	8	461.250	450.000	400.000	450.110
med	10	443.500	425.000	500.000	421.394

Level	Variance	Standard deviation	Standard error	Minimum	Maximum
fil	60476.7	245.920	77.7667	200.000	900.000
mat	11441.1	106.963	37.8171	300.000	600.000
med	23033.6	151.768	47.9933	250.000	750.000

Level	Range	Lower quartile	Upper quartile	Interquartile range	Skewness
fil	700.000	260.000	700.000	440.000	0.50199
mat	300.000	395.000	550.000	155.000	0.06921
med	500.000	310.000	500.000	190.000	0.79478

Level	Standardized skewness	Kurtosis	Standardized kurtosis
fil	0.64807	-1.17613	-0.75919
mat	0.07992	-1.03460	-0.59733
med	1.02606	0.43867	0.28316

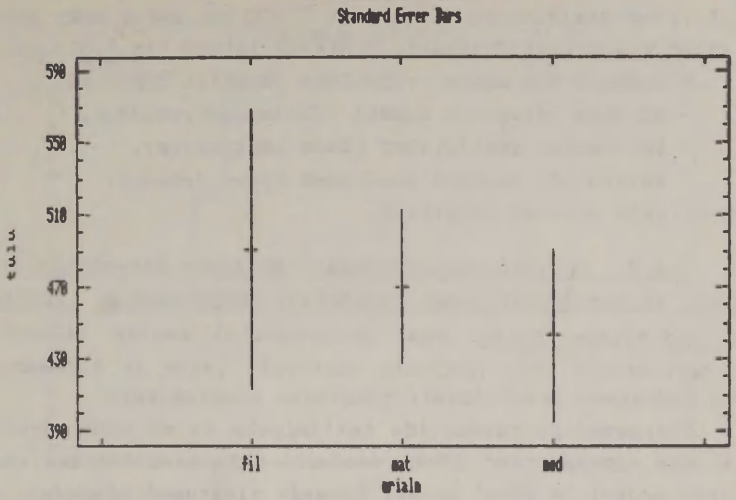
Jn 7.30



Jn 7.31

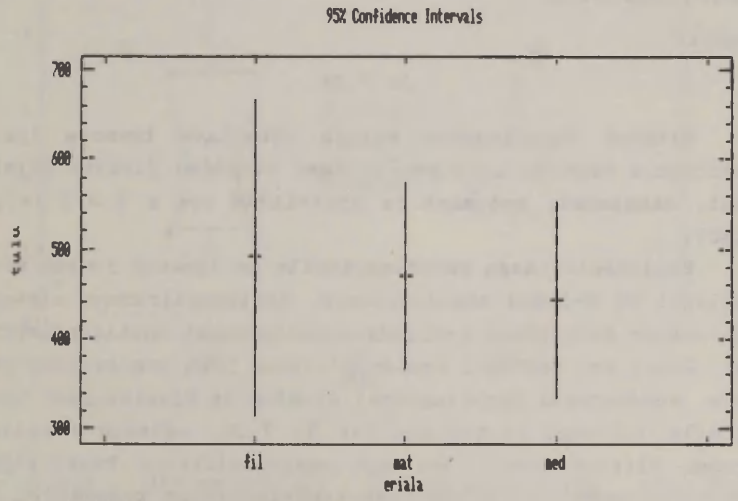
Vaatleme neid võimalusi lähemalt.

Esimene rida - *Plot standard errors* - tähendab graafikut, kus keskmiste juurde on kujutatud tinglike standardvigade pikkused lõigud (vt jn 7.32).



Jn 7.32

Teine rida - *Plot confidence intervals* - tähendab graafikut, kus keskmiste juurde on kujutatud nende usaldusintervallide pikkused lõigud (vt jn 7.33). Kuigi joonised 7.32 ja 7.33 on üsna sarnased, on erinevused lõikude pikkustes.



Jn 7.33

95 %-lised usaldusintervallid (jn 7.33) on umbes kaks korda pikemad kui standardvigadele vastavad lõigud (jn 7.32).

Ülejäänud osa aknast võimaldab järgmisi tegevusi:

- esitada tulemused uuesti (*Redisplay results*),
- salvestada statistikud (*Save statistics*),
- salvestada koodide nimetused (*Save labels*),

mis ei vaja erilist selgitust.

7.4.2. Mitmene karpdiagramm (*Multiple Box-and-Whisker Plot*) ja usalduspiiridega (sälgatud) karpdiagramm (*Notched Box-and-Whisker Plot*). Need protseduurid asuvad allmenüüs *I. Exploratory Data Analysis* vastavalt teise ja kolmandana ning kujutavad graafiliselt tinglikke statistikuid.

Nimetatud protseduuride tellimiseks tuleb vaid märkida uuritava tunnuse nimi (*Data vector*), rühmitava tunnuse nimi (*Level codes*) ja soovi korral rühmade nimetused (*Labels*), vt tellimuspaneeli joonisel 7.34.

Multiple Box-and-Whisker Plot

Data vector: tulu

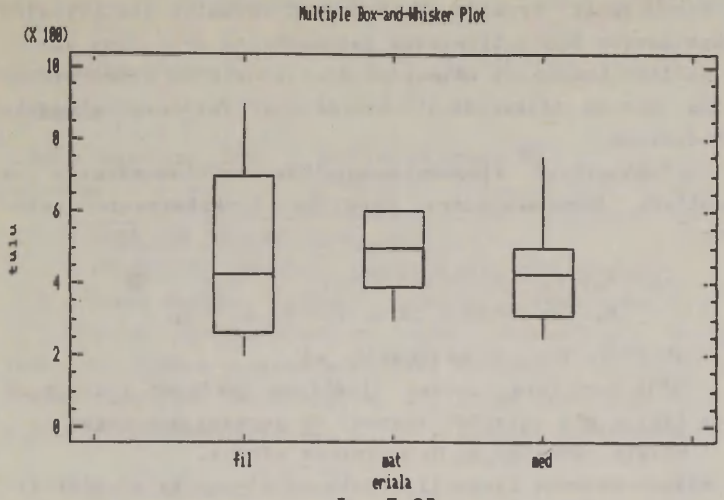
Level codes: eriala

Labels:

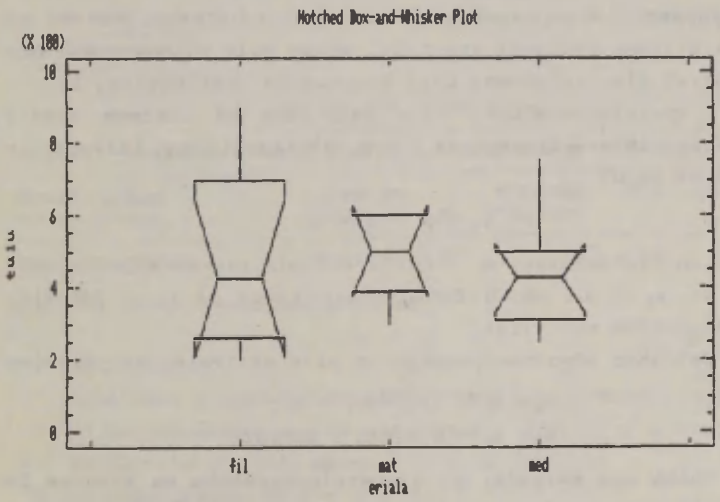
Jn 7.34

Mitmene karpdiagramm esitab rühmitava tunnuse igale väärtusele vastava uuritava tunnuse tingliku jaotuse miinimumi, maksimumi, mediaani ja kvartiilid (vt p 6.3.2 ja jn 7.35).

Usalduspiiridega karpdiagrammile on lisatud juurde veel mediaani 95 %-lised usalduspiirid. Usalduspiirkonna ulatust tähistavad kaldlõigud (sälgud) karpdiagrammi vertikaalservadel. Juhul kui mediaani usalduspiirkond jääb kvartiilide vahele, moodustavad karpdiagrammi alumine ja ülemine pool trapetsile liitunud ristküliku (vt jn 7.36, esimese karpdiagrammi ülemine pool). Kui aga usalduspiirkond kvartiilide vahele ei mahu, tekib omamoodi tagasipööratud kujund (vt jn 7.36, keskmine karpdiagramm).



Jn. 7.35



Jn 7.36

7.4.3. Ühefaktoriline dispersioonanalüüs (*One-Way Analysis of Variance*). Seni kirjelatud protseduurid võimaldasid tinglike jaotuste parameetreid (keskväärtusi, mediaane)

küll visuaalselt võrrelda, kuid puudus võimalus statistiliste hüpoteeside kontrollimiseks parameetrite erinevuse kohta.

Sellist kontrolli võimaldab ühefaktoriline dispersioonanalüüs, mis on allmenüüs *J. Analysis of Variance* esimeseks protseduuriks.

Ühefaktorilise dispersioonanalüüsi põhieesmärgiks on kontrollida hüpoteesipaari tinglike keskväärtuste vahekorrast

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k;$$
$$H_1: \text{leiduvad i ja j, nii et } \mu_i \neq \mu_j.$$

Protseduur tugineb eeldusele, et

- kõik uuritava tunnuse tinglikud jaotused (tingimuse määrab faktor ehk rühmitav tunnus) on normaaljaotusega,
- kõigis rühmades on dispersioon võrdne.

Mõlema eelduse kontrollimiseks on olemas ka statistilised protseduurid. Mis puutub normaalsuse eeldusse, siis tuleb arvesse võtta, et juhul kui rühmades on objekte 10-20 või vähemgi, siis tavaliselt jaotuste sobitamise testid ei suuda mittenormaalsust avastada, seega pole väikeste valimite korral dispersioonanalüüsi kasutamine vastunäidustatud.

Dispersioonanalüüsi võib käsitleda ka teatava mudeli konstrueerimise ülesandena, kus üksikvaatluse tulemus on esitatav kujul

$$y_i = \mu + \alpha_i + \varepsilon_i,$$

kus μ on üldkeskmine, α_i - faktori i-nda taseme mõju (eeldusel, et y_i puhul mõjub faktori i-s tase) ja ε_i - juhuslik prognoosijääk ehk -viga.

Esitatud hüpoteesipaariga on siis ekvivalentne järgmine:

$$H_0: \alpha_1 = \ldots = \alpha_k = 0,$$
$$H_1: \alpha_i \neq 0 \text{ mingi i korral.}$$

Tuleb aga märkida, et dispersioonanalüüs on mahukas ja kompleksne protseduur, millel on lisaks põhilisele - hüpoteeside kontrolli tulemusele - veel terve rida täiendavaid graafilisi ja tekstilisi lisatulemusi.

Dispersioonanalüüsi tellimuspaneelil (vt jn 7.37) on traditsiooniliselt kohad vajalike tunnuste jaoks:

One-Way Analysis of Variance

Data: tulu

Level codes: eriala

Labels:

Range test: Conf. Int. Confidence level: 95

Jn 7.37

- uuritava tunnuse nimi *Data*,
- rühmitava tunnuse e faktori nimi *Level codes*,
- soovi korral lisatakse rühmade nimetused *Labels*.

Edasine informatsioon on vajalik täiendavate tulemuste jaoks. Küsitakse keskmiste mitmese võrdlemise meetodit (vaikimisi pakutakse usalduspiire, s.o *Range test: Conf. Int.*) ja lisaks veel usaldusnivood, mille vaikimisi väärtuseks on 95 %.

7.4.4. Dispersioonanalüüsi tabel. Dispersioonanalüüsi tulemus (vt jn 7.38) ilmub ekraanile vahetult tellimuse alla. See on standardne dispersioonanalüüsi tabel, mis esineb ka edapidi üsna mitmetes trükistes, seetõttu kirjeldame seda lähemalt.

Analysis of variance

Source of variation	Sum of Squares	d.f.	Mean square	F-ratio	Sig. level
Between groups	11486.96	2	5743.482	.173	.8424
Within groups	831680.00	25	33267.200		
Total (corrected)	843166.96	27			

0 missing value(s) have been excluded

Jn 7.38

Täielikus dispersioonanalüüsi tabelis on kuus veergu:

1) hajuvuskomponentide nimetused - *Source of variation*,
2) hälvete ruutude summa - *Sum of Squares*,
3) vabadusastmete arv - *d.f.*,
4) keskruut - *Mean square* (keskmistatud hälvete ruutude summa),
5) F-suhe - *F-ratio*,
6) olulisuse tõenäosus - *Sig. level*.

Ühefaktorilise dispersioonanalüüsi tabelis on kolm rida:

1) rühmadevaheline hajuvus - *Between groups*,

2) rühmasisene hajuvus - *Within groups*,

3) kriipsu all on summarne hajuvus - *Total*.

Dispersioonanalüüsi tabeli täitmiseks vajalikud arvutused on järgmised.

Olgu uuritav tunnus Y ja selle mõõtmistulemused y_i (i = = 1, ..., n). Rühmitava tunnuse X järgi jaotatakse vaatlused k rühma, mida tähistame y ülemise indeksiga. Igas rühmas on üldiselt erinev arv vaatlusi, vaatluste arv i-ndas rühmas olgu n_i, i = 1, ..., k.

Seega saame tunnuse Y väärtuste rühmad

1) $y_1^1, y_2^1, \ldots, y_{n_1}^1$,

2) $y_1^2, y_2^2, \ldots, y_{n_2}^2$,

. . . . . .

k) $y_1^k, y_2^k, \ldots, y_{n_k}^k$.

Igas rühmas arvutame välja ka keskmise

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_j^i ,$$

kuid peale selle leiame ka üldkeskmise

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} \sum_{j=1}^{k} n_i \bar{y}_i .$$

Tunnuse Y hajuvust üldkeskmise ümber kirjeldab hälvete ruutude summa

$$SS = \sum_{i=1}^{n} (y_i - \bar{y})^2 . \qquad (7.13)$$

See summa on jaotatav kaheks osaks (vt nt [10]):

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 . \qquad (7.14)$$

Esimene osa iseloomustab rühmade keskmiste hajuvust üldkeskmise ümber e rühmadevahelist hajuvust, seda tähistatakse sümboliga SS_1. Teises summas eeldame, et vaatlusele y_i mõjub faktori i-s täse.

Teine osa iseloomustab rühmasisest hajuvust, selle tähiseks on SS_o.

Seega on võrdus (7.14) lühidalt esitatav kujul

$$SS = SS_1 + SS_o$$

Vabadusastmete arv tähistab summas paiknevate lineaarselt sõltumatute liidetavate arvu.

Kuivõrd summas SS on üks lineaarne seos (selle tekitab $\bar{y}$ avaldis), siis on summaarne vabadusastmete arv n-1. Samal viisil on SS_1 vabadusastmete arv k-1. Summas SS_o on lineaarseid seoseid k, vabadusastmete arv on seega n-k.

Dispersioonanalüüsis kasutatakse alati selliseid summaarse hajuvuse komponentideks lahutusi, kus komponentidele vastavate vabadusastmete arvude summa võrdub üldhajuvuse vabadusastmete arvuga. Nii ka praegu n - 1 = (k + 1) + (n - k).

Keskruut s^2 arvutatakse mõlemas reas suhtena

$$\text{keskruut} = \frac{\text{hälvete ruutude summa}}{\text{vabadusastmete arv}}$$

Kui peab paika nullhüpotees, st rühmade keskmised on üldkogumis võrdsed, siis on mõlemad keskruudud - niihästi rühmadevaheline kui ka rühmasisene - hinnanguks tunnuse Y dispersioonile ning nende suhe peaks siis olema F-jaotusega vabadusastmete arvudega k-1 ja n-k.

Kui aga nullhüpotees paika ei pea, siis on rühmadevaheline hajuvus märksa suurem rühmasisesest. Järelikult, kui keskruutude suhe (F-suhe) on suurem kriitilisest väärtusest, siis võetakse vastu sisukas hüpotees H_1: faktoril on mõju. See tähendab, et faktori mõjul on uuritava tunnuse keskväärtused rühmades oluliselt erinevad.

Kokkuvõtlikult on äsja tutvustatud suurused esitatud dispersioonanalüüsi tabelis (vt tabel 7.2)

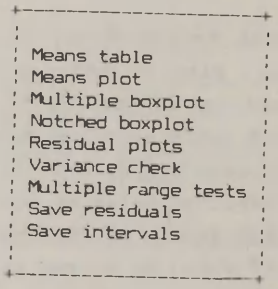
Joonisel 7.38 toodud näites tuleb vastu võtta nullhüpotees: aspirantide tuludes ei ole diferentseeruvust eriala järgi, sest erialasisene hajuvus on märksa suurem erialade vahelisest. Joonise allservas on veel märkus arvestamata jäetud objektide kohta, millel puudub kas mõõdetava tunnuse või faktori väärtus. Kui kõik objektid on mõõdetud, on arvestamata objektide arv 0.

T a b e l 7.2

Dispersioonanalüüsi tabel

Hajuvuse komponent	Ruutude summa	Vabadusastmeid	Keskruut	F-suhe	Olulisuse tõenäosus
Rühmadevaheline	SS_1	k-1	$s_1^2 = \frac{SS_1}{k-1}$	$F = \frac{s_1^2}{s_o^2}$	$p = P(F_{k-1,n-k}>F)$
Rühmasisene	SS_o	n-k	$s_o^2 = \frac{SS_o}{n-k}$		
Summaarne	SS	n-1			

7.4.5. Dispersioonanalüüsi täiendavad võimalused. Joonisel 7.39 on näidatud dispersioonanalüüsi täiendavate võimaluste loetelu aknas, mis saadakse standardsel viisil võtme **F5** abil.



Jn 7.39

Keskmiste tabel (*Means table*) on valitav akna esimesest reast. Selle valimise tulemusena saame tabeli, mis on kujutatud joonisel 7.40.

Selles tabelis on seitse veergu:

- rühma nimetus *Level*,
- objektide arv *Count*,
- keskmine *Average* (vrdl samaga tabelis *Codebook*),
- rühmasisene standardviga *Stnd. Error (internal)* (ka see on sama, mis tabelis *Codebook*),

- rühmade jaoks ühise dispersiooni s_o^2 alusel arvutatud standardviga *Stnd. Error (pooled s)*,

- 95 %-lised usalduspiirid keskmiste jaoks *95 Percent Confidence intervals for mean)*.

Table of means for tulu by eriala

Level	Count	Average	Stnd. Error (internal)	Stnd. Error (pooled s)	95 Percent Confidence intervals for mean	
fil	10	491.00000	77.766745	58.715149	369.78902	612.21098
mat	7	470.00000	42.482489	70.178026	325.12516	614.87484
med	10	443.50000	47.993344	58.715149	322.28902	564.71098
Total	27	467.96296	35.732902	35.732902	394.19631	541.72962

Jn 7.40

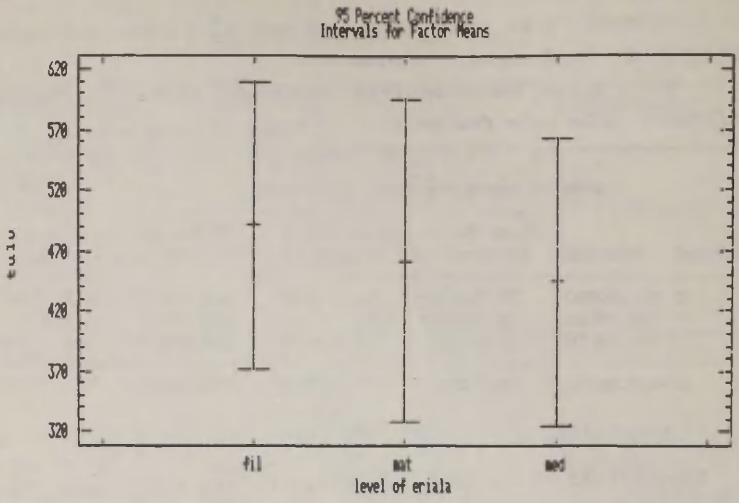
Tabeli read vastavad rühmadele, kriipsu all on aga kogu valimit iseloomustavad summaarsed näitajad - üldkeskmine, üldine standardviga.

Tähelepanu väärib kahe standardvea võrdlus (4. ja 5. veerg tabelis). Neljandas veerus on standardviga arvutatud iga rühma jaoks eraldi, kasutades ainult sellesse rühma kuuluvaid vaatlusi. Viiendas veerus on aga kasutatud dispersioonanalüüsi eeldust, et kõikides rühmades on dispersioonid võrdsed, seetõttu sõltub rühma standardviga ainult rühma arvukusest (tunnuste fil ja med standardviga on võrdne, sest mõlemas rühmas on 10 aspiranti), mitte aga rühma hajuvusest. Ka usalduspiiride arvutamisel on kasutatud ühist standardvea hinnangut.

Järgmisest reast - *Means plot* - on valitav keskmiste graafik. Sellele graafikule on kantud keskmised ja nende usaldusintervallid joonisel 7.40 esitatud tabeli andmetel. Graafikut kujutab joonis 7.41.

Erinevus keskmise usaldusintervallides võrreldes joonisel 7.38 esitatud analoogilise graafikuga tuleneb dispersioonanalüüsi eeldusest võrdsete dispersioonide kohta.

Järgmised read - *Multiple boxplot* (mitmene karpdiagramm) ja *Notched boxplot* (karpdiagramm usalduspiiridega) - esitavad punktis 7.4.2 kirjeldatud graafikuid (vt jn 7.35 ja 7.36).



Jn. 7.41

Järgneb rida *Residual plots* - jääkide graafik, mis esitab graafiliselt nn prognoosijäägid ehk üksikvaatluste hälbed rühmakeskmisest

$$y_j^i - \bar{y}_i, \quad j = 1, \ldots, n_i; \; i = 1, \ldots, k.$$

Jääkide graafiku esitamiseks on kolm standardset võimalust, mille valik toimub tellimuspaneelilt, vt jn 7.42.

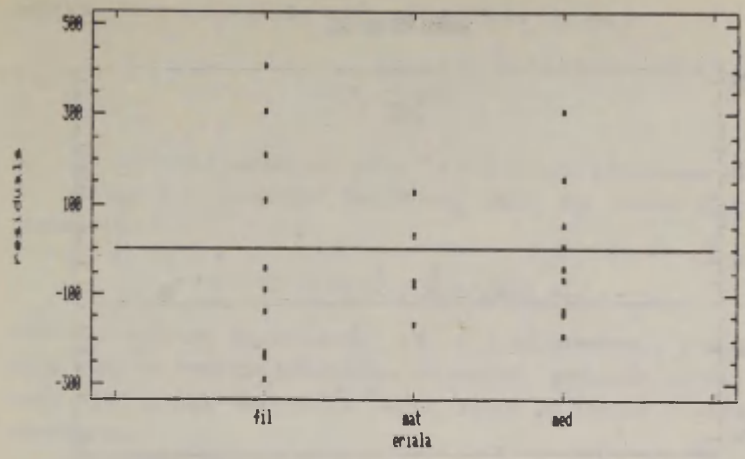
Plot residuals vs factor level, predicted values, or index? (L/P/I):

Jn 7.42

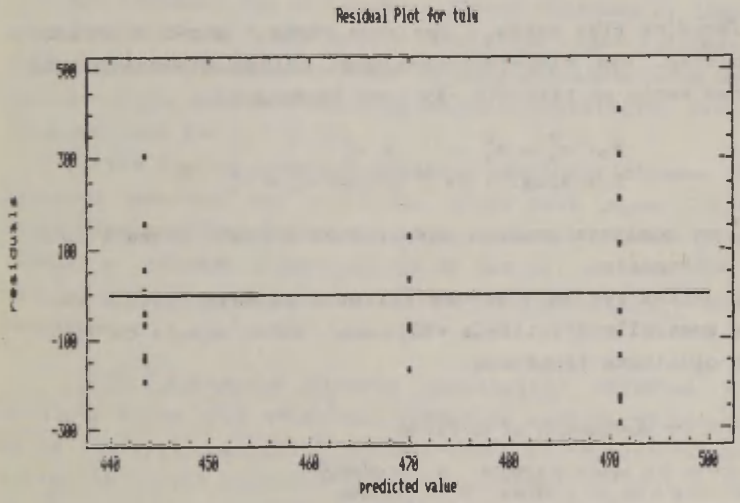
Prognoosijääkide graafiku argumendiks võib olla

- faktori tase (*factor level*),
- prognoositud väärtus, st keskväärtus $\bar{y}_i$ (*predicted values*),
- objekti järjekorranumber valimis l (*index*), l = 1, ..., n.

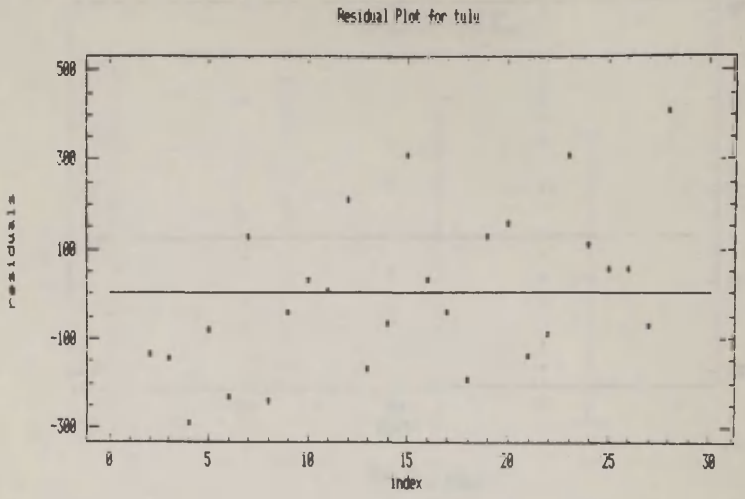
Vastavalt tehtud valikule saame graafikud, mis on kujutatud joonistel 7.43, 7.44 ja 7.45.


residuals
500
300
100
-100
-300
fil
mat
med
erials

Jn. 7.43


Residual Plot for tuls
residuals
500
300
100
-100
-300
440
450
460
470
480
490
500
predicted value

Jn 7.44



Jn 7.45

Järgmine rida aknas - *Variance check* - annab võimaluse kontrollida, kas dispersioonanalüüsi eeldus dispersioonide võrdsuse kohta on täidetud. Esitame hüpoteesid:

$$H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 ,$$
$$H_1: \text{mingi i ja j korral } \sigma_i^2 \neq \sigma_j^2,$$

kus σ_i^2 on uuritava tunnuse dispersioon i-ndale rühmale vastavas üldkogumis.

Trükises (vt jn 7.46) on esitatud kolmele testile vastavate kontrollstatistikute väärtused, neist kahele on lisatud ka olulisuse tõenäosus.

```
Tests for Homogeneity of Variances
----------------------------------
Cochran's C test: 0.629024    P = 0.0539627
Bartlett's test: 1.21099    P = 0.113999
Hartley's test: 4.78707
```

Jn 7.46

Esimesena on märgitud Cochran´i C-test; see test on eriti tundlik ühe, ülejäänutest märksa suurema rühmadispersiooni

suhtes. Test-statistiku väärtus leitakse valemist

$$C = \frac{\max s_i^2}{\Sigma s_i^2} .$$

Selle testi jaoks on antud ka olulisuse tõenäosus p.

Teisena on märgitud Bartletti test, mis võtab aluseks statistiku M,

$$M = (n-k)\ln s^2 - \sum_{i=1}^{k}(n_i-1)\ln s_i^2,$$

sobivalt valitud funktsiooni, mis on ligilähedaselt F-jaotusega ning on tundlik rühmadispersioonide igasuguse mitteühtluse suhtes. Ka Bartletti testi jaoks on antud olulisuse tõenäosus.

Kolmanda, Hartley testi kohta on antud ainult teststatistiku väärtus, kuid puudub olulisuse tõenäosus. Selle kasutamiseks on tarvis eritabeleid.

Kui vähemalt ühe kriteeriumi alusel hüpotees H_1 dispersioonide erinevuse kohta osutub tõestatuks, muutub dispersioonanalüüsi tulemuse korrektsus problemaatiliseks ning selle asemel tuleks kasutada näiteks mitteparameetrilist dispersioonanalüüsi (vt p 7.4.7).

Näites toodud juhul on mõlemate testide olulisuse tõenäosused suuremad kui $\alpha = 0{,}05$, seega pole alust loobuda dispersioonanalüüsi tulemuste tõlgendamisest. Paneme siiski tähele, et rühmade dispersioonid on üpriski erinevad, eriti märkimisväärne on suurima dispersiooni (rühm fil) erinevus ülejäänutest.

7.4.6. Keskmiste mitmene võrdlemine. Järgmine rida *Multiple range test* võimaldab keskmiste mitmest võrdlemist. On ju dispersioonanalüüsi tulemus hüpoteesi H_1 vastuvõtmise korral küllaltki raskesti interpreteeritav: on küll selge, et faktoril teatav mõju on, aga see, missuguste rühmade keskmised erinevad üksteisest oluliselt, pole veel selge. Et täpsustada, missugused rühmad on omavahel erinevad, kasutatakse mitmese võrdluse meetodeid, mille abil rühmadest moodustatakse mitteeristuvate rühmade klassid.

Siin on mitu võimalust (valik nende vahel tehakse dispersioonanalüüsi tellimuspaneelil aknas *Range test* enne protseduuri käivitamist). Võimalused on järgnevad:

- usalduspiirid (vaikimisi),
- Tukey test (*T*),
- vähima olulise erinevuse test (*LSP*),
- Scheffe test (*S*).

Mitmese võrdlemise jaoks on tarvis fikseerida ka usaldusnivoo *Confidence level*, selle vaikimisi väärtuseks on 95 (protsenti). Keskmiste võrdlemise kohta vaata veel p. 10.5.5.

Protseduuri käivitamisel ilmub ekraanile tabel (vt jn 7.47), mis sisaldab päises kõigepealt meetodi nimetuse ja seejärel neli veergu

rühm *Level*,
arvukus *Count*,
keskmine *Average*,
homogeensed rühmad *Homogeneous Groups*.

Multiple range analysis for tulu by eriala

Method: 95 Percent Confidence Intervals

Level	Count	Average	Homogeneous Groups
med	10	443.50000	*
mat	7	470.00000	*
fil	10	491.00000	*

Jn 7.47

Rühmad on tabelisse paigutatud keskmiste kasvamise järjekorras.

Viimases veerus paikneb tärnikestest koosnev tabel, milles igale esialgsele rühmale vastab üks rida ja igale erineva koosseisuga klassile ehk homogeensele rühmale üks veerg (veergude arv võib olla 1 kuni k). Klassid on üldiselt osaliselt kattuvad, st nad võivad sisaldada ka ühiseid rühmi.

Tabeli (maatriksi) lahtrisse on märgitud tärnike * siis, kui reale vastav rühm kuulub veerule vastavasse klassi.

Ühte klassi kuuluvateks loetakse rühmad i_1, i_2, ..., i_r siis, kui nende puhul võetakse (vastava testi alusel) vastu nullhüpotees

$$H_0: \mu_{i_s} = \mu_{i_t}, \quad s,t = 1, \ldots, r,$$

st ühte klassi kuuluvate rühmade puhul on välistatud hüpoteesi

$$H_1: \mu_i \neq \mu_j$$

tõestamise võimalus (vaadeldava kriteeriumi alusel).

Joonisel 7.47 toodud näite puhul kuuluvad kõik rühmad ühte klassi.

Järgmised read aknast joonisel 7.39 annavad võimaluse tulem salvestada, need on

Save residuals - säilita prognoosijäägid,

Save intervals - säilita usaldusintervallid.

7.4.7. Kruskal-Wallise mitteparameetriline ühefaktoriline dispersioonanalüüs (*Kruskal-Wallis One-Way Analysis by Ranks*). Juhul kui dispersioonanalüüsi eeldused ei ole täidetud - uuritav tunnus on järjestustunnus, tema jaotus rühmades erineb oluliselt normaaljaotusest või on dispersioonid erinevad -, on korrektsem kasutada dispersioonanalüüsi mitteparameetrilisi analooge. Üks selline on Kruskal-Wallise meetod, mis paikneb allmenüüs *J. Analysis of Variance* neljandal kohal.

Selle protseduuri tellimuspaneeli (vt jn 7.48) täitmiseks on tarvis märkida uuritava tunnuse nimi (*Data*), rühmitava tunnuse nimi (*Level codes*) ja soovi korral ka rühmade nimetused (*Labels*).

Kruskal-Wallis One-Way Analysis by Ranks

Data: pere

Level codes: synnipaik

Labels:

Jn 7.48

Protseduuri eesmärgiks on kontrollida hüpoteesipaari

H_0: $med_1 = med_2 = \ldots = med_k$,

H_1: mingi i ja j korral $med_i \neq med_j$,

kus med_i tähistab uuritava tunnuse mediaani i-ndale rühmale vastavas üldkogumis.

Selleks järjestatakse kogu valim variatsioonritta ning omistatakse objektidele järjekorranumbrid - astakud.

Mitmele võrdsele objektile $y_i = y_{i+1} = \ldots = y_{i+g}$ (*tie*) omistatakse võrdne järjekorranumber, mis vastab nende järjekorranumbrite keskmisele

$$\frac{1 + (1+1) + \ldots + (1+g)}{g + 1} = 1 + \frac{g}{2} .$$

Kontrollstatistikuks on üksikutesse rühmadesse kuuluvate objektide keskmine astak: kui mediaanid langevad kokku, peaksid kokku langema ka keskmised astakud.

Arvutustulemused on koondatud tabelisse (vt jn 7.49) mis sisaldab kolm veergu:

rühma nimetus (*Level*),

objektide arv rühmas (*Sample Size*) ja

objektide keskmine järjekorranumber ehk astak rühmas (*Average Rank*).

Kruskal-Wallis analysis of pere by synnipaik

Level	Sample Size	Average Rank
alev	5	12.7000
linn	8	14.6250
maa	5	18.1000
suurlinn	10	13.5000

Test statistic = 1.43045 Significance level = 0.698413

Jn 7.49

Tabeli alla on märgitud test-statistiku väärtus (see on asümptootiliselt normaaljaotusega) ja olulisuse tõenäosus.

Joonisel 7.49 toodud näitest järeldub, et aspirantide pere keskmine suurus (täpsemalt, pere suuruse mediaan, st keskmise suurusega pere suurus) ei sõltu oluliselt aspirandi sünnipaigast, kuigi alevipere on ülejäänutest väiksem ja maapere - suurem.

§ 7.5. Kahe (pideva) arvtunnuse ühisanalüüs - regressioonanalüüs

Kuigi käesolevas paragrahvis käsitletavad meetodid on ette nähtud eeskätt pidevate arvtunnuste analüüsimiseks, on nende kasutamine mõeldav ka sel juhul, kui arvtunnused on diskreetsed, kuid mitte liiga vähese väärtuste arvuga.

Vaadeldavasse protseduuride rühma kuuluvad eeskätt regressioonanalüüsiga seotud meetodid.

Alustame lihtsaimast graafikust.

7.5.1. Korrelatsiooniväli ehk hajuvusdiagramm (*X-Y Line and Scatterplots*). See protseduur on esimene allmenüüs *E. Plotting Functions*, seega üldse kõige esimene statistikaprotseduur SG-süsteemis. Protseduuri eesmärgiks on esitada kahe tunnuse **X** ja **Y** abil kirjeldatud valim punktihulgana xy-koordinaadistikus. Sellise punktihulga graafikut nimetatakse korrelatsiooniväljaks e hajuvusdiagrammiks.

Korrelatsioonivälja tellimuspaneel on esitatud joonisel 7.50. Siin on kõigepealt tarvis märkida tunnuste nimed (horisontaalteljele **X** ja vertikaalteljele **Y**).

X-Y Line and Scatterplots

	Log
X: kasv	No
Y: kaal	No

Point labels:

Point codes:
Point colors:
Std. err. X:
Std. err. Y:

Points: Yes
Lines: No

Jn 7.50

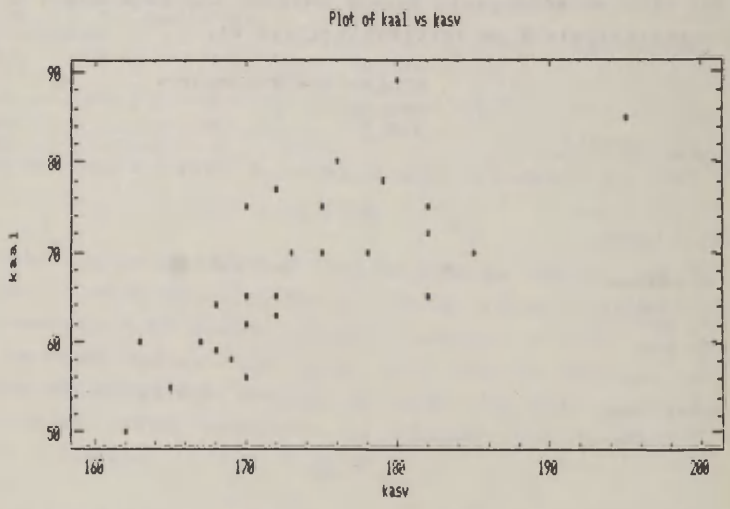
Korrelatsioonivälja varieerimiseks on veel järgmised võimalused:

1) Kasutada kas ühe või mõlema tunnuse väärtuste teisendamist logaritmilise skaala järgi (tellimuspaneelil tunnuse nimetuste rea parempoolses otsas aken *Log* väärtustega *No/Yes*, vaikimisi *No*).

2) Muuta korrelatsiooniväli sisuliselt kolme tunnuse ühisjaotuse analüüsimise vahendiks, võttes kasutusele kolmanda (diskreetse sümbol-) tunnuse, mille väärtused (väärtuste järjekorranumbrid) trükitakse punktide asemele (*Point codes*).

3) Punktidele võib anda spetsiaaltähised (*Point labels*), kasutades (eelnevalt salvestatud) nimede tunnust ja interaktiivse redigeerimise režiimi (vt [7], 2.8.4), samuti muuta nende värvi (*Point colors*).

4) Punktidele võib lisada standardvea piirid (x- või/ /ja y-telje sihis). Selleks peavad iga punkti kohta olema salvestatud vajalikud andmed sobiva pikkusega (n) tunnusena. Tellimuse vormistamisel trükitakse aknasse *Std. err. X* ja/ /või *Std. err. Y* punktide standardhälbeid sisaldava(te) tunnus(t)e nimi (nimed).



Jn 7.51

5) Korrelatsiooniväljal olevad punktid võib ka ühendada sirglõikudega vastavalt nende järjestusele valimis (1 = = 1, ..., n). Selleks tuleb aknasse *Lines* trükkida *Yes*. Märgime, et enamikul juhtudest ei paranda selline täiendus korrelatsioonivälja interpreteerimist, pigem vastupidi. Sellisel joongraafikul on mõtet üksnes siis, kui vähemalt mõni tunnus oluliselt sõltub mõõtmisajast, või siis, kui valim on eelnevalt järjestatud.

6) On võimalik ka loobuda punktidest korrelatsiooniväljal (sellisel graafikul on mõtet muidugi vaid siis, kui punkte tähistab mõni teine graafiline element, näiteks ühendusjoon või standardvea piiridest moodustunud rist).

Kasvu ja kaalu korrelatsioonivälja esitab joonis 7.51.

7.5.2. Regressioonanalüüs (*Simple Regression*). Regressioonanalüüs võimaldab konstrueerida funktsionaalse mudeli pideva tunnuse prognoosimiseks pideva(te) tunnus(t)e kaudu.

Juhtu, kus argumente (sõltumatuid tunnuseid, regressoreid) on ainult üks, nimetatakse lihtregressiooniks.

Lihtregressiooni (*Simple Regression*) protseduur paikneb esimesel kohal allmenüüs *K. Regression Analysis*.

Selle protseduuri tellimuspaneeli kujutab joonis 7.52. Tellimuspaneelile tuleb märkida funktsioontunnuse **Y** (*Dependent variable*), seejärel argumenttunnuse **X** (*Independent variable*) nimed. Mõlemad tunnused peavad olema arvulised ning ühesuguse väärtuste arvuga.

Simple Regression

Dependent variable: kaal

Independent variable: kasv

Model: Linear

Jn 7.52

Seejärel tuleb valida mudeli tüüp järgmisest loetelust:

- lineaarne (*Linear*), $Y = a + bX$;
- multiplikatiivne (*Multiplicative*), $Y = aX^b$;
- eksponentsiaalne (*Exponential*), $Y = \exp(a + bX)$;
- pöördvõrdeline (*Reciprocal*), $Y = 1/(a + bX)$.

Vaikimisi väärtuseks on lineaarne mudel, millega ongi kõige loomulikum analüüsi alustada.

Järgmiseks tuleb märkida usaldustõenäosused

- regressioonijoone usalduspiirkonna (*Confidence limits*),
- väärtuste paiknemise oodatava piirkonna (*Prediction limits*)

konstrueerimiseks. Vaikimisi väärtuseks on 95 %.

On ka võimalik graafikul paiknevaid punkte märgendada. Selleks tuleb sisestada punktide nimetuste loetelu (nominaaltunnus) aknasse *Point labels*. Märgendamine toimub interaktiivse redigeerimise režiimis (vt [7], p 2.8.4).

Regressioonianalüüsi protseduuri käigus on tarvis:

1. Hinnata mudeli parameetreid (näiteks lineaarse regressiooni korral hinnatakse parameetreid a ja b).

2. Kontrollida hüpoteese nende parameetrite olulisuse kohta. Kui ϑ on mudeli parameeter, siis kontrollitakse hüpoteesipaari

$$H_0: \vartheta = 0,$$
$$H_1: \vartheta \neq 0.$$

Esimesel juhul öeldakse, et parameeter ei ole statistiliselt oluline, tihti järeldub sellest ka kogu mudeli mitteolulisus.

3. Hinnata mudeli headust, st seda, kui suure osa sõltuva tunnuse **Y** hajuvusest mudel kirjeldab.

7.5.3. Regressioonanalüüsi tulemused. Protseduuri käivitamise tulemusena ilmub ekraanile kaks tabelit (vt jn 7.53). Esimene nendest on regressioonanalüüsi tabel, mille pealkirjas näidatakse ära ka kasutatud mudeli tüüp, samuti sõltuva ja sõltumatu tunnuse (**Y** ja **X**) nimed. Järgneb 5-veeruline tabel, milles tuuakse järjest ära

1) hinnatava parameetri nimetus (tähendus),
2) parameetri hinnang $\tilde{\vartheta}$,
3) parameetri hinnangu standardviga $s_{\tilde{\vartheta}}$,
4) t-statistiku väärtus $\tilde{\vartheta}/s_{\tilde{\vartheta}}$,
5) t-statistiku olulisuse tõenäosus.

Tabelis on kaks rida, esimene vastab vabaliikmele a (*Intercept*), teine regressioonikordajale b (*Slope*).

Regression Analysis - Linear model: Y = a+bX

Dependent variable: kaal Independent variable: kasv

Parameter	Estimate	Standard Error	T Value	Prob. Level
Intercept	-91.737	27.7803	-3.30223	.00279
Slope	0.919144	0.159737	5.75413	.00000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	Prob. Level
Model	1342.6400	1	1342.6400	33.110	.00000
Error	1054.3243	26	40.5509		
Total (Corr.)	2396.9643	27			

Correlation Coefficient = 0.748426 R-squared = 56.01 percent
Stnd. Error of Est. = 6.36796

Jn 7.53

Regressioonanalüüsi tabeli all on meile juba tuttav dispersioonanalüüsi tabel (vt ka jn 7.38). Veergude tähendus on sama, ainult ruutude summa SS,

$$SS = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

on jaotatud kaheks osaks vastavalt lineaarsele mudelile,

$$y_i = a + bx_i + \varepsilon_i, \qquad (7.15)$$

ehk

$$\tilde{y}_i = a + bx_i,$$

kus x_i on 1-ndal objektil mõõdetud argumendi X väärtus, $\tilde{y}_i$ aga prognoositud väärtus ja ε_i prognoosijääk (viga).

Tähistades

$$SS_1 = \sum_{i=1}^{n} (\tilde{y}_i - \bar{y})^2,$$

$$SS_0 = \sum_{i=1}^{n} (\tilde{y}_i - y_i)^2,$$

saame mudelile vastava hajuvuskomponendi SS_1 (*Model*) ja jäägile vastava hajuvuskonmponendi SS_0 (*Error*). Mudeli vabadusastmete arv on ühe võrra väiksem hinnatavate parameetrite

arvust r, seega käesoleval juhul 2 - 1 = 1. Jäägi vabadusastmete arv on n - r. Dispersioonanalüüsi tabeli põhjal saab teha järeldusi mudeli kui terviku olulisuse kohta, vt p 7.4.4.

Dispersioonanalüüsi tabeli all on korrelatsioonikordaja (*Correlation Coefficient*) r (vt ka p 7.2.4),

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}, \qquad (7.18)$$

kus

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

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

Korrelatsioonikordaja ruutu r^2 (*R-squared*) nimetatakse determinatsioonikordajaks ning see esitatakse tavaliselt protsentides. Determinatsioonikordaja näitab, kui suure osa funktsioontunnuse Y hajuvusest kirjeldab vaadeldav mudel.

Tabeli viimases reas on esitatud veel $s_o = \sqrt{SS_o/(n-r)}$ (*Stnd. Error of Est.*), s.o prognoosi täpsust iseloomustav standardhälbe hinnang.

Joonisel 7.53 toodud tabeli järgi saame välja kirjutada mudeli tunnuse kaal prognoosimiseks tunnuse kasv järgi:

$$\text{kaal} = -91{,}7 + 0{,}92 \cdot \text{kasv}.$$

Dispersioonanalüüsi tabelist näeme, et seos on statistiliselt oluline ning mudel kirjeldab 56 % tunnuse kaal hajuvusest. Esimese tabeli põhjal veendume, et mudelis on mõlemad parameetrid a ja b statistiliselt olulised.

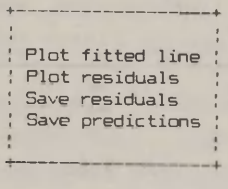
Muutes mudeli kuju, võime saada mõnevõrra erinevad tulemused; esitame need tabelis 7.3.

Nagu näha, on mudeli keerukamaks muutmisest saadav võit tühine, maksimaalselt 1,2 %. Olgu märgitud, et nii väike erinevus lineaarsest mudelist pole joonisel peaaegu üldse nähtav.

T a b e l 7.3

Mudeli tüüp	Mudeli kuju	Kirjeldatuse %
Lineaarne mudel	$Y = -91,7 + 0,92X$	56,0
Multiplikatiivne mudel	$Y = e^{-8,22} \cdot X^{2,41}$	57,2
Eksponentsiaalne mudel	$Y = e^{1,877+0,0136X}$	56,6
Pöördväärtuste mudel	$Y = \frac{1}{0,0504 - 0,000204X}$	56,4

7.5.4. Regressioonanalüüsi täiendavad võimalused. Täiendavad võimalused saab valida aknast, vt jn 7.54.



Jn 7.54

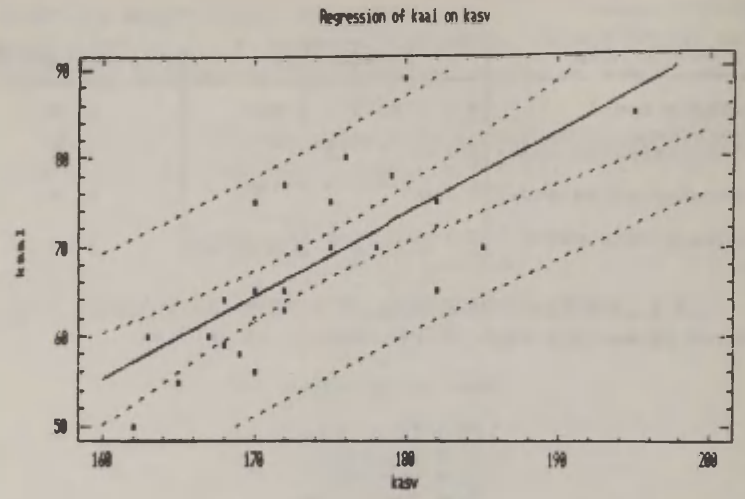
Esimene rida (*Plot fitted line*) annab graafiku, millel korrelatsiooniväljale on kantud mudelit kirjeldav regressioonijoon. Lineaarse mudeli puhul on regressioonijooneks sirge, vt jn 7.55.

Lisaks regressioonijoonele on graafikule kantud ka selle joone usalduspiirkond, st piirkond, milles (tõenäosusega 0,95) paikneb üldkogumit iseloomustav regressioonisirge (see on regressioonijoonele lähemal paiknev joontepaar).

Teine, kaugemal paiknev joontepaar määrab piirkonna, milles tõenäosusega 0,95 paiknevad üldkogumi punktid; ka iga valimi punktidest paikneb selles piirkonnas keskmiselt 95 %.

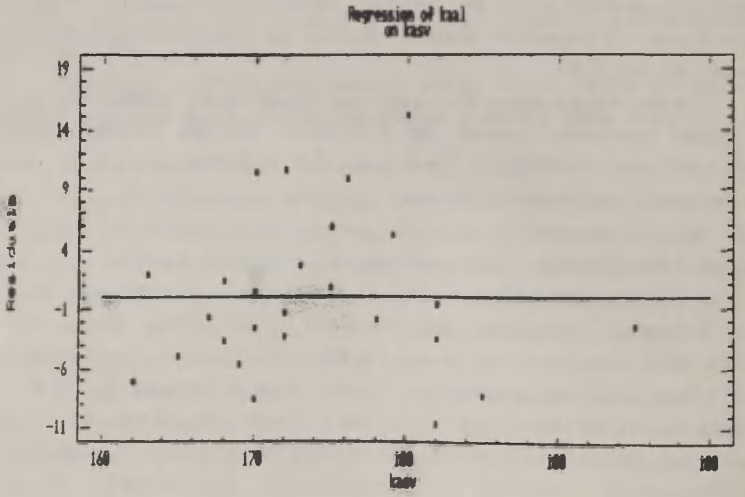
Ekraanil, graafiku all paikneb vasakpoolne sulg. Sellesse võib trükkida argumendi mingi väärtuse x_0; käivitamisel ilmub sulgudesse vastav Y väärtuse prognoos $\tilde{y}$, mis on arvutatud valemist (7.15). Ühtlasi ilmub see punkt koos oma abstsiss- ja ordinaatlõikudega (punktist telgedeni ulatuvate ristlõikudega) ekraanile, loomulikult tingimusel, et ta paikneb ekraanil kujutatud tasandiosal.

Nimetatud tegevust võib korrata.



Jn 7.55

Teine valik (*Plot residuals*) annab graafiku, kus y-teljele kantakse prognoosijääägid $y_i - \tilde{y}_i$, vt jn 7.56.



Jn 7.56

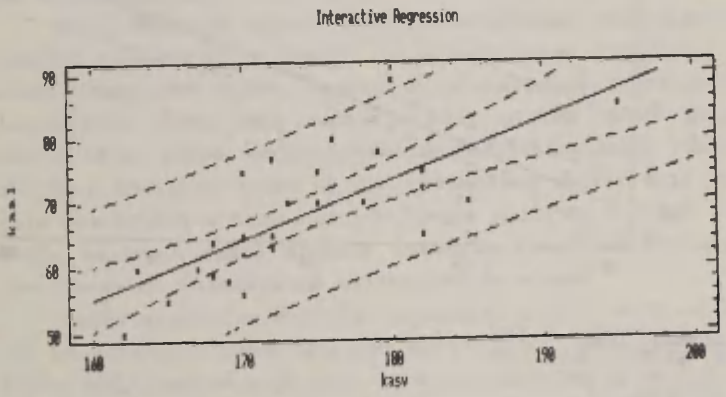
Prognoosijääkide graafik annab teatavat informatsiooni selle üle otsustamiseks, kas mudel kirjeldab tunnust Y täielikult (sel juhul paiknevad prognoosijäägid graafikul juhuslikult) või mitte (siis peaks prognoosijääkide paigutusse ilmuma selge korrapära).

7.5.5. Interaktiivne regressioon (*Interactive Outlier Rejection*). Teatavasti sõltub regressioonimudel sellesse kuuluvate punktide hulgast, kusjuures eriti võivad mudelit mõjutada mõned ülejäänutest tugevasti erinevad punktid (*Outliers*), mille puhul võib oletada ka jämedate mõõtmisvigade olemasolu. Interaktiivse regressiooni protseduur on ette nähtud selleks, et uurida mudeli käitumist üksikute punktide ärajätmisel valimist.

Interaktiivse regressiooni protseduur on teine allmenüüs *K. Regression Analysis*. Selle, nagu ka regressioonanalüüsi puhul peavad analüüsitavad tunnused olema arvulised.

Protseduuri tellimuspaneelil tuleb märkida funktsioon- ja argumenttunnuste nimed.

Seejärel ilmub ekraanile graafik (vt jn 7.57), mis langeb praktiliselt kokku joonisel 7.55 esitatud lineaarse

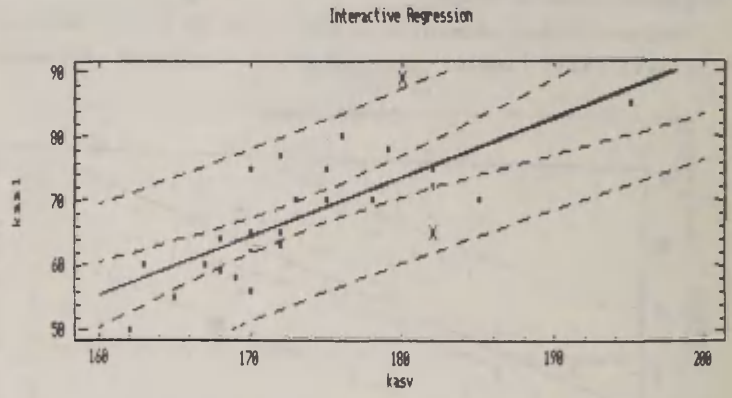


B0: -91.737 SE: 27.78 T: -3.3022
B1: 0.91914 SE: 0.15974 T: 5.7541
CORR: 0.74843 MSE: 48.551 DF: 26
POINTS DELETED:

Jn 7.57

regressioonijoone graafikuga, joonise alla on aga trükitud olulisem informatsioon regressioonivõrrandi kohta: sümboli *B0* järel vabaliikme väärtus, tema .standardviga ja t-statistiku väärtus, sümboli *B1* järel samad andmed regressioonikordaja kohta. Kolmandas reas on korrelatsioonikordaja (*C*orr.), hinnang s_0^2 (vt dispersioonanalüüsi tabel joonisel 7.53) ja sellele vastav vabadusastmete arv. Neljandas reas on ainult päis - välja on jäetud punktid (*Points deleted*); kuivõrd punkte välja jäetud veel ei ole, on see rida edasi tühi.

Graafiku iseärasuseks on sellel paiknev ristikujuline kursor, mis on liigutatav tavaliste kursorit juhtivate nooleklahvide abil. Kui kursor on mõne punkti peale (lähedusse) viidud, siis klahvile **E** vajutamise tulemusena kaob kursorile lähim punkt ja see asendatakse põikristikesega (punkt on maha tõmmatud, vt jn 7.58). Seda protseduuri saab korrata, ka saab kustutatud punkti "tagasi panna", liikudes kursoriga ta peale ja vajutades klahvi I.



B0: -90.684 SE: 24.53 T: -3.6936
B1: 0.91155 SE: 0.1415 T: 6.4421
CORR: 0.79598 MSE: 29.513 DF: 24
POINTS DELETED: 12 17

Jn 7.58

Kui valim on sobival viisil ümber korraldatud, käivitatakse protseduur võtme **F6** abil ning joonise alla ilmuvad ümber arvutatud mudeli parameetrid (vt jn 7.58), vastavalt

muutub ka joonis. Joonisel 7.58 on "välja visatud" kaks punkti, need on järjekorranumbritega 12 ja 17, üks neist kaugeim, 95 % piirist väljaspool asuv, teine regressioonisirgele veidi lähemal.

Väljajätmise tulemusena on regressioonimudeli kõik näitajad mõnevõrra paranenud. Joonisele on kantud ka uus regressioonijoon, see erineb siiski esialgsest vähe.

Tuleb aga märkida, et kuigi nii on võimalik mudelit parandada, saab uurija eetika seisukohast niisugust sammu harva õigustada. Üksnes siis on see samm õige, kui sisuliste kriteeriumide põhjal on selge, et väljajäetavad punktid ei kuulu uuritavasse kogumisse, vaid on nn jämedad vead.

§ 7.6. Mudelid sagedustabelite kirjeldamiseks ja analüüsimiseks

Käesolevas paragrahvis pöördume tagasi paragrahvis 7.2 vaadeldud sagedustabelite juurde ja tutvume nende analüüsimiseks kasutatavate mudelitega. Kõikjal eeldame, et tunnused X ja Y on diskreetsed (arvulised või mittearvulised), tunnuse X väärtuste arv on k, tunnuse Y väärtuste arv h.

Mudel kujutab endast üldiselt mingitele eeldustele tuginevat lihtsustatud seost tunnuste/tunnuse väärtuste või nende sageduste vahel. Sageli on otstarbekas konstrueerida terve rida võimalikke mudeleid ning otsida nende hulgast välja kõik antud objekti/nähtuse kirjeldamiseks sobivad. Mõnikord õnnestub leida ka üks optimaalne mudel, kuid alati pole see võimalik ning vastuvõetavate mudelite hulgast tuleb leida sobivaim mudel mingite (näiteks sisulisest interpreteeritavusest tulenevate) kriteeriumide alusel.

Sagedustabelite analüüsi metoodika on üldiselt rakendatav ka kõrgemat järku tabelite korral, kuid käesolevas peatükis käsitleme me seda vaid kahe tunnuse puhul.

7.6.1. Sagedustabelit kirjeldavad mudelid. Mudelite keerukus sõltub üldiselt selles sisalduvate parameetrite arvust. Alustame seepärast lihtsaimast võimalusest, kasutades punktis 7.2.1 kasutusele võetud tähistusi.

1° Tunnused X ja Y on diskreetse ühtlase jaotusega ning sõltumatud. Siis

$$n_{ij} = \frac{n}{kh} \tag{7.17}$$

ehk

$$\ln n_{ij} = \nu, \tag{7.17'}$$

kus $\nu = \ln n - \ln kh$.

Loobume nüüd järk-järgult kitsendavatest eeldustest.

2° Tunnus Y on diskreetne ühtlase jaotusega, tunnused on sõltumatud. Siis

$$n_{ij} = \frac{n_{i.}}{h}$$

ehk

$$\ln n_{ij} = \nu + \alpha_i \quad (i = 1, \ldots, k),$$

kus $\alpha_i = \ln n_{i.} - \ln (n/h)$.

3° Tunnus X on diskreetse ühtlase jaotusega, tunnused on sõltumatud. Siis

$$n_{ij} = \frac{n_{.j}}{k}$$

ehk

$$\ln n_{ij} = \nu + \beta_j \quad (j = 1, \ldots, h),$$

kus $\beta_j = \ln n_{.j} - \ln (n/k)$.

4° Tunnused X ja Y on sõltumatud. Siis

$$n_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}$$

ehk

$$\ln n_{ij} = \nu + \alpha_i + \beta_j \quad (i = 1, \ldots, k;\ j = 1, \ldots, h),$$

kus α_i ja β_j on juhtudel 2° ja 3° defineeritud väärtustega.

Lõpuks saame välja kirjutada ka kõige üldisema mudeli, mis kitsendusi ei sisalda.

5°

$$\ln n_{ij} = \nu + \alpha_i + \beta_j + \gamma_{ij}, \tag{7.18}$$

kus $\gamma_{ij} = \ln n_{ij} + \ln n - \ln n_{i.} - \ln n_{.j}$.

Märgime, et mudelites 2°-4° esitatud liikmeid nimetatakse esimest järku mõjudeks. Mudelisse 5° lisandub ka üks teist järku mõju, nimelt γ_{ij}. Üldiselt määrab mõju järgu see, mitme tunnuse väärtusi temas arvestatakse.

Mudelite konstrueerimiel tuleb meil lahendada järgmised põhiküsimused:

1) hinnata mudeli parameetreid (ν, α_i, β_j, γ_{ij} vastavalt mudeli tüübile),

2) kontrollida mudeli sobivust.

Selleks arvutatakse vastavalt mudelile teoreetilised sagedused n^*_{ij} ning kontrollitakse hüpoteesipaari

$$H_0: n_{ij} = n^*_{ij}, \quad i = 1, \ldots, k, \; j = 1, \ldots, h,$$
$$H_1: \text{mingi } i \text{ ja } j \text{ korral } n_{ij} \neq n^*_{ij}.$$

Nimetatud hüpoteesipaari kontrollimiseks kasutatakse kõige sagedamini χ^2-statistikut kujul

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{h} \frac{(n^*_{ij} - n_{ij})^2}{n^*_{ij}}, \tag{7.19}$$

kusjuures vabadusastmete arv f avaldub seosega

$$f = kh - r,$$

kus r on hinnatavate parameetrite arv.

Märgime, et tegemist on meile juba tuntud valemiga (7.11), mille korral $n^*_{ij} = (n_{i.} n_{.j})/n$.

Valemiga (7.19) arvutatavat χ^2-statistikut nimetatakse ka Pearsoni χ^2-statistikuks. Selle kõrval kasutatakse tihti ka teist, samuti asümptootiliselt χ^2-jaotusega statistikut kujul

$$\tilde{\chi}^2 = 2 \sum_{i=1}^{k} \sum_{j=1}^{h} n_{ij} \ln \left(\frac{n_{ij}}{n^*_{ij}} \right). \tag{7.20}$$

Statistiku $\tilde{\chi}^2$ kohta ütleme edaspidi, et see on tõepärasuhte χ^2; vabadusastmete arv arvutatakse selle jaoks samal viisil kui Pearsoni χ^2-statistiku jaoks.

7.6.2. Log-lineaarse analüüsi protseduur kahe tunnuse puhul (*Log-Linear Analysis*). Log-lineaarne analüüs (sagedustabelite analüüsimiseks) on neljas protseduur allmenüüs *P. Categorical Data Analysis*. See protseduur võimaldab põhimõtteliselt ka kõrgema dimensiooniga tabeleid analüüsida, käesolevas punktis piirdume aga kahedimensionaalsetega.

Protseduuri tellimuspaneeli kujutab joonis 7.59.

Log-Linear Analysis

Frequency table: TABLE

Label matrix: LEGENDS

Generators:

Effects:

Convergence criterion: 0.01

Maximum number of iterations: 20

Jn 7.59

Tellimuse esitamine toimub järgmiselt.

1° Aknasse *Frequency Table* märgitakse analüüsitav tabel, see võib olla salvestatud näiteks protseduuri *Crosstabulation* tulemusena.

2° Aknasse *Label Matrix* märgitakse soovi korral tunnus(t)e väärtuste nimetused/koodid, ka see maatriks on salvestatav tabeli valmistamise käigus (vt p 7.2.4).

3° Aknasse *Generators* tuleb märkida, milliste tunnuste mõju sagedustabelile soovitakse uurida, st milliste tunnuste suhtes loobutakse eeldusest, et nad on diskreetse ühtlase jaotusega. Tunnused tähistatakse siin järjest tähestiku algustähtedega A, B, ...

4. Aknasse *Effects* tuleb märkida, milliseid mõjukomponente täielikust mudelist (7.18) soovitakse uurida (st mille

kohta ei eeldata nulliga võrdumist). Tähistus on siin järgmine:

$$\alpha_i \rightarrow A, \quad \beta_j \rightarrow B, \quad \gamma_{ij} \rightarrow AB.$$

5° Kuivõrd parameetrite ν, α_i, β_j, γ_{ij} leidmine toimub üldjuhul iteratiivselt, tuleb tellimuspaneelil näidata ära ka protsessi koonduvust juhtivad parameetrid - koonduvuskriteerium (*Convergence criterion*) ja maksimaalne iteratsioonide arv (*Maximum number of iterations*).

Vaikimisi määratakse parameetritele järgmised väärtused:

- parameetrile *Generators* tühi hulk, st kontrollitakse lihtsaimat mudelit (7.17); vastavalt sellele on loomulikult tühi ka mõjude lahter *Effects*;

- kui generaatorite lahter on täidetud, siis paigutatakse lahtrisse *Effects* vaikimisi nende generaatorite kõikvõimalikud mõjud ja mõjude kombinatsioonid;

- koonduvuskriteerium on 0,01 ja iteratsioonide arv 20; enamasti on need väärtused sobivad ja ei vaja asendamist.

Joonisel 7.59 on uuritavaks tabeliks võetud tunnuste synnipaik ja eriala jaotustabel, mis on salvestatud nimega **TABLE** (vt jn 7.10). Tunnuste väärtuste loetelud on salvestatud sümboltunnusena (maatriksina) **LEGENDS**. Muus osas tellimusest on kasutatud vaikimisi väärtusi.

Protseduuri tulemused esitatakse kõigepealt standardse tabelina (vt jn 7.60).

Tabeli pealkirjaks on "Valim- ja teoreetiliste sageduste võrdlus" (*Comparison of observed and expected frequencies*). Tabeli päises esitatakse generaatorite loetelu, tabelis endas on kuus veergu:

1) faktorite (tunnuste väärtuste ja nende kombinatsioonide) loetelu (*Factor*);

2) vastavate tunnuste väärtuste nimetused (*Level*);

3) valimi empiirilised sagedused ja suhtelised sagedused - viimased on esitatud protsentides ja paigutatud sulgudesse (*Observed Freq.*);

4) teoreetilised sagedused vastavalt oletatavale mudelile (*Expected Freq.*);

5) valim- ja teoreetilise sageduse vahe ehk jääk e;

6) standardiseeritud jääk e/[s(e)] (*Std. Resid.*).

Comparison of observed and expected frequencies

Generator:

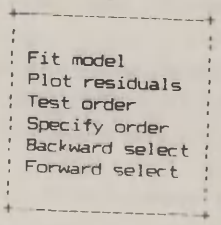
Factor Level	Observed Freq.	Expected Freq.	Residual	Std. Resid.
Factor A: alev				
Factor B: fil	1 (3.57)	2.3 (8.33)	-1.3	-0.873
Factor B: mat	1 (3.57)	2.3 (8.33)	-1.3	-0.873
Factor B: med	3 (10.71)	2.3 (8.33)	0.7	0.436
Factor A: linn				
Factor B: fil	2 (7.14)	2.3 (8.33)	-0.3	-0.218
Factor B: mat	5 (17.86)	2.3 (8.33)	2.7	1.746
Factor B: med	1 (3.57)	2.3 (8.33)	-1.3	-0.873
Factor A: maa				
Factor B: fil	2 (7.14)	2.3 (8.33)	-0.3	-0.218
Factor B: mat	2 (7.14)	2.3 (8.33)	-0.3	-0.218
Factor B: med	1 (3.57)	2.3 (8.33)	-1.3	-0.873
Factor A: suurlinn				
Factor B: fil	5 (17.86)	2.3 (8.33)	2.7	1.746
Factor B: mat	0 (0.00)	2.3 (8.33)	-2.3	-1.528
Factor B: med	5 (17.86)	2.3 (8.33)	2.7	1.746

Jn 7.60

Kuivõrd tabelis joonisel 7.59 esitatud tellimuses ühtegi generaatorit ei ole märgitud, on eeldatavaks mudeliks diskreetne ühtlane jaotus (mudel 1°). Seega teoreetilisteks sagedusteks on kõikides lahtrites 28/12 = 2,33 objekti ehk 1/12 = 8,33 % vaatluste koguarvust.

Järgmised töötlusetapid toimuvad vastavalt valikule aknast joonisel 7.61.

Märgime, et mudeli parameetreid on võimalik hinnata ainult siis, kui lähtetabelis ei ole ühtegi nullsagedust (tühja lahtrit). Vastasel korral jäävad paratamatult tegemata ka järgmised, parameetrihinnangutele tuginevad töötlussammud.



Jn 7.61

Esimene rida - *Fit model* - tähistab hüpoteesipaari kontrollimist:

H_0: uuritav tabel on esitatud teoreetilise mudeliga kooskõlas,

H_1: uuritav tabel ei ole esitatud teoreetilise mudeliga kooskõlas.

Selle rea valimisel saadakse joonisel 7.62 esitatud tabel.

Summary of fitted model

Maximum Cell Difference: 0 Number of iterations: 2

Goodness-of-fit test statistics for model adequacy

	Statistic	d.f.	P-value
Likelihood ratio chi-square	15.7439	11	.1509
Pearson chi-square	14.8571	11	.1891

Jn 7.62

Tabelis on veerus *P-value* trükitud olulisuse tõenäosused, mille põhjal saab kontrollida hüpoteesi H_1 uuritava tabeli ja teoreetilise mudeli kooskõla kohta.

Vaatleme veel tunnuste **sugu** ja **eriala** ühisjaotust, mida esitab järgmine tabel.

Sugu \ Eriala	fil	mat	med
1	6	2	4
2	4	6	6

Mudeli 1° kontrollimisel saame sobivuse nullhüpoteesiga; kuivõrd siin tühje lahtreid ei ole, on siin hinnatud ka mudeli parameetreid (praegu on neid ainult üks).

Hindamistulemused on esitatud tabelis joonisel 7.63 pealkirjaga *Estimated model coefficients* - "mudeli kordajate hindamine". Tabelis on veerud *Factor* (mõju), *Coeff.* (kordaja), *Std. Err.* (koradaja standardviga), selle kordaja standardiseeritud väärtus *Z-value* ning alumine ja ülemine usalduspiir *Lower 95 CI* ja *Upper 95 CI*. Viimaste abil on lihtne

kontrollida parameetri olulisust: kui 0-punkt ei kuulu usalduspiirkonda (st ülemine ja alumine usalduspiir on samamärgilised), siis võetakse vastu sisukas hüpotees

H_1: see parameeter erineb nullist,

vastasel korral parameeter ei erine oluliselt nullist ehk on mitteoluline.

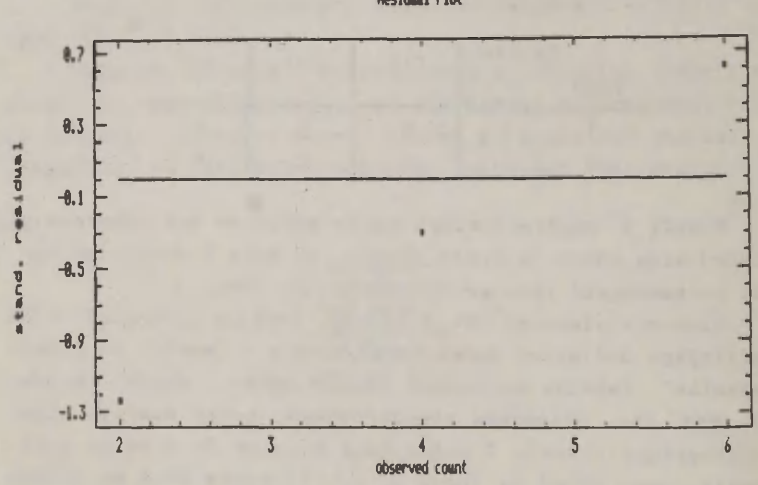
Estimated model coefficients

Generator:

Factor	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
	1.54045	0.20412	7.54661	1.14036	1.94053

Jn 7.63

Aknast joonisel 7.61 on võimalik valida ka standardne jääkide graafik (*Plot residuals*), vt jn 7.64. Järgmine võimalus on kontrollida mõjude olulisust mudelis (*Test order*), sellele järgneb võimalus muuta mudeli järku. Viimased kaks rida võimaldavad mudeli automaatset valikut, mis on aga efektiivne kõrgemat järku tabeli korral (lähemalt vt p 8.3.2).



Jn 7.64

7.6.3. Mediaanidele baseeruv kahemõõtmelise tabeli analüüs (*Median Polish of Two-Way Tables*). Sagedustabeli analüüsimiseks mediaanide põhjal on SG-süsteemis protseduur *Median Polish of Two-Way Tables*, mis paikneb allmenüüs *I. Exploratory Data Analysis* neljandal kohal.

Protseduuri aluseks on koosmõjuta mudel (vt ka valemit (7.18))

$$n_{ij} = \nu + \alpha_i + \beta_j,$$

kusjuures parameetrite ν, α_i ja β_j väärtusi hinnatakse interaktiivselt.

Joonisel 7.65 on sisestatud maatriks **TABLE** jooniselt 7.10. Käivitamisel saadud tulemusi esitavad joonised 7.66 ja 7.67).

Median Polish

Data matrix: TABLE

Number of iterations: 10

Save polished table
Save row effects
Save column effects

Jn 7.65

Polished table for TABLE after 10 it

	Col 1	Col 2	Col 3
Row 1	.0000	.0000	2.50000
Row 2	.0000	3.0000	-.50000
Row 3	.0000	.0000	-.50000
Row 4	.0000	-5.0000	.50000

Jn 7.66

Joonisel 7.66 on esitatud prognoosijääkide w_{ij} tabel,

$$w_{ij} = n_{ij} - \nu - \alpha_i - \beta j.$$

```
Typical value: 2

Row effects for TABLE after 10 iterations.
-1 0 0 3

Column effects for TABLE after 10 iterations.
0 0 -0.5
```

Jn 7.67

Joonisel 7.67 on esitatud mõjusid kirjeldavate parameetrite ν, α_i, β_j hinnangud. Aknas (vt jn 7.65) on pakutud salvestamise võimalusi jääkide tabeli (*polished table*), veeru- ja reamõjude (*row effects*, *column effects*) jaoks.

Näeme, et käesolevas näites on $\nu = 2$ (keskmine objektide arv sagedustabeli lahtris). Ridade **mõjud** on -1, 0, 0 ja 3, st esimeses reas (**alev**) on objekte ühe võrra vähem kui ülejäänutes, viimases (**suurlinn**) aga kolme võrra rohkem. Veergudele vastavad mõjud on 0, 0 ja -0,5, st ainult viimase veeru igas lahtris on objekte keskmiselt 0,5 võrra vähem kui mujal.

Seega mõjude järgi arvutatud tabel oleks järgmine:

2 - 1	2 - 1	2 - 1 - 0,5
2	2	2 - 0,5
2	2	2 - 0,5
2 + 3	2 + 3	2 + 3 - 0,5

Võrreldes saadud tabelit joonisel 7.10 esitatuga, näeme, et tabelite vahe annab tõepoolest joonisel 7.66 esitatud tabeli, millest järeldub, et vaadeldava tabeli kõrvalekalded sõltumatuse eeldusele rajatud mudelist on üsna märgatavad, eriti teises veerus. Ranget hüpoteeside kontrolli see protseduur ei võimalda.

8. MITME TUNNUSE ÜHISJAOTUSE ANALÜÜS JA NÄITLIKUSTAMINE

Mitmemõõtmelise statistilise analüüsi metoodika tegeleb tunnushulga kirjeldamise, prognoosimise ja objektide struktuuri analüüsimisega juhul, kui tunnuste arv m on suvaline kahest suurem lõplik arv. Sel juhul tavaliselt tunnuste ühisjaotust ei kasutata, vaid lähtutakse teatavatest ühisjaotuse põhjal arvutatud statistikutest (nt korrelatsioonimaatriksist vmt) ning püütakse leida mingit mudelit andmestiku kirjeldamiseks.

Ometi pakub ka tunnuste ühisjaotus huvi, eriti sel juhul, kui tunnuste arv on 3-4. Niisuguse ühisjaotuse näitlikustamiseks ja analüüsimiseks on välja töötatud rida meetodeid, mida standardse mitmemõõtmelise statistilise analüüsi hulka kuuluvaks ei loetagi, vaid mis on enamasti teatavateks üldistusteks vastavatele kahemõõtmelise analüüsi meetoditele. Et mõned sellised mudelid on realiseeritud ka süsteemis SG, pühendamegi käesoleva peatüki, mis eelneb klassikalise mitmemõõtmelise statistilise analüüsi käsitlusele, nendele mudelitele.

§ 8.1. Kolme ja nelja tunnuse ühisjaotuse graafiline esitus

Kaasaegne arvutustehnika võimaldab ruumilisi kujundeid näitlikustada arvutil dünaamiliste (nt pöörlevate) kujunditena. Niisuguseid võimalusi SG-süsteem ei paku, küll aga kasutatakse mitmesuguseid projektsioone, värve, erinevaid joone jne ruumilisuse efekti tekitamiseks ja kujundi näitlikustamiseks.

8.1.1. Kolme tunnuse korrelatsiooniväli ehk hajuvusdiagramm (*X-Y Line and Scatterplots*). Korrelatsioonivälja (vt p 7.5.1) kirjeldamisel nimetasime punktide märgendamise võimalust. Kui selleks kasutada kolmandat tunnust (mis peaks olema diskreetne), saame sisuliselt kolme tunnuse ühisjaotuse.

X-Y Line and Scatterplots

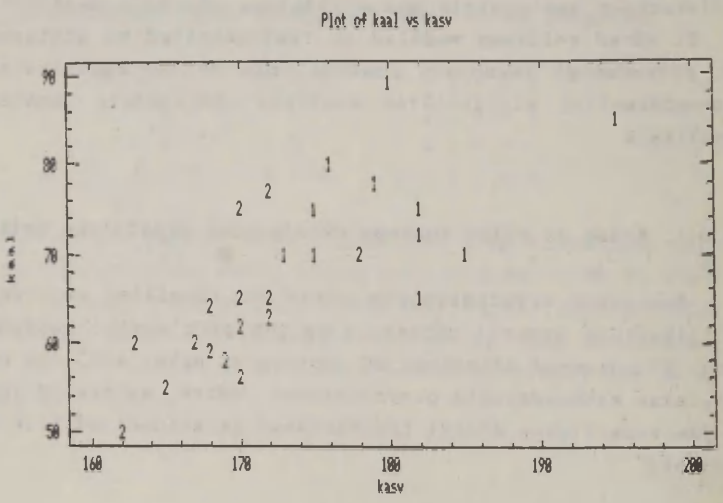
	Log
X: kasv	No
Y: kaal	No

Point labels:

Point codes: sugu
Point colors:
Std. err. X:
Std. err. Y:

Points: Yes
Lines: No

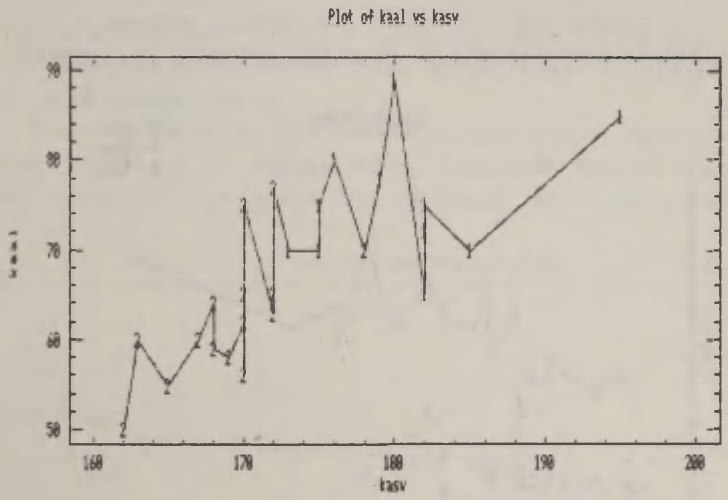
Jn 8.1



Jn 8.2

Joonisel 8.1 ongi esitatud tellimus, kus lahter *Point codes* on täidetud tunnusega **sugu**. Kuivõrd see tunnus on kaheväärtuseline väärtustega 1 (mees) ja 2 (naine), siis tulemis trükisel (vt jn 8.2) on punktid asendatud arvudega *1* ja *2* (samuti võinuks nendele kohtadele olla trükitud tähed *M* ja *N*, kui tunnus **sugu** olnuks vastavalt kodeeritud).

Korrelatsiooniväljal võib naaberpunktid (valimi objektide järjekorranumbrite järjestuses) ka sirglõikudega ühendada. Sellel protseduuril on aga mõtet ainult sel juhul, kui andmestik on järjestatud **X**-tunnuse väärtuste kasvavas järjekorras (seda on võimalik teha protseduuriga *File Operations*, C, *Sort in Descending/Ascending Order*, vt [7], lk 132-135). Selliselt korrastatud andmestiku jaoks ongi joonisel 8.1 esitatud tellimust täiendatud, märkides aknasse *Lines* vastuseks *Yes*, mille tulemusena saadakse joonisel 8.3 kujutatud graafik.



Jn 8.3

8.1.2. Mitme funktsioontunnusega korrelatsiooniväli ühes teljestikus (*Multiple X-Y Plots*). Protseduur *Multiple X-Y Plots* (teine allmenüüs *E. Plotting Functions*) võimaldab ühte teljestikku, seega sõltuvana samast argumenttunnusest

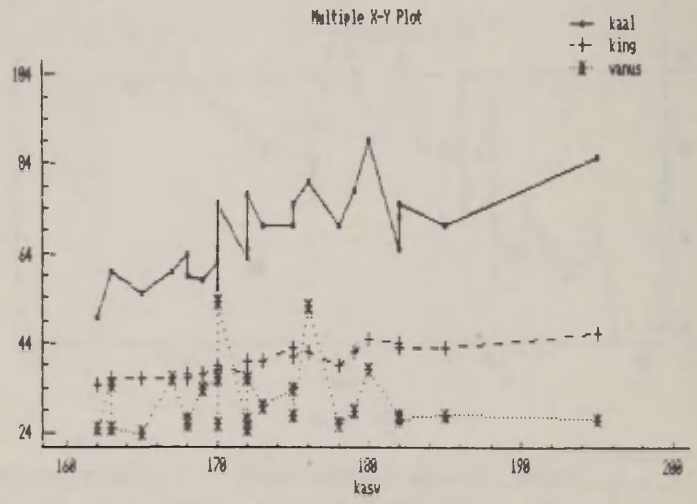
X, paigutada mitme funktsioontunnuse Y korrelatsiooniväljad, kusjuures samale funktsioontunnusele vastavad punktid on joontega ühendatud, nagu see on näha ka joonisel 8.3. Protseduuri tellimuspaneeli kujutab joonis 8.4.

Multiple X-Y Plots

	Axis	Log
X:kasv		No
Y:kaal	Left	No
Y:king	Left	
Y:vanus	Left	
Y:	Left	
Y:	Left	
Y:	Left	
Y:	Left	
Y:	Left	
Y:	Left	
Y:	Left	
Y:	Left	
Y:	Left	

Jn 8.4

Tellimuspaneelil on näha, et niihästi argument- kui ka funktsioontunnuse puhul on võimalik kasutada ka logaritmi-



Jn 8.5

list teisendust, samuti on võimalik mõne argumendi puhul y-telg tõsta joonise paremasse serva.

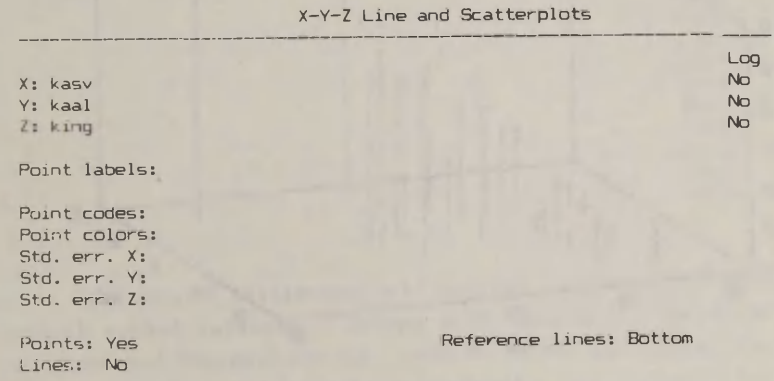
Eriti sobiv on selline esitus ajas kulgevate protsessi-de võrdlemiseks; sel juhul oleks argumenttunnuseks aeg, uu-ritavaid protsesse aga kirjeldaksid funktsioontunnused.

Käesolevas näites esitame me kolm funktsioontunnust sõl-tuvana argumenttunnusest **kasv**, kusjuures andmestik on jär-jestatud argumenttunnuse väärtuste põhjal (see nõue on tar-vilik hästi interpreteeritava joonise saamiseks), vt jn 8.5.

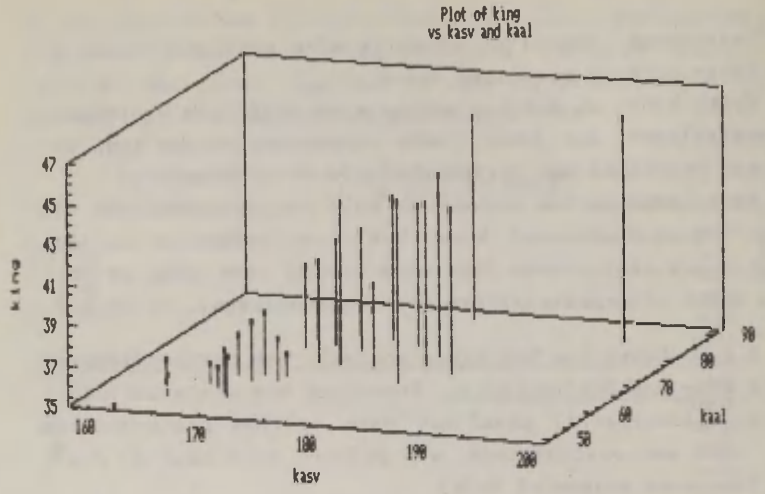
8.1.3. Ruumiline korrelatsiooniväli ehk hajuvusdiagramm (*X-Y-Z Line and Scatterplots*). Ruumilise korrelatsiooniväl-ja (korrelatsiooniparve) graafikut saab tellida protseduuriga *X-Y-Z Line and Scatterplots*, mis paikneb allmenüüs *E. Plot-ting Functions* kolmandal kohal.

Protseduuri tellimuspaneel on üsna sarnane eelmistega, vt jn 8.6, tulemusena saadud graafikut kirjeldab joonis 8.7. Näeme, et siin kujutavad kolmanda tunnuse (**Z**) väärtusi püst-lõigud ("nööpnõelad"), mis on "torgatud" vajalikesse kohta-desse xy-korrelatsiooniväljal ning mille "pead" kujutavadki uuritavaid punkte ruumis.

Samale graafikule võib lisada informatsiooni veel nel-jandagi (diskreetse) tunnuse kohta, kasutades selleks akent *Point codes*. Trükime sinna tunnuse **sugu**, selle tulemusena

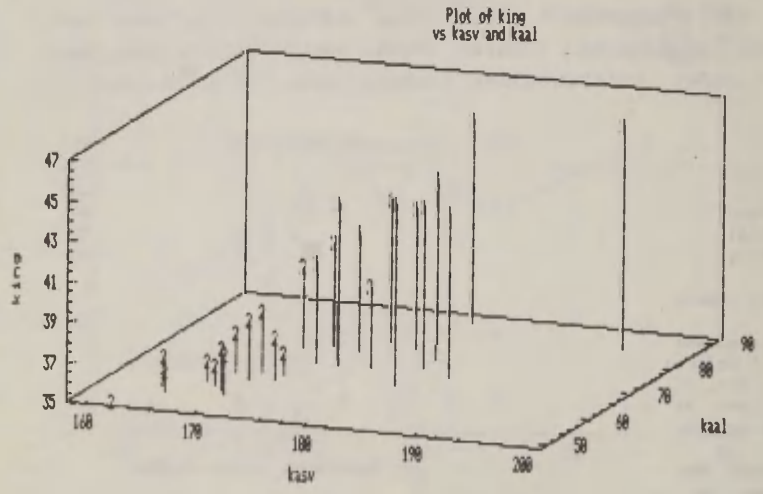


Jn 8.6



Jn 8.7

saame joonisel 8.8 esitatud ruumilise graafiku, kus "nööpnõelapeade" asemel on nööpnõelte tippudes vastava tunnuse koodid, antud juhul arvud *1* (joonisel halvasti nähtavad) ja *2*.



Jn 8.8

Olgu märgitud, et sellistele "korrelatsiooniparvedele" võib lisada ka täiendavat informatsiooni: ühendada punktid joontega (*Lines: Yes*), kanda peale tunnuste standardvead (selleks peavad need aga olema varem iseseisva tunnusena salvestatud), mis kujutataks graafikul lõikude, ristide või ruumiliste ristidena "nööpnõelapeade" asemel jne. Üldjuhul siiski halvendab selline täiendav teave niigi koormatud graafikute vaadeldavust.

8.1.4. Mitme tunnuse ruumiline korrelatsiooniväli (*Multiple X-Y-Z Plots*). Põhimõtteliselt võib ka ruumis kirjeldada mitut tunnust (Z_1, Z_2, ...) sõltuvana samadest argumentidest X ja Y (kujutleksime, et "nööpnõeltel" on erineval kõrgusel mitu pead). Sellise graafiku joonistabki protseduur *Multiple X-Y-Z Plots*, mis paikneb neljandal kohal allmenüüs *E. Plotting Functions*.

Selles protseduuris on protseduurisiseseid kujundamisvõimalusi suhteliselt vähe. Punkte ühendavad jooned (vastavalt järjekorranumbritele andmestikus) on obligatoorsed.

Multiple X-Y-Z Plots

	Axis	Log
X:kasv		No
Y:kaal		No
Z:king	Left	No
Z:varus	Left	
Z:	Left	
Z:	Left	
Z:	Left	
Z:	Left	
Z:	Left	
Z:	Left	
Z:	Left	
Z:	Left	
Z:	Left	
Z:	Left	

Jn 8.9

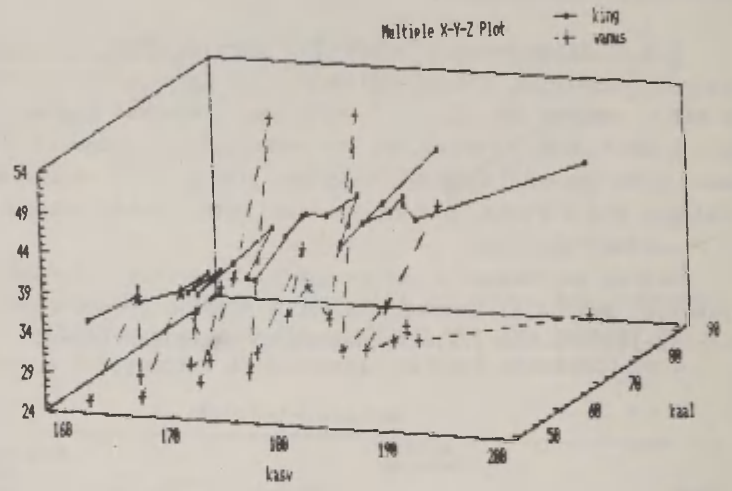
Protseduuri tellimuspaneeli kujutab joonis 8.9 ja selle põhjal saadud tulemust - joonis 8.10. Kuivõrd "nööpnõelad" on asendatud ühendusjoontega, on pilt suhteliselt vähem ülevaatlik joonisel 8.8 esitatuga võrreldes. Hea tahtmise korral õnnestub siiski jooniselt välja lugeda:

tunnuste **X** ja **Y** arvestatav omavaheline positiivne sõltuvus,

tunnuse Z_1 positiivne sõltuvus tunnustest **X** ja **Y**,

tunnuse Z_2 suhteliselt vähene sõltuvus tunnustest **X** ja **Y**,

mõningate tunnuse Z_2 eriti suurte väärtustega objektide olemasolu andmestikus.



Jn 8.10

§ 8.2. Mitme tunnuse ühisjaotuse näitlikustamine marginaalsete ja tinglike korrelatsiooniväljade kaudu

Kui paragrahvis 8.1 esitatud meetodite eesmärgiks oli mitme tunnuse ühisjaotuse esitamine ühel graafikul, siis teine võimalik lähenemine püstitatud ülesandele seisneb ühisjaotuse esitamises kahemõõtmeliste marginaal- või tinglike jaotuste korrelatsiooniväljade komplekti kaudu. Selliseid protseduure on SG-süsteemis kaks, ning nendega me järgnevas tutvumegi.

8.2.1. Mitme tunnuse ühisjaotuse esitamine marginaalsete korrelatsiooniväljade kaudu (*Draftsman Plot*). Protseduur *Draftsman Plot* asub allmenüüs *Q. Multivariate Methods* 11. kohal.

Protseduuri tellimisel tuleb sisestada uuritavate tunnuste (need peavad olema arvtunnused) nimed, vt jn 8.11.

Draftsman Plot

Data vectors: kasv
kaal
king

Full array: No

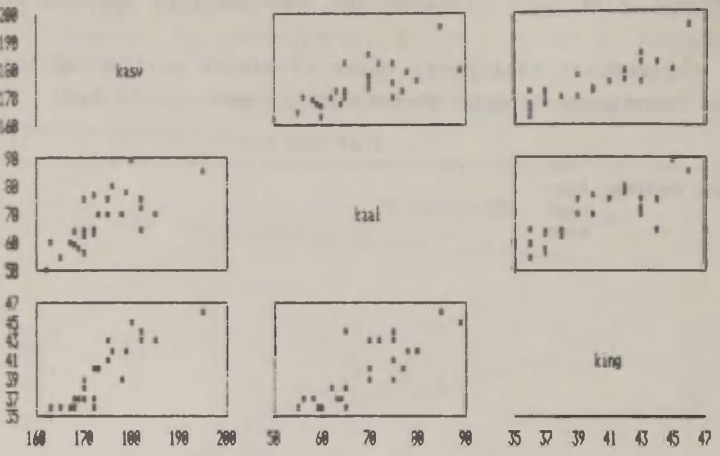
Jn 8.11

Nimede loetelu X_1, X_2, ..., X_m ($m \geq 3$) põhjal moodustatakse $m \times m$ maatriks, mille elemendiks indeksitega (i,j) ($1 \leq i < j \leq m$) on tunnuste X_i ja X_j korrelatsiooniväli (vertikaalteljel tunnus X_i, horisontaalteljel aga X_j).

Parameetri *Full array: Yes* puhul trükitakse väljundina terve maatriks, kusjuures diagonaalielementide asemele on trükitud vastavate tunnuste nimed, vt jn 8.12. Tunnuste väärtuste skaalad on trükitud joonise vasakpoolsesse ja alumisse serva, ning kuna vastav tunnus on ühine maatriksi rea/veeru jaoks, on see skaala kasutatav kõigi selles reas/veerus paiknevate korrelatsiooniväljade jaoks, seega on skaalad olemas kõigi maatriksis esitatud korrelatsiooniväljade jaoks.

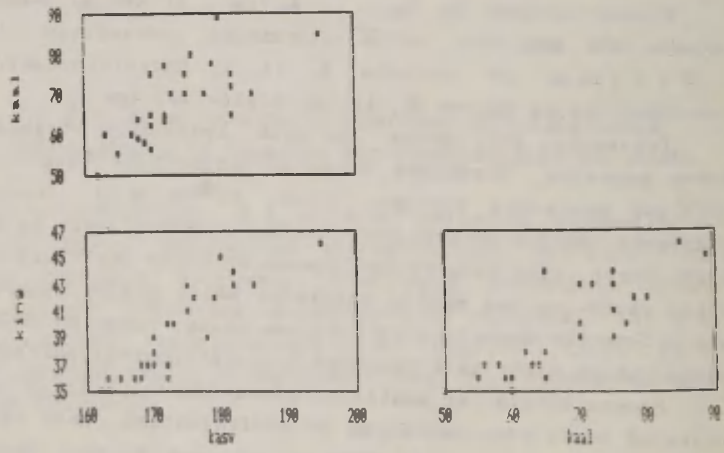
Paneme tähele, et maatriksi elementidele (i,j) ja (j,i) vastavad korrelatsiooniväljad on ekvivalentsed, vaid teljed on vahetatud. Mõnesugune näiline erinevus tuleneb ühikute erinevast pikkusest horisontaal- ja vertikaalteljel (millest

on võimalik protseduuri *Graphics Options* abil ka vabaneda, vt [7], p 2.8.2).



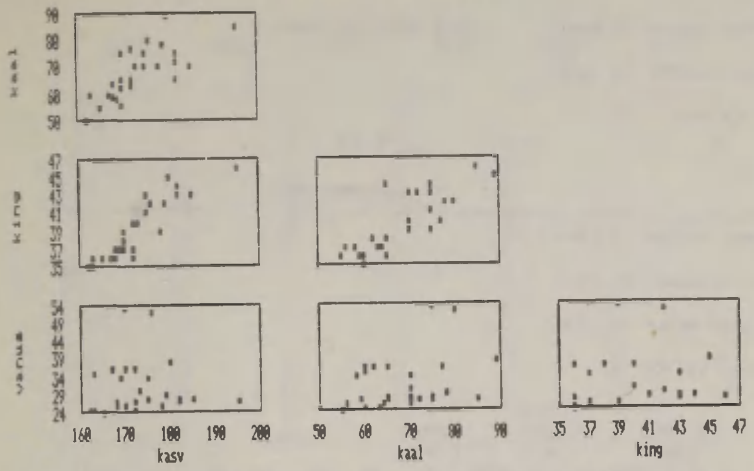
Jn 8.12

Kui parameetrile *Full array* omistatakse väärtus *No*, siis sisaldab trükis ainult maatriksi alumist kolmnurka (ilma peadiagonaalita), vt jn 8.13. Vaikimisi on nimetatud parameetri väärtuseks *No*.



Jn 8.13

Joonisel 8.14 on kujutatud marginaalsed korrelatsiooniväljad nelja tunnuse - **kasv**, **kaal**, **king**, **vanus** - jaoks.



Jn 8.14

8.2.2. Mitme tunnuse ühisjaotuse esitamine tinglike korrelatsiooniväljade kaudu (*Casement Plot*). Kui marginaalsete korrelatsiooniväljade kasutamist võib käsitleda mitmemõõtmelise jaotuse projekteerimisena erinevate tunnuspaaridega määratud tasanditele, siis teine võimalus mitmemõõtmelise jaotuse esitamiseks on selle tükeldamine mingi kolmanda (neljanda, ...) tunnuse väärtuste järgi. See tükeldamine tähendabki teatud tingimust täitvate objektide väljavalimist, ning vastavad korrelatsiooniväljad ongi siis tinglikud.

Tinglike korrelatsiooniväljade konstrueerimiseks on SG-süsteemis protseduur *Casement Plot*, mis paikneb allmenüüs *Q. Multivariate Methods* viimasel, 12. kohal.

Selle protseduuri tellimuspaneeli kujutab joonis 8.15, tellimuses tuleb märkida 3 või 4 tunnuse nimed.

Järgmisel sammul täieneb tellimuspaneel joonisel 8.16 kujutatud klassifitseerimispaneeliga, kus märgitakse kolmanda tunnuse väärtuste järgi moodustatud "kihid", milleks punktihulk jaotatakse.

Casement Plot

Data vector 1: kasv

Data vector 2: kaal

Data vector 3: king

Data vector 4:

Jn 8.15

Casement Plot

Data vector 1: kasv

Data vector 2: kaal

Data vector 3: king

Data vector 4:

Strata Number	king Lower Limit	Upper Limit
1	35	37
2	37	39
3	39	41
4	41	43
5	43	45
6	45	47

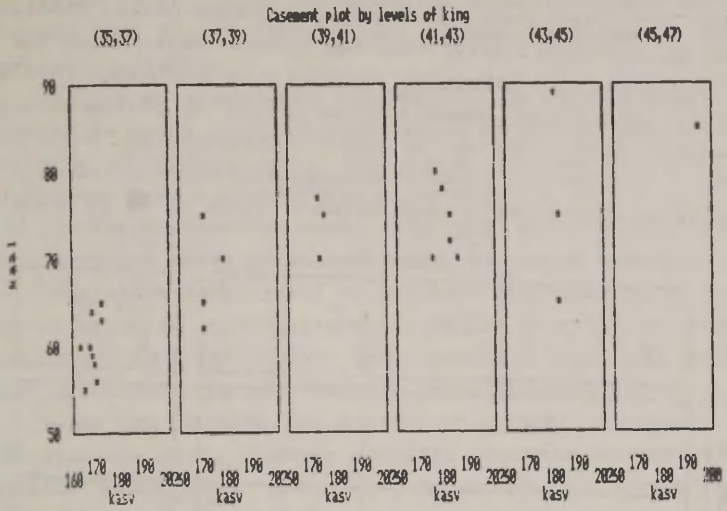
Jn 8.16

Klassifikatsiooni algvariant määratakse vaikimisi; sellesse tuleb suhtuda teatava kriitikaga, sest protseduur määrab klassid nõnda, et alumisel piiril paiknevad objektid jäävad klassist välja. Seetõttu olukorras, kus esimese klassi alumine piir langeb ühte vastava tunnuse miinimumväärtusega, läheb osa materjali kaotsi.

Nii on ka joonisel 8.16 esitatud tellimuse puhul: klassidesse kuuluvad vastavalt objektid kinganumbriga 36-37, 38-39 jne; kuid kaks objekti kinganumbriga 35 jäävad lihtsalt välja.

Protseduuri tulemust kujutab joonis 8.17, millel on kujutatud **kasv**u-**kaal**u korrelatsiooniväljad vastavalt neile objektidele, kelle kinganumber on 36-37, neile, kelle kinganumber on 38-39 jne. Graafikud on samas mastaabis, teljed on

samad. Joonisel on ilmne kasvu ja kaalu märgatav suurenemine sõltuvalt kinganumbrist.



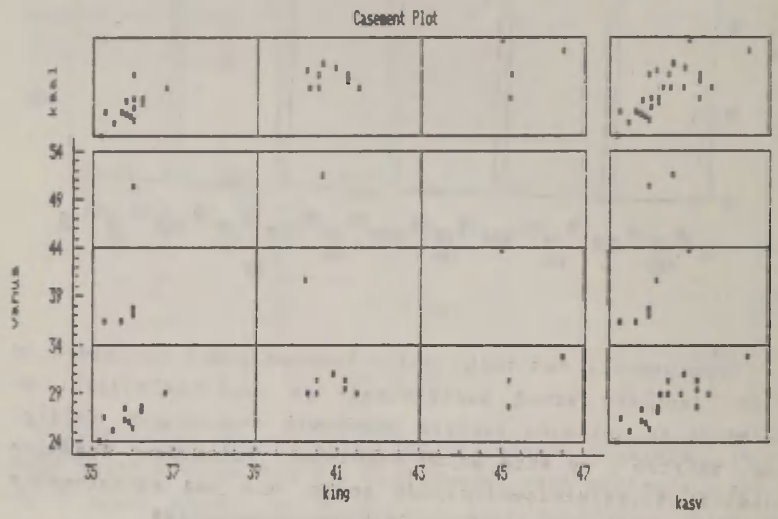
Jn 8.17

Protseduuri tellimust nelja tunnuse puhul kirjeldab jn 8.18; jaotiste arvud horisontaal- ja vertikaalteljel, st kolmanda ja neljanda tunnuse väärtuste skaalal, on valitavad, väärtus 3 on ette antud vaikimisi. Tulemusena saadakse tinglike korrelatsiooniväljade kogum, mis oma struktuurilt meenutab sagedustabelit koos marginaalsagedustega.

```
                        Casement Plot
-------------------------------------------------------------

Data vector 1: kasv

Data vector 2: kaal

Data vector 3: king

Data vector 4: vanus

Number of horizontal divisions:     3

Number of vertical divisions:       3
```

Jn 8.18

Põhitabel (vt jn 8.19) kujutab endast andmestikku, mis on jaotatud osadeks vastavalt tellimuses määratud osade arvudele; käesoleval juhul on tegemist 9 osaga (3×3). Joonise allservas ja vasakul äärel esitatud skaala iseloomustab jaotavaid tunnuseid - kolmandat (allserv) ja neljandat (vasak serv) -, ning sellel skaalal on märgitud ka jaotuspunktid (esimesse veergu jäävad inimesed kinganumbriga kuni 39 (incl.), neid on 15, teise 9 inimest kingadega nr 40-43 ja kolmandasse neli inimest, kes kannavad kingi nr 44 ja enam).



Jn 8.19

Samal viisil on objektid jaotatud vanuse järgi kolmeks: kuni 34-aastased, 35-44 aastased ja üle 44-aastased.

Tabeli vasakpoolne alumine väike lahter näiteks annab kasvu-kaalu korrelatsioonivälja nende objektide jaoks, kes on kuni 34-aastased ja kannavad kingi kuni nr 39.

Kahjuks pole sellele korrelatsiooniväljale tunnuste **kasv** ja **kaal** väärtuste skaalat kantud, küll aga on näha, et **kasv** on horisontaalsel ja **kaal** vertikaalsel teljel.

Samal viisil on ka teistes lahtrites objektide alamhulkade korrelatsiooniväljad. Kogu tabeli kõigi objektide summa

on n = 28, ning pole haruldane seegi, kui mõni lahter (praegu ülemine parempoolne) jääb tühjaks.

Väike tabel joonise parempoolses ülemises nurgas esitab tunnuste **kasv** (horisontaalteljel) ja **kaal** (vertikaalteljel) korrelatsioonivälja.

Joonise ülemises reas (ülejäänud kolm väikest tabelit) on **kaalu** ja **kasvu** korrelatsiooniväljad sõltuvalt **kinganumbri** jaotusest (põhimõtteliselt sama, mis joonisel 8.17, kuid **kinganumbri** jaotus klassidesse on erinev).

Joonise parempoolne veerg (ülejäänud kolm väikest tabelit) kirjeldab **kasvu ja kaalu** jaotust sõltuvalt **vanusest**.

Seega käesoleval juhul on joonisel kujutatud nelja tunnuse ühisjaotust iseloomustavate tinglike korrelatsiooniväljade kõrval ka kolme tunnust iseloomustavad tinglikud korrelatsiooniväljad ning kahe tunnuse korrelatsiooniväli.

Tuleb aga märkida, et siiski on sellegi protseduuriga kõikvõimalike nelja tunnuse tinglike jaotuste hulgast kirjeldatud suhteliselt väike osa: täielikuma informatsiooni saamiseks tuleks 4 uuritava tunnuse seast välja valida kõik võimalused tingimust määrava 2 tunnuse valikuks (neid võimalusi on 6) ning sel viisil saada 6 erinevat tinglike korrelatsiooniväljade peret sarnaselt joonisel 8.19 esitatuga.

§ 8.3. Kolme- ja enamamõõtmeline jaotustabel ning selle analüüsimine

8.3.1. Mitme diskreetse tunnuse ühisjaotuse moodustamine *Crosstabulation*. Varem kirjeldatud (p 7.2.1.) protseduur *Crosstabulation* võimaldab moodustada ka rohkem kui kahe tunnuse ühisjaotust. Selleks tuleb tellimuspaneelile (vt jn 7.2 ja 8.20) märkida nende (diskreetsete) tunnuste nimed. Käesolevas näites vaatleme kolme tunnuse ühisjaotust, lisades tunnustele **synnipaik** ja **eriala** veel tunnuse **sugu**.

Kuna tunnus **sugu** on esitatud kaheväärtuselise arvtunnusena, kodeerime tema väärtused selguse mõttes sümboliteks. Seda oleks võimalik teha koodide tunnuse abil, mille pikkuseks on tunnuse erinevate väärtuste arv ning väärtusteks

Crosstabulation

```
Data vector A: synnipaik
Labels:
Data vector B: eriala
Labels:
Data vector C: sugu
Labels: 2 2 RESHAPE'M N '
Data vector D:
Labels:                                       +----------------------+
Data vector E:                                |                      |
Labels:                                       | Display table        |
Data vector F:                                | Display statistics   |
Labels:                                       | Save counts          |
Data vector G:                                | Save labels          |
Labels:                                       |                      |
Data vector H:                                +----------------------+
Labels:
Data vector I:
Labels:

Percentages: Tablewise
```

Jn 8.20

kodeeritava tunnuse väärtuste koodid (sümbolite jadad). Niisugune koodide tunnus peab olema **SG**-maatriksmuutuja. Et meil sobivat koodide tunnust varem moodustatud ei ole, moodustame selle paneeli täitmise käigus operatsiooni **RESHAPE** abil, vt [7], p 4.9.11).

Protseduuri täitmisel moodustatakse 1. ja 2. tunnuse tinglikud ühisjaotused vastavalt 3. tunnuse väärtustele.

Crosstabulation of synnipaik by eria sugu = M

eriala synnipaik	fil	mat	med	Row Total
alev	0 .0	1 8.3	2 16.7	3 25.0
linn	2 16.7	1 8.3	0 .0	3 25.0
maa	2 16.7	0 .0	0 .0	2 16.7
suurlinn	2 16.7	0 .0	2 16.7	4 33.3
Column Total	6 50.0	2 16.7	4 33.3	12 100.0

Jn 8.21

Seega joonisel 8.21 on esitatud tabel meeste sünnipaikade ja erialade ühisjaotuse kohta.

Tabelites sisalduv informatsioon ja ka tabelite üldine vormistus langeb kokku punktis 7.2.1 kirjeldatuga.

Järgmine tabel (mis vastab järgmisele tinglikule jaotusele) on saadav klahvi **N** (*Next*) vajutamisel. Käesoleval juhul on see naissoost vastajate sünnipaiga-eriala jaotust kirjeldav tabel joonisel 8.22.

Crosstabulation of synnipaik by eria sugu = N

eriala synnipaik	fil	mat	med	Row Total
alev	1	0	1	2
	6.3	.0	6.3	12.5
linn	0	4	1	5
	.0	25.0	6.3	31.3
maa	0	2	1	3
	.0	12.5	6.3	18.8
suurlinn	3	0	3	6
	18.8	.0	18.8	37.5
Column Total	4 25.0	6 37.5	6 37.5	16 100.0

Jn 8.22

Seosekordajad, mis käesoleval juhul arvutatakse, iseloomustavad aga ainult kahe esimese tunnuse omavahelist sõltuvust, seega saaksime ka käesoleval juhul joonistel 7.7 ja 7.8 esitatud väljundi.

Tabelite ja vastavate väärtuste/koodide loetelu on ka käesoleval juhul võimalik salvestada akna (vt jn 8.20) kolmanda ja neljanda rea valimise tulemusena.

8.3.2. Loglineaarsed mudelid mitme tunnuse puhul (*Log-Linear Analysis*). Loglineaarse analüüsi meetod on rakendatav üldiselt m diskreetse tunnuse puhul, kusjuures kõige detailsemas mudelis (mis põhimõtteliselt sarnaneb valemiga (7.18) esitatule) on $\ln n_{i_1 i_2 \ldots i_m}$ avaldises järgmised liikmed:

- vabaliige ν;

- m esimest järku liiget α_i^1, α_i^2, ..., α_i^m, mis vastavad m tunnusele; nende hindamiseks vajalike parameetrite arv on $k_1 + k_2 + \ldots + k_m$, kus k_i on i-nda tunnuse väärtuste arv;

- $[m(m-1)]/2$ teist järku liiget $\gamma_{i_1 i_2}^{j_1 j_2}$, mis vastavad kõikvõimalikele tunnuspaaridele X_{j_1}, X_{j_2} ($j_1, j_2 = 1, \ldots, m$, $j_1 \neq j_2$) ning mille hindamiseks vajalike parameetrite arv on $\Sigma k_{j_1} k_{j_2}$;

- ...;

- üks m-järku liige.

Liikmete arvu vähendamine on seotud teatavasti kitsendavate eelduste tegemisega mudeli kohta; nendest olulisemaid (teatavad jaotused on ühtlased, teatavad tunnused sõltumatud) me juba vaatlesime punktis 7.6.2. Samalaadsed on ka ülejäänud kitsendused, mis enamasti väidavad teatavate tunnusrühmade (vastastikust) sõltumatust.

Üldjuhul koosneb loglineaarne analüüs järgmistest sammudest:

– valitakse teatav mudel (fikseerides oletatavad mõjuvad tunnused *Generators* ja nende mõju järgu *Effects*);

– hinnatakse mudeli parameetreid;

– kontrollitakse andmestiku sobivust vastava mudeliga.

Seda protseduuri on põhimõtteliselt võimalik korrata kuni mingis mõttes sobivaima mudeli (mudelite hulga) leidmiseni. Mudelite otsimise protsessi on võimalik läbi viia ka vastava programmi abil, kusjuures SG-süsteemis on realiseeritud kaks otsimise strateegiat, nimelt

- mudelist liikmeid järk-järgult kõrvale jättes (*Backward select*),

- mudelisse liikmeid samm-sammult lisades (*Forward select*).

Vaatleme järgnevalt loglineaarse analüüsi protseduuri konkreetse kolmemõõtmelise sagedustabeli põhjal.

Analüüsi aluseks on kolme tunnuse
eriala (väärtused fil, mat, med),
sugu (väärtused m, n),
abielus (väärtused ei, ja)
ühisjaotus, mis on moodustatud protseduuriga *Crosstabulation*

ja salvestatud maatriksisse **TABLEpluss** (vt jn 8.23). Tellimus on esitatud joonisel 8.24.

```
: TABLEpluss
 1 5
 1 3

 1 1
 1 5

 2 2
 3 3
:
```

Jn 8.23

```
                         Log-Linear Analysis
------------------------------------------------------------------------
Frequency table: TABLEpluss

Label matrix: LEGENDSplu

Generators: A B C


                                                 +-----------------+
                                                 :                 :
Effects: A B C                                   : Fit model       :
                                                 : Plot residuals  :
                                                 : Test order      :
                                                 : Specify order   :
                                                 : Backward select :
                                                 : Forward select  :
Convergence criterion: 0.01                      :                 :
                                                 +-----------------+
Maximum number of iterations: 20
```

Jn 8.24

Esimest järku mudelile vastav tabel on trükitud joonisel 8.25 (tabeli kirjeldus vt p 7.6.2), võrdlustulemuste (hüpoteesi H_0: jaotus vastab mudelile) kontrollimise tulemus (vt jn 8.26) kinnitab, et andmestik on küllaltki hästi kooskõlas nullhüpoteesile vastava oletusega, mille kohaselt sagedused n_{ijl} vastavad mudelile

$$\ln n_{ijl} = \nu + \alpha_i + \beta_j + \delta_l,$$

kus α_i on I tunnuse, β_j II tunnuse ja δ_l III tunnuse esimest järku mõjud.

Comparison of observed and expected frequencies

Generator: A B C

Factor	Level	Observed Freq.	Expected Freq.	Residual	Std. Resid.
Factor A: fil					
Factor B: m					
Factor C: ei		1 (3.57)	1.4 (4.92)	-0.4	-0.322
Factor C: ja		5 (17.86)	2.9 (10.39)	2.1	1.227
Factor B: n					
Factor C: ei		1 (3.57)	1.8 (6.56)	-0.8	-0.617
Factor C: ja		3 (10.71)	3.9 (13.85)	-0.9	-0.446
Factor A: mat					
Factor B: m					
Factor C: ei		1 (3.57)	1.1 (3.94)	-0.1	-0.097
Factor C: ja		1 (3.57)	2.3 (8.31)	-1.3	-0.870
Factor B: n					
Factor C: ei		1 (3.57)	1.5 (5.25)	-0.5	-0.387
Factor C: ja		5 (17.86)	3.1 (11.08)	1.9	1.078
Factor A: med					
Factor B: m					
Factor C: ei		2 (7.14)	1.4 (4.92)	0.6	0.530
Factor C: ja		2 (7.14)	2.9 (10.39)	-0.9	-0.533
Factor B: n					
Factor C: ei		3 (10.71)	1.8 (6.56)	1.2	0.858
Factor C: ja		3 (10.71)	3.9 (13.85)	-0.9	-0.446

Jn 8.25

Summary of fitted model

Maximum Cell Difference: 8.88178E-16 Number of iterations: 2

Goodness-of-fit test statistics for model adequacy

	Statistic	d.f.	P-value
Likelihood ratio chi-square	5.54207	7	.5941
Pearson chi-square	5.76511	7	.5674

Jn 8.26

Joonisel 8.27 on hinnatud mudeli kõigi parameetrite ν, α_i ($i = 1, \ldots, 3$), β_j ($j = 1, 2$) ja δ_l ($1 = 1, 2$) väärtusi, samuti on arvutatud nende standardhälbed ja (kasutades normaalset lähendit) usaldusintervallid.

Tähelepanu väärib siin tõsiasi, et peale esimese parameetri ν sisaldavad kõigi ülejäänud parameetrite usalduspiirid nullpunkti, seega pole need statistiliselt olulised (nullist erinevad).

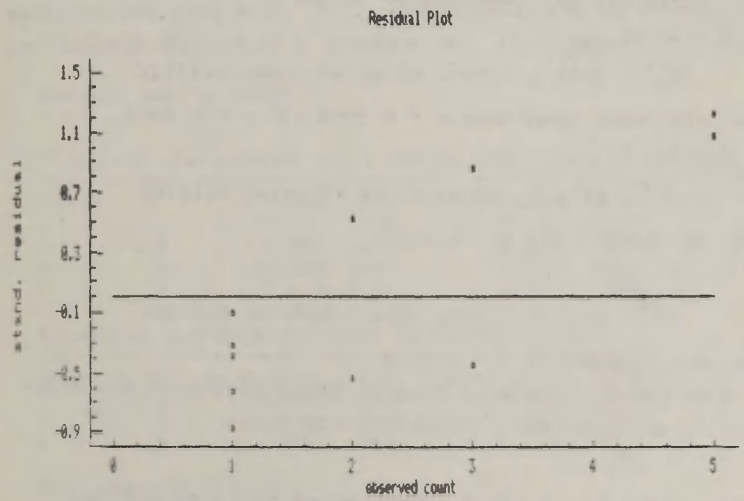
Estimated model coefficients

Generator: A B C

Factor	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
	0.76337	0.22669	3.36746	0.31906	1.20769
A					
1	0.07438	0.32275	0.23046	-0.55821	0.70697
2	-0.14876	0.34359	-0.43296	-0.82220	0.52468
3	0.07438	0.29345	0.25347	-0.50077	0.64954
B					
1	-0.14384	0.22669	-0.63452	-0.58816	0.30047
2	0.14384	0.22669	0.63452	-0.30047	0.58816
C					
1	-0.37361	0.22669	-1.64809	-0.81792	0.07071
2	0.37361	0.22669	1.64809	-0.07071	0.81792

Jn 8.27

Joonisel 8.28 on esitatud mudeli prognoosijääkide graafik (saadud aknast valikuga *Plot residuals*), kus x-teljele on kantud tabelis paiknevad sagedused, y-teljele prognoosivead. Siin ilmneb (täiesti seaduspäraselt), et võrreldes tunnuste sõltumatuse eeldusele rajatud mudeliga on suured



Jn 8.28

tabelisagedused (näiteks 5) liiga suured - jääk on positiivne, väikesed (näiteks 1) aga - liiga väikesed, nende puhul on jääk negatiivne.

8.3.3. Mudeli järgu valik. Sobiva mudeli järgu valimisel aitab valiku *Test order* (kolmas rida aknas) tulemusena saadud koondtabel. vt jn 8.29. Seda tabelit on sobiv hakata uurima alt üles.

Probabilities that K-way and higher effects are zero

K	DF	L. R. Chisq	Prob	Pearson Chisq	Prob	Iteration
3	2	.8496	.6539	.8884	.6413	4
2	7	5.5421	.5941	5.7651	.5674	2
1	11	10.0603	.5250	10.5714	.4798	2

Probabilities that all K-way effects are zero

K	DF	L. R. Chisq	Prob	Pearson Chisq	Prob
3	2	.8496	.6539	.8884	.6413
2	5	4.6925	.4545	4.8767	.4311
1	4	4.5182	.3404	4.8063	.3078

Jn 8.29

Tabeli alumisest blokist on näha olulisuse tõenäosused (*Prob*) järgmiste nullhüpoteeside korral:

K = 1, st

$H_o^{(1)}$: kõik 1. järku mõjud võrduvad nulliga

korral olulisuse tõenäosus p = 0,3078 või p = 0,3404;

K = 2, st

$H_o^{(2)}$: kõik 2. järku mõjud võrduvad nulliga

korral p = 0,4545 või p = 0,4311;

K = 3, st

$H_o^{(3)}$: kõik 3. järku mõjud võrduvad nulliga

korral p = 0,6539 või p = 0,6413.

Sama tabeli ülemisest blokist saame olulisuse tõenäosused ka veidi karmimate nullhüpoteeside jaoks:

K = 1, st

$H_o^{(1)}$: kõik 1., 2. ja 3. järku mõjud võrduvad nulliga

korral p = 0,4798 või p = 0,5250;

K = 2, st

$H_o^{(2)}$: kõik 2. ja 3. järku mõjud võrduvad nulliga korral p = 0,5941 või p = 0,5674;

K = 3, st

$H_o^{(3)}$: kõik 3. järku mõjud võrduvad nulliga korral p = 0,6539 või p = 0,6413.

Erinevad olulisuse tõenäosused tulenevad kahe erineva asümptootiliselt χ^2-jaotusega statistiku kasutamisest, vt p 7.6.1. Kuivõrd praegusel juhul järeldused on kooskõlas, pole nende eristamine probleemiks.

Antud materjali puhul kehtivad kõik nullhüpoteesid, st mudelis ei ole ei 1., 2. ega 3. järku mõjud statistiliselt olulised.

8.3.4. Sobivaima mudeli valik üksiktunnuste tasemel. Kuigi kõigi tunnuste esimest järku mõjud üheskoos ei ole statistiliselt olulised, on siiski mõtet vaadelda igaühe mõju eraldi, kasutades operatsiooni *Backward select* (aknas eelviimane) ning alustades kõigi tunnuste (*Generators*) I järku mõjudest. Vastav esimesel sammul saadav väljund on esitatud joonisel 8.30. Selle alumises blokis on I järku mõjud jagatud vastavalt I tunnuse (*A*), II tunnuse (*B*) ja III

Generator set for model:
A B C

Probability that previous model less excluded effect is adequate

Effect	DF	L. R. Chisq	Prob	Pearson Chisq	Prob	Iteration
	7	5.5421	.5941	5.7651	.5674	2
C	8	9.1936	.3262	10.2083	.2507	2
B	8	6.1155	.6343	6.3205	.6114	2
A	9	5.8354	.7563	6.0175	.7382	2

Probability that excluded effect has coefficient zero

Effect	DF	L. R. Chisq	Prob	Pearson Chisq	Prob
C	1	3.6515	.0560	4.4432	.0350
B	1	.5734	.4489	.5554	.4561
A	2	.2933	.8636	.2524	.8814

Eliminate effect? (Y/N):

Jn 8.30

tunnuse (*C*) mõjudeks; vastavalt lahutuvad komponentideks ka χ^2-statistik (joonisel 8.29 kõige alumine rida) ning vabadusastmete arvud. Siit ilmneb tunnuste erinev mõju. Kui tellija soovib ühe tunnuse mõju kõrvale jätta (*Eliminate effect - Yes*), siis jäetakse vastavalt standardsele kokkuleppele kõrvale nõrgima mõjuga tunnus *A* ning järgmisel sammul (vt jn 8.31) toimub valik kahe mõju - *B* ja *C* - vahel. Viimasel sammul (vt jn 8.32) on käes lõplik tulemus - tunnusele *C* (abielus - ei, jah) vastav parameeter üksi võetuna osutub statistiliselt oluliseks (kui kasutada Pearsoni χ^2-kriteeriumi).

```
Generator set for model:
B C

Probability that previous model less excluded effect is adequate
```

Effect	DF	L. R. Chisq	Prob	Pearson Chisq	Prob	Iteration
	9	5.8754	.7563	6.0175	.7382	2
C	10	9.4869	.4866	10.2500	.4188	2
B	10	6.4088	.7798	6.3860	.7819	2

```
Probability that excluded effect has coefficient zero
```

Effect	DF	L. R. Chisq	Prob	Pearson Chisq	Prob
C	1	3.6515	.0560	4.2325	.0397
B	1	.5734	.4489	.3684	.5439

```
Eliminate effect? (Y/N):
```

Jn 8.31

```
Generator set for model:
C

Probability that previous model less excluded effect is adequate
```

Effect	DF	L. R. Chisq	Prob	Pearson Chisq	Prob	Iteration
	10	6.4088	.7798	6.3860	.7819	2
C	11	10.0603	.5250	10.5714	.4798	2

```
Probability that excluded effect has coefficient zero
```

Effect	DF	L. R. Chisq	Prob	Pearson Chisq	Prob
C	1	3.6515	.0560	4.1855	.0408

```
Eliminate effect? (Y/N):
```

Jn 8.32

Paneme tähele ka joonistel 8.29-8.31 esitatud tabelite ülemisi blokke.

Esimeses neist kontrollitakse kõrgemat järku koosmõjude olulisust, käesolevas näites on need, nagu juba nägime, ebaolulised.

Jooniste 8.31 ja 8.32 ülemistes blokkides kontrollitakse, kas potentsiaalselt ärajäetavate mõjude puudumisel mudel säilitab adekvaatsuse. Käesoleval juhul on kõikjal jõus nullhüpotees, st parandatud mudel ei erine esialgsest oluliselt.

Analoogiline protseduur, kuid üksikmõjude samm-sammult lisamisega, toimub valiku *Forward select* tulemusena.

9. TUNNUSTEVAHELISTE SEOSTE STRUKTUURI ANALÜÜS

Käesolevas peatükis alustame mitmemõõtmelise statistilise analüüsi (MSA) käsitlemist.

MSA korral vaadeldakse tunnusvektorit $\mathbf{X} = (X_1, \ldots, X_m)$, mis üldjuhul on mingi alamvektor lähtetunnusvektorist $\mathbf{Z}$ (vt [7], § 1.2).

Enamikul juhtudest (erandid märgime hiljemas tekstis spetsiaalselt) eeldatakse, et kõik tunnused X_i ($i = 1, \ldots, m$) on arvulised. Paljudel juhtudel (kui tegemist on hüpoteeside kontrollimisega) lisandub veel nõue, et $\mathbf{X}$ on mitmemõõtmelise normaaljaotusega. Parameetrite punkthinnangute puhul aga normaalsuse eeldus ei ole vajalik.

Jämedates joontes võime MSA ülesanded jaotada järgmisse kolme klassi:

- tunnuste vaheliste seoste struktuuri analüüs ja kirjeldamine (korrelatsioon-, komponent-, faktoranalüüs);
- prognoosiülesanne, mille eesmärgiks on funktsioontunnust (-tunnuseid) argumenttunnuste vektori järgi prognoosida (regressioon- ja dispersioonanalüüs);

Objektide struktuuri analüüs ja kirjeldamine (diskriminant- ja klasteranalüüs).

Nimetatud meetodite rühmi ja neile vastavaid SG-protseduure käsitlemegi käesoleva raamatu 9., 10. ja 11. peatükis.

§ 9.1. Korrelatsiooni- ja kovariatsioonimaatriks ning nende analüüsimine

9.1.1. Teoreetiline ja valimkovariatsioonimaatriks. Käesolevas punktis kasutame uuritava tunnusvektori tähisena tähte $\mathbf{X}$ ning mõtleme selle all veeruvektorit

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \dots \\ X_m \end{bmatrix}.$$

Juhusliku vektori kovariatsioonimaatriks Σ on defineeritud järgmise seosega:

$$\Sigma = E(\mathbf{X} - \mu)(\mathbf{X} - \mu)',$$

kus μ on vektori $\mathbf{X}$ keskväärtuste vektor ($\mu = E\mathbf{X}$), E tähistab keskväärtuse leidmise operatsiooni ja ′ maatriksi transponeerimist. Kovariatsioonimaatriksil on m rida ja m veergu. Ta on sümmeetriline maatriks elementidega σ_{ij}, kus

$$\sigma_{ij} = \text{cov}\,(X_i, X_j) = E(X_i - \mu_i)(X_j - \mu_j), \quad i \neq j,$$

ning

$$\sigma_{ii} = \sigma_i^2 = DX_i,$$

seega maatriksi Σ elementideks on tunnuste kovariatsioonid ehk teist järku tsentraalsed segamomendid $\bar{\mu}_{ij}$ ning peadiagonaalil - dispersioonid.

Valimi põhjal arvutatakse valimkovariatsioonimaatriks S elementidega s_{ij},

$$s_{ij} = \frac{1}{n-1} \sum_{l=1}^{n} (x_{il} - \bar{x}_i)(x_{jl} - \bar{x}_j), \quad i \neq j,$$

kusjuures peadiagonaalil on tunnuste valimdispersioonid s_i^2 (vt ka p 6.2.3).

9.1.2. Kovariatsioonimaatriksi leidmine *Covariance Analysis*. Tühikute arvestamine. Kovariatsioonimaatriksi arvutamiseks leidub SG-süsteemis protseduur *Covariance Analysis* (teine allmenüüs *Q. Multivariate Methods*). Olgu märgitud, et protseduuris mingit kovariatsioonimaatriksi (struktuuri) analüüsi ei toimu, selle käigus on võimalik üksnes vajalik kovariatsioonimaatriks leida ja välja trükkida.

Joonisel 9.1 on esitatud kovariatsioonimaatriksi tellimuspaneel. Sellele tuleb trükkida uuritavate tunnuste nimed (kas üksteise alla või ritta, kasutades eraldajat &). Tunnused peavad olema arvulised ja võrdse pikkusega (valimi

mahuga). Teine võimalus on trükkida andmestiku nimi, et siis moodustataks kovariatsioonimaatriks kõigist selle andmestiku tunnustest. See on võimalik vaid sel juhul, kui andmestik koosneb ainult arvtunnustest ning on objekt-tunnus-maatriksi kujuga, st et kõigi tunnuste pikkused on võrdsed.

```
                              Covariance Analysis
------------------------------------------------------------------
Data vectors kaal
or filename: kasv
             king
             vanus
             sugu
             abielus
             pere
             tulu
             teadtoid
             arvutoskus
             eritulu



Missing values: Listwise
```

Jn 9.1

Paneeli alumises servas on parameeter *Missing values*, mis määrab tühikute arvestamise režiimi. Selgitame seda järgmises tabelis esitatava andmestiku varal.

	vanus	**sugu**	**kasv**	**kaal**
Ants	17	1	182	
Peeter	22	1	170	70
Mall		2	162	60
Eva	19	2		55
Rein	23	1		
Mann	24			
Mats	16		185	
Triin	20	2	177	
Anu	18	2	170	53
Tõnu	25	1	178	30

Kui tühikud eraldatakse loeteluna (*Listwise*), siis saab kovariatsioonimaatriksi arvutamisel kasutada ainult kolme objekti - need on Peeter, Anu ja Tõnu, kellel on olemas kõik mõõtmistulemused.

Selle režiimi korral moodustatakse täielik alammassiiv, st selline objektide alamhulk, millel kõik tunnused on täielikult mõõdetud.

Teine võimalus on tühikute eraldamine paarikaupa (*Pairwise*). Sel juhul vanuse ja soo kovariatsiooni arvutamisel arvatakse massiivist välja Mann, Mall ja Mats; vanuse ja kasvu kovariatsiooni arvutamisel Mall, Eva, Rein ja Mann, jne. Kokkuvõttes saadakse kovariatsioonimaatriksisse järgmised osavalimite mahud:

	vanus	sugu	kasv	kaal
vanus	9	7	6	5
sugu	7	8	6	5
kasv	6	6	7	4
kaal	5	5	4	6

Käesolevas näites on iga kovariatsioonikordaja arvutamiseks kasutatud erinevat valimit (kuigi mõnel juhul valimite mahud ühte langevad).

Kuigi tühikute paarikaupa väljajätmine võimaldab valimis sisalduvat informatsiooni maksimaalselt kasutada, võivad asjaolust, et erinevad seosekordajad on erinevate osavalimite põhjal arvutatud, tuleneda mitmesugused täiendavad raskused, mille põhjuseks on tekkinud maatriksi mittekooskõlalisus.

Teatavad raskused tekivad ka tühikute väljajätmisel loetelu alusel. Nimelt, kui samade tunnuste kovariatsioonivõi korrelatsioonikordaja arvutatakse kahe erineva maatriksi koosseisus ning tühikud on välja jäetud loeteluna, saadakse üldiselt erinevad väärtused, sest tavaliselt on erinevad maatriksid arvutatud mõnevõrra erinevate objektihulkade järgi.

Kui puuduvaid väärtusi on vähe või pole neid üldse, võib kasutada mõlemat valikut. Vaikimisi väärtus on *Listwise*.

Joonisel 9.2 on esitatud kovariatsioonimaatriksi fragment (selle 4 esimest veergu). Maatriksi igas lahtris on kaks arvu: ülemine on kovariatsioonimaatriksi element, alumine - valimi maht.

Kovariatsioonimaatriksi võib salvestada, et seda kasutada edasises tegevuses.

Covariances

	kaal	kasv	king	vanus
kaal	88.7765	54.1019	26.4656	25.1548
	(28)	(28)	(28)	(28)
kasv	54.1019	58.8611	23.6296	-4.06481
	(28)	(28)	(28)	(28)
king	26.4656	23.6296	11.4392	2.82011
	(28)	(28)	(28)	(28)
vanus	25.1548	-4.06481	2.82011	54.0780
	(28)	(28)	(28)	(28)
sugu	-3.27513	-3.00000	-1.52381	-0.06878
	(28)	(28)	(28)	(28)
abielus	-0.53042	-0.52778	-0.25397	0.39286
	(28)	(28)	(28)	(28)
pere	0.25397	0.70370	0.08466	0.43386
	(28)	(28)	(28)	(28)
tulu	330.946	448.102	187.460	-153.247
	(28)	(28)	(28)	(28)
teadtoid	59.7897	27.2685	20.5873	54.3479
	(28)	(28)	(28)	(28)
arvutoskus	-1.30291	-0.37963	-0.52381	-1.48545
	(28)	(28)	(28)	(28)
eritulu	18.4877	15.2778	28.3642	3.64198
	(28)	(28)	(28)	(28)

Covariance (sample size)

Jn 9.2

9.1.3. Teoreetiline korrelatsioonimaatriks. Olgu $\mathbf{X}$ tunnusvektor ja Σ tema kovariatsioonimaatriks. Tähistame veel $D = \text{diag}\,\Sigma$, st D on maatriks, mille diagonaalielemendid langevad ühte Σ diagonaalielementidega, ülejäänud elemendid aga võrduvad nulliga.

Maatriksit P (suur roo) elementidega ρ_{ij}, mis on defineeritud valemiga

$$P = D^{-1/2}\Sigma D^{-1/2},$$

nimetatakse korrelatsioonimaatriksiks. Korrelatsioonimaatriksi elemendid on lineaarsed korrelatsioonikordajad (vt p 7.2.4 ja 7.5.2), neile valimi põhjal saadud hinnangud r_{ij} arvutatakse järgmiselt:

$$r_{ij} = \frac{s_{ij}}{s_i \cdot s_j},$$

kus s_i ja s_j on tunnuste X_i ja X_j standardhälbe hinnangud.

Valimi põhjal arvutatud korrelatsioonimaatriksi tähiseks on R; maatriksil R on järgmised omadused:

1. Maatriks R on sümmeetriline, $r_{ij} = r_{ji}$.

2. Maatriksi R peadiagonaalil on ühed, seega maatriksi R jälg (peadiagonaali elementide summa) võrdub tema järguga m:

$$\mathrm{Tr}\ R = m.$$

3. Maatriks R on mittenegatiivselt määratud, $\det R \geq 0$.

4. Järgmised väited on samaväärsed:

- maatriksi R astak on m-r,
- tunnuste hulgas on r (omavahel sõltumatut) lineaarset sõltuvust

$$\sum_{i=1}^{m} \alpha_i^{(j)} X_i = 0, \quad j = 1, \ldots, r,$$

kus iga j korral mingi $\alpha_i^{(j)}$ erineb nullist.

Siit järeldub, et juhul kui r = 0, siis

- maatriks R on regulaarne, eksisteerib R^{-1};
- tunnuste $X_1, \ldots, X_m$ vahel ei ole täpseid lineaarseid sõltuvusi.

9.1.4. Korrelatsioonimaatriksi leidmine (*Correlation Analysis*) ja analüüs. Korrelatsioonimaatriksi leidmise protseduur on esimene allmenüüs *Q. Multivariate Methods*. Selle tellimuspaneel on sarnane kovariatsioonimaatriksi tellimuspaneeliga. Saadud korrelatsioonimaatriksit kujutab joonis 9.3.

Sample Correlations

	kaal	kasv	king	vanus	sugu	abielus
kaal	1.0000	.7484	.8305	.3630	-.6897	-.1184
	(28)	(28)	(28)	(28)	(28)	(28)
	.0000	.0000	.0000	.0576	.0000	.5486
kasv	.7484	1.0000	.9106	-.0720	-.7759	-.1446
	(28)	(28)	(28)	(28)	(28)	(28)
	.0000	.0000	.0000	.7156	.0000	.4627
king	.8305	.9106	1.0000	.1134	-.8940	-.1579
	(28)	(28)	(28)	(28)	(28)	(28)
	.0000	.0000	.0000	.5656	.0000	.4223
vanus	.3630	-.0720	.1134	1.0000	-.0186	.1123
	(28)	(28)	(28)	(28)	(28)	(28)
	.0576	.7156	.5656	.0000	.9253	.5693
sugu	-.6897	-.7759	-.8940	-.0186	1.0000	.0221
	(28)	(28)	(28)	(28)	(28)	(28)
	.0000	.0000	.0000	.9253	.0000	.9112
abielus	-.1184	-.1446	-.1579	.1123	.0221	1.0000
	(28)	(28)	(28)	(28)	(28)	(28)
	.5486	.4627	.4223	.5693	.9112	.0000
pere	.0213	.0723	.0197	.0465	-.0993	.5351
	(28)	(28)	(28)	(28)	(28)	(28)
	.9145	.7146	.9206	.8142	.6150	.0033
tulu	.1988	.3305	.3136	-.1179	-.3280	.2379
	(28)	(28)	(28)	(28)	(28)	(28)
	.3106	.0858	.1041	.5501	.0884	.2228
teadtoid	.4926	.2759	.4725	.5737	-.4458	.1686
	(28)	(28)	(28)	(28)	(28)	(28)
	.0077	.1552	.0111	.0014	.0174	.3910
arvutoskus	-.1164	-.0417	-.1304	-.1700	.1767	.2224
	(28)	(28)	(28)	(28)	(28)	(28)
	.5553	.8333	.5085	.3870	.3683	.2553
eritulu	.0180	.0182	.0768	.0045	-.0437	-.3577
	(28)	(28)	(28)	(28)	(28)	(28)
	.9277	.9266	.6978	.9817	.8251	.0617

Coefficient (sample size) significance level

Joonis

	pere	tulu	teadtoid	arvutoskus	eritulu
kaal	.0213	.1988	.4926	-.1164	.0180
	(28)	(28)	(28)	(28)	(28)
	.9145	.3106	.0077	.5553	.9277
kasv	.0723	.3305	.2759	-.0417	.0182
	(28)	(28)	(28)	(28)	(28)
	.7146	.0858	.1552	.8333	.9266
king	.0197	.3136	.4725	-.1304	.0768
	(28)	(28)	(28)	(28)	(28)
	.9206	.1041	.0111	.5085	.6978
vanus	.0465	-.1179	.5737	-.1700	.0045
	(28)	(28)	(28)	(28)	(28)
	.8142	.5501	.0014	.3870	.9817
sugu	-.0993	-.3280	-.4458	.1767	-.0437
	(28)	(28)	(28)	(28)	(28)
	.6150	.0884	.0174	.3683	.8251
abielus	.5351	.2379	.1686	.2224	-.3577
	(28)	(28)	(28)	(28)	(28)
	.0033	.2228	.3910	.2553	.0617
pere	1.0000	.2433	.1525	.1791	-.8284
	(28)	(28)	(28)	(28)	(28)
	.0000	.2122	.4384	.3618	.0000
tulu	.2433	1.0000	.5510	.4345	.2381
	(28)	(28)	(28)	(28)	(28)
	.2122	.0000	.0024	.0209	.2225
teadtoid	.1525	.5510	1.0000	-.0445	.1503
	(28)	(28)	(28)	(28)	(28)
	.4384	.0024	.0000	.8220	.4451
arvutoskus	.1791	.4345	-.0445	1.0000	.0718
	(28)	(28)	(28)	(28)	(28)
	.3618	.0209	.8220	.0000	.7164
eritulu	-.8284	.2381	.1503	.0718	1.0000
	(28)	(28)	(28)	(28)	(28)
	.0000	.2225	.4451	.7164	.0000

9.3

Korrelatsioonimaatriksi iga lahter sisaldab kolme arvu. Kõige esimene neist on korrelatsioonikordaja, teine - valimi maht ja kolmas - olulisuse tõenäosus p kontrollimaks hüpoteesipaari

H_o: $\rho_{ij} = 0$ (st korrelatsioon on mitteoluline, üldkogumis on vastav korrelatsioon 0),

H_1: $\rho_{ij} \neq 0$ (st korrelatsioon on statistiliselt oluline, tunnuste vahel on korrelatiivne seos ka üldkogumis).

Hüpotees H_1 võetakse vastu siis, kui olulisuse tõenäosus p pole suurem soovitavast olulisuse nivoost α (mille standardne väärtus on 0,05).

Korrelatsioonanalüüsi protseduuri tulemusi on võimalik salvestada, vt akent joonisel 9.4. Kui vaikimisi salvestatakse tulemused tööfaili (*WORKAREA*), siis korrelatsioonimaatriks on kasulik salvestada nimelisse püsifaili, et see säiliks ka järgmisteks tööetappideks.

```
+-------------------------+
|                         |
| Save the correlations    |
| Save the variable labels |
|                         |
+-------------------------+
```

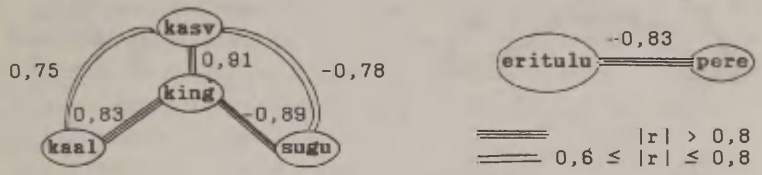
Jn 9.4

Korrelatsioonimaatriksi analüüsimise hõlbustamiseks on sobiv kasutada mitmesuguseid korrelatsioonigraafe, mille abil saab illustreerida tunnuste rühmitumist omavahel tihedalt seotud tunnusrühmadesse, mis ongi tunnustevaheliste seoste struktuuri aluseks.

9.1.5. Korrelatsioonigraafid. Esitame näitena nn tasemegraafi; selle puhul ühendame kaarega need tunnused (tipud), mille vaheline seos ületab teatavat taset. Kasutades erinevaid tasemeid, saame seeria graafe, mida saab ühisel joonisel kujutada, kasutades erinevaid ühendusjooni.

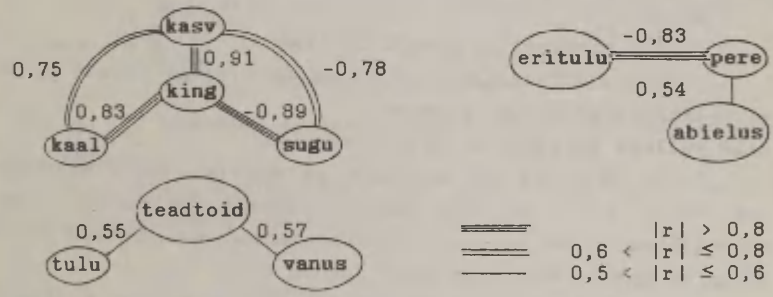
Konstrueerime tasemegraafi joonisel 9.3 esitatud korrelatsioonimaatriksi põhjal.

Jooniselt 9.5 hakkab silma kaks tugevasti korreleeritud tunnusterühma: üks, milles on inimese keha suurust iseloomustavad näitajad kasv, kaal ja kinganumber ning sugu (ülejäänutega negatiivselt korreleeritud, sest on kodeeritud mees - 1, naine - 2), ning teine, mis seob negatiivselt pere suurust ja sissetulekut pereliikme kohta - eritulu. Kõik joonisel 9.5 kujutatud korrelatsioonid on suuremad kui 0,75.



Jn 9.5

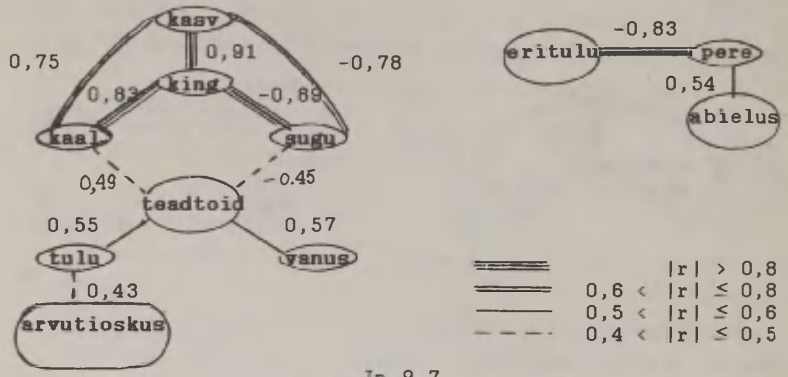
Joonisele 9.6 on kantud korrelatsioonid alates väärtusest 0,5 (tegelikult 0,54). Näeme, et tunnusega pere seostub (positiivselt seotud) tunnus abielus, mis on ka täiesti ootuspärane. Lisaks sellele formeerub uus tunnusterühm, mis seob teadustööde arvu vastaja vanusega ja sissetulekuga - ootuspäraselt on korrelatsioon positiivne.



Jn 9.6

Joonisele 9.7 on kantud ka korrelatsioonid, mis on suuremad kui 0,4. Näeme, et teadustööde arv seostub soo, kinganumbri ja kaaluga. See, et meestel on teadustöid rohkem (korrelatsioon on negatiivne), on igati ootuspärane. Seos teadustööde arvu ja kehakaalu, veel enam aga kinganumbri vahel tundub esialgu kummalisena. Meenutagem aga seda, et

niihästi kinganumber kui ka kaal on tugevasti korreleeritud sooga. Siit tulebki, et teaduslikus mõttes on viljakamad enamasti mehed, kuid mehed on raskemad ja kannavad suuremaid jalavarje.



Jn 9.7

Peale selle ilmneb veel huvitav seos sissetuleku ja arvutioskuse vahel. Lõpptulemusena koosneb graaf jällegi kahest osast, ühes neist 8, teises 3 tunnust.

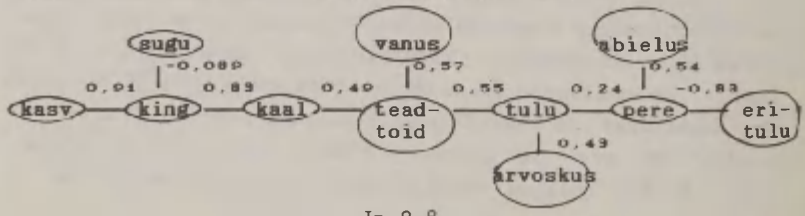
Kuivõrd

- kõik tunnused on graafi lülitatud,
- kõik statistiliselt olulised (olulisuse nivool 0,05) korrelatsioonid on kirjeldatud,

siis sellega lõpetame graafi koostamise.

Teine võimalus on koostada nn suurima korrelatsiooni tee, vt nt [10]; sellele vastavat graafi (silmusteta teed, millele vastavate korrelatsioonikordajate absoluutväärtuste summa on maksimaalne), kujutab joonis 9.8.

Näeme siin samuti kolme suhteliselt iseseisvat tunnuserühma, mida omavahel ühendavad nõrgemad seosed.



Jn 9.8

9.1.6. Osakorrelatsioonide arvutamine (*Partial Correlation Analysis*). Korrelatsioonide analüüsimisel paneme tihti tähele seda, et kahe tunnuse vaheline seos võib olla põhjustatud hoopis mingi kolmanda tunnuse mõjust mõlemale vaadeldavale tunnusele. Selline oli olukord teadustööde arvukuse ja kinganumbri korral, mis tulenes soo mõjust mõlemale tunnusele.

Seosekordajaid, mis kirjeldavad kahe tunnuse omavahelist korrelatiivset seost tingimusel, et kolmanda tunnuse/ /mingi tunnuste rühma lineaarne mõju on elimineeritud, nimetatakse osakorrelatsioonikordajateks. Osakorrelatsioonikordajate arvutamiseks on ette nähtud protseduur *Partial Correlation Analysis*, kolmas allmenüüst *Q. Multivariate Methods*.

Olgu X ja Y uuritavad tunnused, Z - segav tunnus, mille mõju soovitakse kõrvaldada. Vastava osakorrelatsioonikordaja saamiseks arvutatakse tunnuste X ja Y prognoosijäägid lineaarse regressiooniseose (vt p 7.5.2) abil:

$$\varepsilon_X = X - (a + bZ),$$
$$\varepsilon_Y = Y - (c + dZ),$$

ning leitakse korrelatsioonikordaja prognoosijääkide ε_X ja ε_Y vahel, mis ongi otsitavaks osakorrelatsioonikordajaks $r_{XY/Z}$,

$$r_{XY/Z} = r(\varepsilon_X, \varepsilon_Y),$$

kus r(X, Y) tähistab suuruste X ja Y lineaarset korrelatsioonikordajat.

Korrelatsioonikordajate r_{XY}, r_{XZ} ja r_{YZ} kaudu avaldub osakorrelatsioonikordaja järgmiselt:

$$r_{XY/Z} = \frac{r_{XY} - r_{XZ} \cdot r_{YZ}}{\sqrt{1-r_{XZ}^2} \cdot \sqrt{1-r_{YZ}^2}}.$$

Seda korrelatsioonikordajat nimetatakse esimest järku osakorrelatsioonikordajaks.

Samal viisil võib tunnustest X ja Y elimineerida ka mitme tunnuse Z_1, Z_2, ..., Z_q mõju. Sel juhul on tulemuseks q-järku osakorrelatsioonikordaja $r_{XY/Z_1 \ldots Z_q}$

Protseduuri *Partial Correlation Analysis* tellimuspaneel on analoogiline eelmistega. Tulemuseks on maatriksid, mis põhimõtteliselt sarnanevad korrelatsioonimaatriksiga (vt jn 9.9 ja 9.10). Peadiagonaalil paiknevad väärtused -1 on süsteemi SG ebakorrektsus, tegelikult on ka iga tunnuse osakorrelatsioon iseendaga võrdne ühega.

Partial Correlations

	kasv	kaal	sugu
kasv	-1.00000	0.46685	-0.54080
	(28)	(28)	(28)
kaal	0.46685	-1.00000	-0.26060
	(28)	(28)	(28)
sugu	-0.54080	-0.26060	-1.00000
	(28)	(28)	(28)

Correlation (sample size)

Jn 9.9

Partial Correlations

	kasv	kaal	sugu	king
kasv	-1.00000	-0.08125	0.21861	0.69152
	(28)	(28)	(28)	(28)
kaal	-0.08125	-1.00000	0.22323	0.53079
	(28)	(28)	(28)	(28)
sugu	0.21861	0.22323	-1.00000	-0.71251
	(28)	(28)	(28)	(28)
king	0.69152	0.53079	-0.71251	-1.00000
	(28)	(28)	(28)	(28)

Correlation (sample size)

Jn 9.10

SG-süsteemis arvutatakse tellimuses märgitud tunnusvektori (X_1, ..., X_m) puhul maatriksi elemendiks indeksitega i,j osakorrelatsioonikordaja

$$r_{i,j/(1, \ldots, m)}*,$$

kus jada (1, ..., m)* = (1, ..., i-1, i+1, ..., j-1, j+1, ..., m), st iga kahe tunnuse osakorrelatsioon on arvutatud nõnda, et kõigi ülejäänute mõju on elimineeritud.

Nii on joonisel 9.9 antud kasvu ja kaalu osakorrelatsioon tingimusel, et soo mõju on elimineeritud, st on arvestatud kasvu ja kaalu seos meestel ja naistel eraldi võetuna; on näha (võrreldes joonisega 9.3), et see on mõnevõrra nõrgem kui kasvu ja kaalu seos üldkogumis. Samuti väheneb kasvu ja soo omavaheline seos kaalu mõju kõrvaldamisel, st samakaaluliste objektide vaatlemisel. Üsna nõrgaks jääb kaalu ja soo korrelatsioon kasvu mõju elimineerimisel, st sama pikkusega mehed ja naised ei erine kaalu poolest kuigivõrd.

Joonisel 9.10 on esitatud teist järku osakorrelatsioonid, st iga tunnuspaari puhul on kahe ülejäänud tunnuse mõju kõrvaldatud.

9.1.7. Astakkorrelatsioonid (*Rank Correlation Coefficients*). Hüpoteeside kontrollimisel korrelatsioonide olulisuse kohta kasutatakse eeldust, et tunnused on arvulised ja ühegi tunnuspaari jaotus ei erine oluliselt kahemõõtmelisest normaaljaotusest. Kui need eeldused ei ole täidetud, on sobiv tunnustevaheliste seoste kirjeldamiseks kasutada astakkorrelatsioonikordajaid. Märgime siin, et astakkorrelatsioonikordajad mõõdavad tunnustevahelise monotoonse seose tugevust (vt p 7.1.4).

Astakkorrelatsioonikordajad esitatakse samuti kui lineaarsed korrelatsioonikordajadki maatriksina. Astakkorrelatsioonikordajate maatriksi arvutamiseks on SG-süsteemis ette nähtud protseduur *Rank Correlation Coefficients* - viies allmenüüst *R. Nonparametric Methods*.

Jooniselt 9.11 näeme, et selle protseduuri tellimus sarnaneb teiste korrelatsioonitellimustega, erinev on aga üks parameeter nimega *Procedure*, millel on kaks võimalikku väärtust *Spearman* (vaikimisi) ja *Kendall*. Joonistel 9.12 ja 9.13 on toodud protseduuri tulemused korrelatsioonimaatriksite näol, milles, nagu ka joonisel 9.3 esitatud maatriksis, on lisaks kordaja väärtustele antud ka valimi mahud ja olulisuse tõenäosused.

Rank Correlation Coefficients

Data vectors: kaal
kasv
king
vanus
sugu
abielus

Missing values: Listwise

Procedure: Spearman

Jn 9.11

Spearman Rank Correlations

	kaal	kasv	king	vanus	sugu	abielus
kaal	1.0000	.7761	.8196	.4450	-.6996	-.1806
	(28)	(28)	(28)	(28)	(28)	(28)
	1.0000	.0001	.0000	.0208	.0003	.3481
kasv	.7761	1.0000	.9259	.1305	-.8250	-.2090
	(28)	(28)	(28)	(28)	(28)	(28)
	.0001	1.0000	.0000	.4976	.0000	.2774
king	.8196	.9259	1.0000	.2967	-.8586	-.1953
	(28)	(28)	(28)	(28)	(28)	(28)
	.0000	.0000	1.0000	.1232	.0000	.3102
vanus	.4450	.1305	.2967	1.0000	-.1934	-.0381
	(28)	(28)	(28)	(28)	(28)	(28)
	.0208	.4976	.1232	1.0000	.3150	.8430
sugu	-.6996	-.8250	-.8586	-.1934	1.0000	.0221
	(28)	(28)	(28)	(28)	(28)	(28)
	.0003	.0000	.0000	.3150	1.0000	.9087
abielus	-.1806	-.2090	-.1953	-.0381	.0221	1.0000
	(28)	(28)	(28)	(28)	(28)	(28)
	.3481	.2774	.3102	.8430	.9087	1.0000

Coefficient (sample size) significance level

Jn 9.12

Kendall Rank Correlations

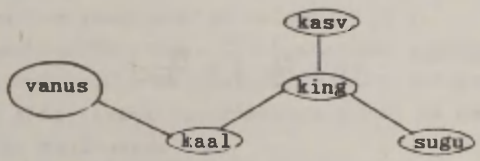
	kaal	kasv	king	vanus	sugu	abielus
kaal	1.0000	.6000	.6480	.3385	-.5934	-.1532
	(28)	(28)	(28)	(28)	(28)	(28)
	1.0000	.0000	.0000	.0061	.0000	.1714
kasv	.6000	1.0000	.8142	.0734	-.6999	-.1773
	(28)	(28)	(28)	(28)	(28)	(28)
	.0000	1.0000	.0000	.5536	.0000	.1166
king	.6480	.8142	1.0000	.1946	-.7368	-.1676
	(28)	(28)	(28)	(28)	(28)	(28)
	.0000	.0000	1.0000	.1156	.0000	.1349
vanus	.3385	.0734	.1946	1.0000	-.1661	-.0327
	(28)	(28)	(28)	(28)	(28)	(28)
	.0061	.5536	.1156	1.0000	.1454	.7654
sugu	-.5934	-.6999	-.7368	-.1661	1.0000	.0221
	(28)	(28)	(28)	(28)	(28)	(28)
	.0000	.0000	.0000	.1454	1.0000	.8650
abielus	-.1532	-.1773	-.1676	-.0327	.0221	1.0000
	(28)	(28)	(28)	(28)	(28)	(28)
	.1714	.1166	.1349	.7654	.8650	1.0000

Coefficient (sample size) significance level

Jn 9.13

Võrreldes maatrikseid joonistelt 9.3, 9.12 ja 9.13, näeme, et keskmiselt on Spearmani korrelatsioonikordajad absoluutväärtuselt pisut suuremad kui lineaarsed, Kendalli omad aga väiksemad. Samal ajal on aga Kendalli seosekordajatel tihti väikseimad olulisuse tõenäosused.

Isegi olulisi korrelatsioone sisaldav korrelatsioonigraaf tuleks astakkorrelatsioonikordajate baasil pisut erinev (vt jn 9.14), sest lisandub statistiliselt oluline seos



Jn 9.14

tunnuste vanus ja kaal vahel. See tulemus illustreerib tõsiasja, et kasutades lineaarseid korrelatsioone astakkorrelatsioonide asemel, võime me küll seoseid alahinnata (jääda põhjendamatult nullhüpoteesi juurde), kuid enamasti pole erilist ohtu, et põhjendamatult võetakse vastu sisukas hüpotees.

Ka astakkorrelatsioonimaatriksi võib salvestada. Seda võib valikuliselt kasutada ka edasiste arvutuste juures. Tuleb aga olla tähelepanelik selliste protseduuride puhul, mis matemaatiliste eelduste kohaselt peaksid kasutama lähteandmestikuna lineaarset korrelatsioonimaatriksit (nt faktoranalüüs); kindlasti ei saa sellisel juhul teha tõestatud järeldusi, küll aga võib genereerida uusi hüpoteese.

§ 9.2. Peakomponentide analüüs ja faktoranalüüs

9.2.1. Informatsiooni "tihendamine" peakomponentide analüüsi (ehk komponentanalüüsi) abil. Tunnusvektor **X**, milles leidub omavahel tugevasti korreleeritud üksiktunnuseid, sisaldab "liiga palju" tunnuseid - nendes tunnustes sisalduv informatsioon (dispersioonide ja kovariatsioonide, s.o hajuvuse mõttes) on üldiselt edasi antav ka märksa väiksema arvu tunnuste kaudu, kui valida nendeks sobival viisil lähtetunnuste lineaarkombinatsioonid.

Illustreerime seda joonise 9.15 abil. Siin on kujutatud kahe korreleeritud ($r = 0,5$) tunnuse **X** ja **Y** korrelatsiooniväli, kusjuures mõlema tunnuse dispersioonid on võrdsed. Sarnasuguse hajuvusdiagrammi võiksime saada ka tunnuste **U** ja **V**, st tunnuste **X** ja **Y** lineaarteisenduse abil saadud tunnuste jaoks, kus

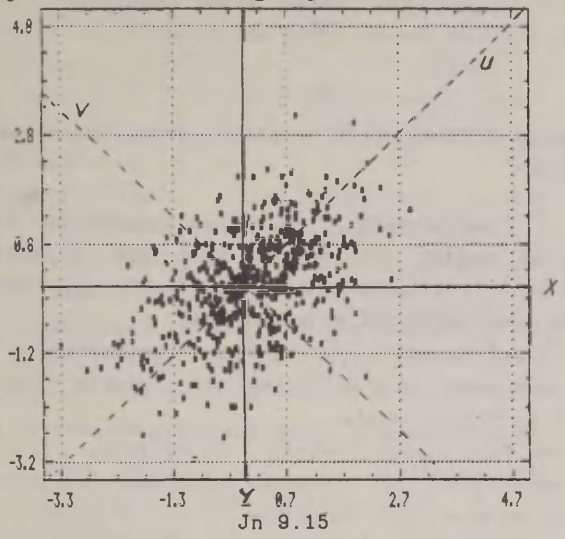
$$U = \frac{\sqrt{2}}{2} X + \frac{\sqrt{2}}{2} Y$$

ja

$$V = \frac{\sqrt{2}}{2} Y - \frac{\sqrt{2}}{2} X.$$

Viimase hajuvusdiagrammi teljed on aga 45° võrra pööratud, vt jn 9.15.

Paneme tähele, et DU = 1,5, DV = 0,5, seega U esitab 75 % tunnuspaari informatsioonist hajuvuse mõttes, teise tunnuse V arvele jääb ainult 25 % koguhajuvusest.



Jn 9.15

Peakomponentide analüüsi idee seisnebki selles, et leida niisugused lähtetunnuste lineaarkombinatsioonid

$$U_i = \sum_{j=1}^{m} c_{ij} X_j, \quad i = 1, \ldots, m, \tag{9.1}$$

mis rahuldaksid järgmisi tingimusi:

1) $\sum_{j=1}^{m} c_{ij}^2 = 1$, i = 1, ..., m,

2) $EU_i = 0$, i = 1, ..., m,

3) DU_1 on maksimaalne,

4) $E(U_i U_j) = 0$ (ehk $U_i \perp U_j$), kui j = 1, ..., i-1,

5) DU_i on maksimaalne kõigi tingimusi 2) ja 4) rahuldavate lineaarkombinatsioonide hulgas, i = 2, ..., j.

Lineaarkombinatsioone U_i nimetatakse peakomponentideks.

Seega esimene peakomponent määrab jaotuse n-ö suurima hajuvusega sihi, teine on esimesega risti ja omakorda suurima võimaliku hajuvusega jne.

Osutub, et peakomponentide leidmine taandub analüüsitava korrelatsiooni- või kovariatsioonimaatriksi omaväärtuste ja omavektorite leidmisele.

Omaväärtused (järjestatuna kahanevalt) võrduvad peakomponentide dispersioonidega:

$$DU_1 = \lambda_1,$$
$$DU_2 = \lambda_2,$$
$$\ldots\ldots$$
$$DU_m = \lambda_m.$$

Peakomponentide sihid on määratud vastavate omavektorite sihtidega.

9.2.2. Peakomponendid *Principal Components*. Komponentanalüüs on neljas protseduur allmenüüs *Q. Multivariate Methods*. Selle tellimuspaneel (vt jn 9.16) võimaldab sisestada algandmed järgneval kolmel viisil:

1° korrelatsiooni- või kovariatsioonimaatriks *Correlation or covariance matrix* ja tunnuste loetelu sisaldav sümboltunnus *Variable labels*;

2° tunnuste nimede loetelu *Data vectors*;

3° andmestiku nimi *filename* (eeldusel, et andmestik on objekt-tunnus-maatriksi kujuline, vt [7], § 2.6).

Principal Components

Correlation or covariance matrix: ASS.korel
and
Variable labels: loetelu
or
Data vectors
or filename:

Standardize: Yes Missing values: Listwise Point labels:

Jn 9.16

Kõigi esituste korral kehtivad nõuded:

- kõik tunnused peavad olema arvulised;
- tunnused peavad olema võrdse pikkusega.

Täiendavate parameetrite käsitlemisel märgime järgmist.

Parameeter *Standardize* (standardiseerida) omab mõtet ainult algandmete ja kovariatsioonimaatriksi puhul ning selle kasutamise korral on faktoranalüüsi lähtematerjaliks korrelatsioonimaatriks.

Parameetri *Missing values* käsitlus on standardne, vt p 9.1.2.

Valik *Point labels* annab võimaluse märgendada graafikul punkte (esialgseid tunnuseid).

Standardtrükist peakomponentide analüüsile kujutab joonis 9.17. Tabeli teises veerus (*Percent of Variance*) on esitatud komponentidele vastavad hajuvuse kirjeldatuse protsendid (mis on arvutatud komponentide dispersioonide λ_i järgi)

$$\frac{\lambda_i}{m} \cdot 100\ \%,$$

kus m on tunnuste arv ja $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$ on maatriksi omaväärtused. Viimane veerg kujutab parimate 1-, 2-, 3- jne komponendiliste komplektide poolt kirjeldatavat hajuvuse osa; näeme, et kolme komponendi abil on võimalik kirjeldada üle 70 % 11 tunnuse hajuvusest (seega teatud mõttes nendega edastatavast informatsioonist).

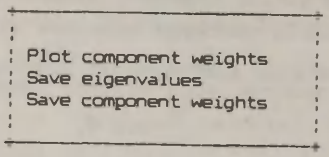
Principal Components Analysis for ASS.korel

Component Number	Percent of Variance	Cumulative Percentage
1	35.89778	35.89778
2	20.66370	56.56148
3	14.69849	71.25997
4	13.48695	84.74693
5	5.95016	90.69708
6	4.51409	95.21117
7	1.96793	97.17910
8	1.31326	98.49236
9	.97825	99.47061
10	.31147	99.78208
11	.21792	100.00000

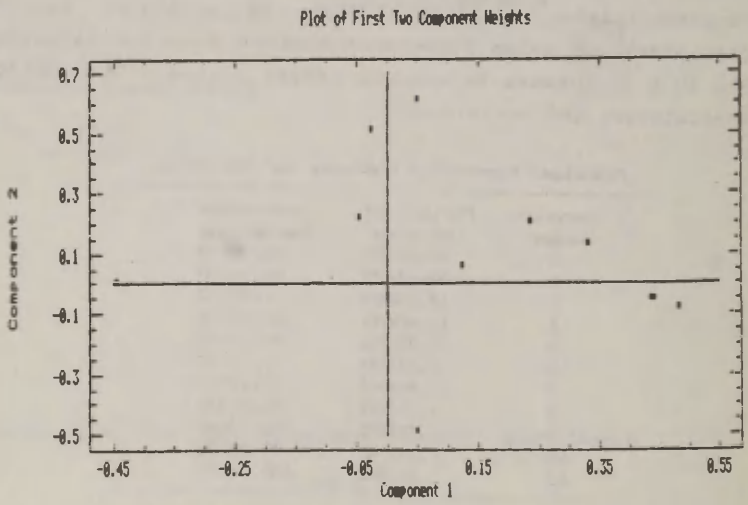
Jn 9.17

Täiendavate võimaluste loetelu on esitatud aknas, vt jn 9.18. Esimesele reale (*Plot component weights*) vastab tunnuste (kui punktide) graafik (vt jn 9.19), millel telgedeks kaks esimest peakomponenti. Punktide vastavuse tunnustele saab kindlaks teha, kasutades interaktiivset režiimi (vt [7], p 2.8.4). Järgmine rida *Save eigenvalues* salvestab omaväärtused, mida aga tulemuste tõlgendamiseks enam otseselt tarvis ei ole. Seevastu viimane rida *Save component weights* võimaldab salvestada teisendusmaatriksi B = (b_{ij}) (i, j = = 1, ..., m), mille põhjal esialgsed tunnused $\mathbf{X}_i$ avalduvad peakomponentide kaudu järgmiselt:

$$\mathbf{X}_i = \sum_{j=1}^{m} b_{ij}\mathbf{U}_j, \quad i = 1, \ldots, m. \tag{9.2}$$



Jn 9.18



Jn. 9.19

Salvestatud maatriksi võib ekraanile kutsuda ja välja trükkida standardsel viisil, kasutades selleks näiteks protseduuri *Display Data Directory* (vt [7], p 4.1.8). Esimest osa sellest salvestatud maatriksist näeme joonisel 9.20.

Variable: KORR.WEIGHTS (length = 11 11)

(1, 1) 0.439854	(1, 2) −0.0475512	(1, 3) −0.152615
(2, 1) 0.435195	(2, 2) −0.0473811	(2, 3) −0.0191925
(3, 1) 0.48108	(3, 2) −0.0744725	(3, 3) −0.0645306
(4, 1) 0.12231	(4, 2) 0.059105	(4, 3) −0.249848
(5, 1) −0.445531	(5, 2) 0.0139321	(5, 3) 0.0571586
(6, 1) −0.0260643	(6, 2) 0.514186	(6, 3) 0.0818693
(7, 1) 0.0492651	(7, 2) 0.614136	(7, 3) −0.145907
(8, 1) 0.235036	(8, 2) 0.210056	(8, 3) 0.581368
(9, 1) 0.330658	(9, 2) 0.133281	(9, 3) 0.120632
(10, 1) −0.0466865	(10, 2) 0.223594	(10, 3) 0.569068
(11, 1) 0.0487338	(11, 2) −0.482522	(11, 3) 0.449572

Jn 9.20

Selle tabeli (maatriksi) kaks esimest veergu määravadki punktide koordinaadid joonisel 9.19. Nii veendume näiteks, et kõige parempoolne punkt joonisel vastab tunnusele nr 3, mis meie loetelus (vt nt jn 9.3) on **king**, kõige vasakpoolsem aga tunnusele 5 – **sugu**. Kõige ülemine ja kõige alumine punkt tähistavad vastavalt tunnust 7 (pere, st pere suurus) ja tunnust 11 (**eritulu**, st tulu pereliikme kohta). Kõige lähemal 0-punktile, st kõige halvemini kirjeldatud on tunnus 4 (**vanus**).

Peakomponentide analüüsis esitatakse maatriksi kõik m veergu, kuigi viimaste veergude osatähtsus andmestiku kirjeldamisel võib olla väga väike (joonisel 9.17 esitatud tabelist on näha, et eelviimane ja viimane peakomponent kirjeldavad tunnuste komplekti hajuvusest vastavalt 0,3 ja 0,2 %).

9.2.3. Faktoranalüüs. Peakomponentide analüüs on küll matemaatiliselt teatud mõttes optimaalne protseduur tunnustes sisalduva informatsiooni tihendamiseks ja ühtlasi ka tunnustevaheliste seoste struktuuri uurimiseks, kuid ometi ei ole ta praktilises andmeanalüüsis eriti populaarne. Tunduvalt kasutatavam on põhimõtteliselt sama matemaatilist ideed rakendav, kuid informatsioonilisest seisukohast hoopiski avaramalt välja arendatud faktoranalüüs.

Faktoranalüüsi puhul aktsepteeritakse tõsiasja, et uued tunnused $U_1, \ldots, U_m$ ei tarvitse kirjeldada lähtetunnuste dispersioone täielikult, kuna teatav osa tunnuste summaarsest hajuvusest ei ole süstemaatiline (seetõttu ei paku ka huvi), vaid on tingitud juhusest - seda osa nimetatakse tunnuste omapäraks. Seega faktoranalüüsi põhivõrrand

$$X_i = \sum_{j=1}^{k} a_{ij}F_j + \varepsilon_i, \quad i = 1, \ldots, k \qquad (8.3)$$

kus F_j on uued, nn faktortunnused, erineb komponentanalüüsi põhivõrrandist (8.2). Kordajaid a_{ij} (maatriksi A elemente) nimetatakse faktorkaaludeks (ka faktorlaadungiteks, *loadings*). Erinevus on selles, et

- faktorite arv k on väiksem lähtetunnuste arvust m,
- teatav osa tunnusest on jäänud kirjeldamata, moodustades nn jääkliikme ε_i.

Teatav kokkuleppeline, mitte küll põhimõtteline erinevus on ka uute tunnuste kohta tehtud eeldustes. Lisaks tsentreeritusele ja sõltumatusele

$$EF_i = 0; \quad EF_iF_j = 0, \text{ kui } i \neq j,$$

eeldatakse veel, et

$$DF_j = 1, \quad j = 1, \ldots, k. \qquad (8.4)$$

Seega on faktorite kovariatsioonimaatriksiks ühikmaatriks I ning ülesanne taandub sellise võimalikult väikese veergude arvuga k maatriksi A määramisele, et kehtiks võrdus

$$R = A'A + G,$$

kus R on lähtetunnusvektori X korrelatsioonimaatriks ja G diagonaalmaatriks.

Faktorite F_j ja neid määrava maatriksi $A = (a_{ij})$ määramisel on võimalik lähtuda kahest põhimõttest.

1° Peakomponentide printsiip.

Faktorid leitakse, kasutades peakomponentide analüüsi ning valides välja sobiva komponentide arvu k nii, et k esimest peakomponenti kirjeldaksid uuritavat andmestikku piisava määrani. Iga tunnuse korral see osa, mis jääb k faktoriga kirjeldamata, moodustabki omapära ε_i.

Paneme tähele seda, et tingimuse (9.4) tõttu faktorid erinevad peakomponentidest,

$$F_j = \frac{1}{\sqrt{\lambda_j}} U_j,$$

ning sellest tulenevalt

$$a_{ij} = \sqrt{\lambda_j}\, b_{ij}, \quad i = 1, \ldots, k; \; j = 1, \ldots, k.$$

See protseduur on alati teostatav ning selle aluseks on maatriksi R (erijuhul kovariatsioonimaatriksi S) omaväärtuste ja omavektorite leidmine.

2° Klassikaline faktoranalüüs.

Klassikalise faktoranalüüsi puhul hinnatakse juba ette ära, milline osa (protsentides) iga tunnuse dispersioonist jääb faktorite poolt kirjeldamata, st hinnatakse suurusi

$$D\varepsilon_i = d_i^2, \quad 0 \le d_i^2 < 1$$

(siin me kasutame asjaolu, et korrelatsioonimaatriksi arvutamisel käsitletakse kõiki tunnuseid normeeritutena, $DX_i = 1$, $i = 1, \ldots, m$).

Tähistades

$$\operatorname{diag}(d_1^2, \ldots, d_m^2) = G \tag{9.5}$$

ning eeldades omapärade sõltumatust faktoritest, saame ülesande taandada nn redutseeritud korrelatsioonimaatriksi

$$\bar{R} = R - G \tag{9.6}$$

omaväärtusülesande lahendamisele. Maatriks $\bar{R}$ erineb esialgsest korrelatsioonimaatriksist R ainult selle poolest, et tema peadiagonaalil paiknevad ühtede asemel vahed $1-d_i^2$.

Selleks, et omaväärtusülesanne oleks lahenduv, peab ka maatriks (9.6) olema mittenegatiivselt määratud. Kui maatriks A on leitud, on suurusi $1-d_i^2$ lihtne leida: faktorid $F_1, \ldots, F_k$ kirjeldavad tunnusest osa

$$h_i^2 = \sum_{j=1}^{k} a_{ij}^2 \quad (i = 1, \ldots, m). \tag{9.7}$$

Suurust h_i^2 nimetatakse kommunaliteediks,

$$0 \le h_i^2 \le 1. \tag{9.8}$$

Kommunaliteet näitab seda osa i-nda tunnuse X_i hajuvusest,

mis on faktorite F_1, ..., F_k poolt kirjeldatud, seega $h_i^2 = 1 - d_i^2$.

Siit tulenebki idee klassikalise faktoranalüüsi iteratiivseks lahenduseks - esimesel sammul hinnatakse mingil viisil maatriks G, seejärel lahendatakse maatriksi $\bar{R}$ jaoks omaväärtusülesanne, määratakse maatriks A. Maatriksi A järgi arvutatakse kommunaliteedid valemist (9.7), nende arvel täpsustatakse maatriksi $\bar{R}$ hinnangut jne.

Saadud iteratiivne protsess enamasti koondub, kuid mitte alati, on olemas ka võimalus, et saadakse võõrlahend (mõni arvutustulemustest h_i^2 või a_{ij} ei kuulu lubatud piirkonda). Meenutame, et kommunaliteetide h_i^2 jaoks peab kehtima seos (9.8), seetõttu ka

$$|a_{ij}| \leq 1 \quad (i = 1, \ldots, m;\ j = 1, \ldots, k) \qquad (9.9)$$

Mittekoonduvust võib põhjustada muuhulgas ka see, et korrelatsioonimaatriks ei ole kooskõlaline, sest kasutatud on tühikute eraldamist paarikaupa.

Võrreldes peakomponentide ja klassikalist faktoranalüüsi tuleb märkida, et teisel meetodil ei ole erilisi eeliseid peale teatavate traditsioonide (eriti psühholoogias jne), kuid ta on mõnevõrra töömahukam, seetõttu aeglasem (iteratiivne protsess) ning võib vastuseks anda ka võõrlahendi. Sellisel juhul tuleb soovitada kasutada peakomponentide meetodit. **SG** klassikaline faktoranalüüs piirdub esimese sammuga.

9.2.4. Faktoranalüüsi protseduur (*Factor Analysis*). Faktoranalüüs on viies protseduur allmenüüs *Q. Multivariate Methods*. Selle tellimuspaneel (vt jn 9.21) võimaldab sisestada algandmed järgneval kolmel viisil:

1° korrelatsiooni- või kovariatsioonimaatriks *Correlation or covariance matrix* ja tunnuste loetelu sisaldav sümboltunnus *Variable labels*;

2° tunnuste nimede loetelu *Data vectors*;

3° andmestiku nimi *filename* (eeldusel, et andmestik on objekt-tunnus-maatriksi kujuline, vt [7], § 2.6).

Kõigi esituste korral kehtivad nõuded:

- kõik tunnused peavad olema arvulised;
- tunnused peavad olema võrdse pikkusega.

Factor Analysis

Correlation or covariance matrix: korel
and
Variable labels: loetelu
or
Data vectors
or filename:

Standardize: Yes Missing values: Listwise Point labels:
Type of rotation: Varimax Convergence criterion: 1E-5 Iterations: 100

Jn 9.21

Täiendavate parameetrite kohta märgime järgmist.

Parameeter *Standardize* (standardiseerida) omab mõtet ainult algandmete ja kovariatsioonimaatriksi korral, selle kasutamise korral on faktoranalüüsi lähtematerjaliks korrelatsioonimaatriks.

Parameetri *Missing values* käsitlus on standardne, vt p 9.1.2.

Valik *Point labels* annab võimaluse märgendada graafikul punkte (esialgseid tunnuseid).

Pööramismeetoditele *Type of rotation* pöörame tähelepanu punktis 9.2.6. Üldiselt sobivad niihästi sellele kui ka iteratsiooniprotsessi reguleerivatele parameetritele *Convergence criterion* ja *Iteration* täiesti hästi vaikimisi pakutud väärtused, seega pole tavaliselt põhjust neid muuta.

Pärast tellimuspaneeli täitmist ja käivitamist ilmub ekraanile küsimus selle kohta, millist faktoranalüüsi tüüpi valida (vt jn 9.22) - kas te soovite diagonaalielemendid asendada kommunaliteetidega. Vastus "jah" (*Yes*) tähistab klassikalist, vastus "ei" (*No*) peakomponentide faktoranalüüsi.

Do you wish to replace diagonal elements with communalities? (N/Y):

Jn 9.22

9.2.5. Peakomponentide faktoranalüüs. Vaatleme kõigepealt teist varianti. Esimene tulemuste trükis siin on tabel jooniselt 9.23. Pöörame tähelepanu sellele, et sisuliselt on siin tegemist kahe kõrvuti asetatud tabeliga, millel omavahel mingit seost ei ole.

Variable	Communality	Factor	Eigenvalue	Percent Var	Cum Percent
kaal	0.79871	1	3.94876	35.9	35.9
kasv	0.90215	2	2.27301	20.7	56.6
king	0.95777	3	1.61683	14.7	71.3
vanus	0.77802	4	1.48356	13.5	84.7
sugu	0.87771	5	.65452	6.0	90.7
abielus	0.41556	6	.49655	4.5	95.2
pere	0.93331	7	.21647	2.0	97.2
tulu	0.86867	8	.14446	1.3	98.5
teadtoid	0.82694	9	.10761	1.0	99.5
arvutoskus	0.43918	10	.03426	.3	99.8
eritulu	0.92918	11	.02397	.2	100.0

Jn 9.23

Esimene tabel haarab kaht esimest veergu ning sisaldab tunnuste loetelu ja nendele vastavaid esialgselt hinnatud kommunaliteete(hinnangud on arvutatud mitmeste korrelatsioonikordajate baasil, vt p 10.1.3).

Vaadeldes konkreetset andmestikku tabelis, näeme, et esialgse hinnangu kohaselt on faktorite abil võimalik kirjeldada üle 95 % tunnusest **king** ja 93 % tunnusest **pere**, kuid ainult 42 % tunnusest **abielus**. Kommunaliteete summeerides näeme, et kokku prognoositakse kirjeldatust ca 80 % (täpsemalt 79,3 %) ulatuses.

Kuivõrd see hinnang ei ole seotud konkreetse faktorite arvuga, ei tarvitse lõppjäreldus sellega ka ühtida.

Teine tabel sisaldab esialgse informatsiooni kõigi 11 peakomponendi järgi arvutatud faktori kohta. See informatsioon langeb suurel määral ühte teabega tabelis joonisel 9.17. Trükitud on omaväärtused *Eigenvalues*, faktorite poolt kirjeldatud hajuvuse protsendid (*Percent Var*) ja summeeritud hajuvuse protsendid (*Cum Percent*). Kõik need näitajad on võrdsed vastavate näitajatega peakomponentide analüüsi korral. Siin on kõige olulisemaks veeruks omaväärtuste

veerg, mille põhjal saab otsustada valitavate faktorite arvu üle, mis tuleb enesel määrata.

Kui ei ole alust lähtuda mõnest muust kriteeriumist (näiteks summaarsest kirjeldatuse protsendist), on sobiv võtta kasutusele just nii palju faktoreid, kui on ühest suuremaid omaväärtusi. Faktorid, mis vastavad ühest väiksematele omaväärtustele, kirjeldavad vähem kui esialgsed tunnused, seetõttu pole nende kasutamine tavaliselt õigustatud.

Käesoleva näite puhul on ilmselt otstarbekas vähemalt esimeses lähenduses võtta kasutusele 4 faktorit, mis tagab peaaegu 85 %-lise esialgse materjali kirjeldatuse.

Järgmisel sammul ilmubki ekraanile küsimus soovitavate faktorite arvu kohta (vt jn 9.24) - sisestada faktorite arv (vaikimisi väärtuseks on 2). Käesolevas näites sisestame arvu 4, mille tulemusena saame ekraanile nn faktormaatriksi A, vt jn 9.25.

```
Enter number of factors to be extracted (2):
```

Jn 9.24

Factor Matrix

Variable/Factor	1	2	3	4
kaal	0.87405	-0.07169	-0.19406	-0.06011
kasv	0.86480	-0.07143	-0.02440	0.38877
king	0.95598	-0.11228	-0.08205	0.17973
vanus	0.24305	0.08911	-0.31769	-0.84859
sugu	-0.88534	0.02100	0.07268	-0.21323
abielus	-0.05179	0.77521	0.10410	-0.16620
pere	0.09790	0.92590	-0.18553	0.15459
tulu	0.46705	0.31669	0.73924	-0.03393
teadtoid	0.65707	0.20094	0.15339	-0.62039
arvutoskus	-0.09277	0.33710	0.72360	0.11847
eritulu	0.09684	-0.72747	0.57165	-0.28168

Jn 9.25

Faktormaatriksi tõlgendamisel on väga soodus kasutada asjaolu, et kordajad a_{ij} on i-nda tunnuse X_i ja j-nda faktori F_j korrelatsioonikordajad:

$$a_{ij} = r(X_i, F_j). \qquad (9.10)$$

Faktoranalüüsi tõlgendamisel on põhiküsimus faktorite identifitseerimine, st faktoritele nime omistamine. Selleks

soovitame koostada iga faktori jaoks järgmise tabeli, mis iseloomustab faktori pluss-poolele (suurtele väärtustele) ja miinus-poolele (väikestele väärtustele) vastavaid tunnuste tüüpilisi väärtusi.

Faktori F_1 jaoks saame

F_1 +	F_1 -
suur kaal	väike kaal
suur kasv	väike kasv
suur king	väike king
soo väärtus väike: 1, st mees	soo väärtus suur: 2, st naine
. . . . .	. . . . .
teadtoid keskmisest pisut suurem	teadtoid keskmisest natuke väiksem
(tulu keskmisest veidi suurem)	(tulu keskmisest veidi väiksem)

Tulemusena võime nimetada faktorit F_1 soo-kehaehituse faktoriks; ilmselt soo mõjul lülituvad sisse mõned kehaehitusse mitte puutuvad tunnused.

Tabeli koostamisel peame meeles reegleid:

1° pluss-poolele kanname nende tunnuste suured väärtused, millele vastavad faktorkaalud a_{ij} on positiivsed;

2° pluss-poolele kanname nende tunnuste väikesed väärtused, millele vastavad faktorkaalud on negatiivsed;

3° kodeeritud tunnuste (sugu) puhul arvestame suurte ja väikeste koodide tähendusi (2 - naine > 1 - mees);

4° miinus-poolele kanname tunnuste vastassuunalised väärtused võrreldes pluss-poolega.

Kahe poolega (+ ja -) tabeli koostamine on eriti vajalik siis, kui faktormaatriksi analüüsitavas veerus on niihästi positiivseid kui negatiivseid elemente.

Tabelisse paigutatakse ainult absoluutväärtuse poolest suhteliselt suured faktorkaalud, neid on sobiv grupeerida, näiteks suured - üle 0,8 või 0,7, mõõdukad - üle 0,5, kuid mitte suured.

Samal viisil koostame tabelid ka järgmiste faktorite jaoks.

F_2 +	F_2 −
suur pere	väike pere
abielus (väärtus 1)	vallaline (väärtus 0)
väike sissetulek pereliikme kohta	suur sissetulek pereliikme kohta

F_3 +	F_3 −
suur tulu	väike tulu
hea arvutioskus	halb arvutioskus
suur tulu pereliikme kohta	väike tulu pereliikme kohta

F_4 +	F_4 −
noor (väike vanus)	vana
vähe teadustöid	palju teadustöid

Näeme, et faktor F_2 on interpreteeritav kui pere-, faktor F_3 kui sissetuleku ja faktor F_4 kui vanuse faktor.

Asjaolu, et faktori F_4 absoluutväärtuselt suured väärtused on negatiivsed, ei ole tõlgendamise seisukohalt oluline. Tõlgendamise hõlbustamiseks võib faktori kõik kaalud (nii suured kui väikesed) korrutada soovi korral läbi arvuga -1, ilma et tõlgendus sellest muutuks. Küll aga muutub selle tagajärjel näiteks graafik, teisendusmaatriks jne.

Arvutades faktormaatriksi veergude kaupa faktorkaalude ruutude summad, võime veenduda seose

$$\sum_{i=1}^{m} a_{ij}^2 = \lambda_j, \quad j = 1, \ldots, k, \tag{9.10}$$

paikapidavuses. Sellepärast ongi esimestes faktorites märksa rohkem suuri faktorkaale kui viimastes.

Valemiga (9.10) määratud suurust nimetatakse j-nda faktori kirjeldusmääraks. Kirjeldusmäär näitab, kui suure osa lähtetunnuste koguhajuvusest e varieeruvusest (mis võrdub tunnuste arvuga k) kirjeldab j-s faktor. Pööramata faktorite puhul võrduvad kirjeldusmäärad omaväärtustega. Kõrvuti kirjeldusmääradega pakuvad huvi ka faktorite suhtelised kirjeldusmäärad $(\sum_{i=1}^{m} a_{ij}^2)/m$, mis tavaliselt avaldatakse protsentides. Faktorite kirjeldusmäärad on esitatud tabelis jn 9.23.

Järgmisel ekraanil (vt jn 9.26) esitatakse kommunaliteedid, mis on arvutatud valemi (9.8) põhjal juba leitud faktormaatriksist. Käesoleval juhul on arvutatud kommunaliteedid esialgselt hinnatutest mõnevõrra suuremad. Selle põhjuseks on h_j^2 definitsioonist tulenev seos

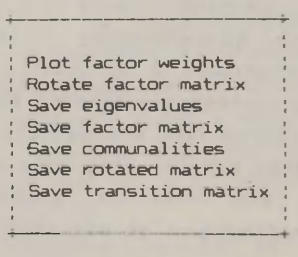
$$\sum_{j=1}^{k} h_j^2 - \sum_{j=1}^{k}\sum_{i=1}^{m} a_{ij}^2 = \sum_{j=1}^{k} \lambda_j$$

ning tõsiasi, et käesoleval juhul $\Sigma\lambda_j$ = 9,32 ja 9,32/11 = = 84,7 %.

Variable	Est Communality
kaal	0.81038
kasv	0.90471
king	0.96553
vanus	0.88804
sugu	0.83501
abielus	0.64210
pere	0.92520
tulu	0.86605
teadtoid	0.88053
arvutoskus	0.65787
eritulu	0.94472

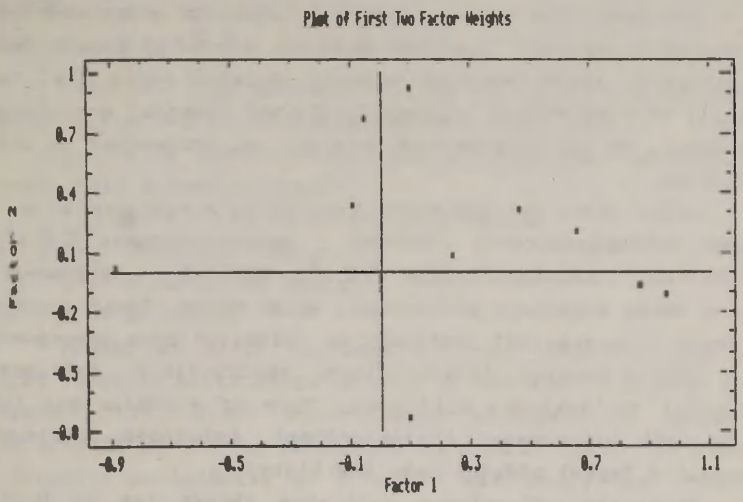
Jn 9.26

Faktoranalüüsi täiendavate võimaluste loetelu on esitatud aknas joonisel 9.27. Esimese valiku (*Plot factor weights*) tulemusena saame joonise 9.28, mis on väga sarnane joonisel 9.19 esitatud graafikuga. Ainsaks erinevuseks on skaala erinevus: faktorkaalude väärtused ulatuvad ±1 lähedusse, olles horisontaalteljel ca 2 (täpsemalt, $\sqrt{3,95}$) ja vertikaalteljel



Jn 9.27

ca 1,5 (täpsemalt, $\sqrt{2,27}$) korda suuremad peakomponentide teljestikus määratud punktide vastavatest koordinaatidest.



Jn 9.28

Järgmist protseduuri - faktorite pööramist - vaatleme punktis 9.2.6.

9.2.6. Faktorite pööramine. Kuigi faktoranalüüs leiab faktorid matemaatilises mõttes optimaalsete lineaarkombinatsioonidena, ei ole sellisel viisil saadud faktorid alati kõige hõlpsamini tõlgendatavad. Enamasti kipub esimene faktor olema seotud väga paljude tunnustega. Tihti on mitmed tunnused olulisel määral korreleeritud mitme erineva faktoriga jne.

Kõige paremini on faktoranalüüsi tulemused tõlgendatavad siis, kui faktorkaalud a_{ij} omavad väärtusi, mis on kas võimalikult lähedased nullile või absoluutväärtuselt ühe lähedal. Tõlgendamisel arvestatakse just viimaseid, pidades silmas seost (9.10).

Graafiliselt võiksime seda ülesannet sõnastada nõnda: tuleb pöörata faktoritele vastavaid koordinaattelgi (vt jn 9.28), säilitades nende omavahelise asendi, nii et teljestiku

uues asendis paikneks võimalikult palju punkte telgedel või nende vahetus läheduses.

Kui faktorite arv on k, siis tuleb telgi pöörata $\frac{k(k-1)}{2}$ erineval tasapinnal (joonisel 9.28 on kujutatud ainult faktoritega F_1 ja F_2 määratud tasand). Kuivõrd pööre ühel tasandil võib mõnevõrra "rikkuda" eelmisel tasandil saavutatud tulemusi, on tarvis pööramist korrata, st protseduur on iteratiivne.

Erinevatele optimaalsuskriteeriumidele vastavad ka erinevad pööramismeetodid (VARIMAX - maksimeeritakse $\sum_i \sum_j a_{ij}^2$, QUARTIMAX - maksimeeritakse $\sum_i \sum_j a_{ij}^4$, EQUIMAX - maksimeeritakse nende avaldiste poolsumma), kuigi nende üldine eesmärk on sama ning enamikul juhtudest on tulemused üsna lähedased.

Pööramismeetod fikseeritakse faktoranalüüsi tellimuspaneelil tühikuklahvi abil aknas *Type of rotation* kas töö algul või hilisematel tööperioodidel, kasutades tellimuspaneelile tagasi pöördumiseks Esc-klahvi.

Faktorite pööramiseks valitakse aknast (vt jn 9.27) teine rida *Rotate factor matrix*.

Pööramise tulemus on esitatud esialgse faktormaatriksiga sarnases tabelis (vt jn 9.29).

VARIMAX ROTATED FACTOR MATRIX

Variable/Factor	1	2	3	4
kaal	0.83115	-0.02279	-0.07703	0.33633
kasv	0.94230	-0.00509	0.07686	-0.10414
king	0.97317	-0.05979	0.02304	0.11983
vanus	-0.00607	0.04406	-0.22328	0.91445
sugu	-0.90839	-0.02782	-0.05384	-0.07847
abielus	-0.17772	0.64878	0.36551	0.23665
pere	0.08306	0.94194	0.17280	0.03460
tulu	0.31683	0.01715	0.86330	0.14173
teadtoid	0.39037	0.01312	0.31096	0.79453
arvutoskus	-0.16194	0.07712	0.77405	-0.16893
eritulu	-0.00057	-0.92055	0.28307	0.13110

Jn 9.29

Lihtne arvutus näitab, et $_k$ pööramise tulemusena

- kommunaliteedid $h_i^2 = \sum_{j=1}^{k} a_{ij}^2$ ei muutu, st iga tunnuse kirjeldatus (ja seega ka summaarne kirjeldatus) jäävad endiseks ;

- muutuvad üksikute faktorite kirjeldusmäärad, kusjuures üldreeglina esimese faktori kaalude ruutude summa $\sum_{i=1}^{m} a_{i1}^2$ väheneb, seevastu aga viimastele faktoritele vastavad summad, sh eriti $\sum_{i=1}^{m} a_{ik}^2$, suurenevad.

Joonisel 9.28 esitatud näites on $\sum_{i=1}^{m} a_{i1}^2 = 3,48$, seega esimese faktori kirjeldusmäär vähenes 0,46 võrra.

Faktorite interpreteerimisel pilt muutus suhteliselt vähe, kuid siiski selgines:

1. faktor - soo ja kehaehituse faktor,
2. - perekonna faktor,
3. - sissetuleku faktor,
4. - ea faktor.

Näeme, et kõige olulisem erinevus võrreldes joonisel 9.25 esitatud maatriksiga on see, et keskmise suurusega (vahemikus 0,4...0,75) korrelatsioonikordajaid on pööratud maatriksis ainult 1, pööramata maatriksis aga 7. Erinevalt pööramata maatriksist on pööratud maatriksis iga mõõdetud tunnus oluliselt korreleeritud ainult ühe faktoriga, mis märgatavalt hõlbustab tõlgendamist. Pööratud faktormaatriksile vastavat graafikut näeme joonisel 9.31.

Seda, et pööramine muutis antud juhul faktormaatriksit vähe, näitab ka teisendusmaatriks T, mis on üsna lähedane ühikmaatriksile, vt jn 9.30,

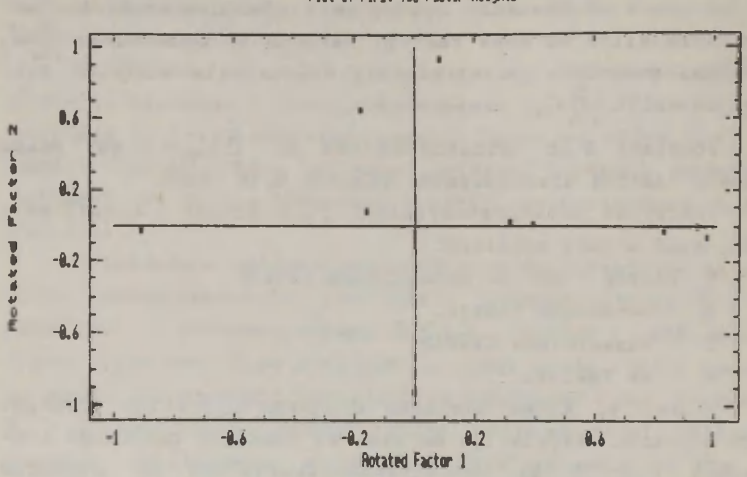
$$A^* = AT'.$$

Eelduse kohaselt on T ortogonaalmaatriks, $T'T = I$.

FINAL TRANSITION MATRIX

Factor	1	2	3	4
1	0.93985	-0.01702	0.14579	0.30844
2	-0.08449	0.91987	0.35665	0.13964
3	-0.11477	-0.35224	0.92279	-0.10589
4	0.31043	0.17167	-0.00315	-0.93496

Jn 9.30



Jn 9.31

Teised kaks pööramismeetodit annavad tõlgenduse poolest üsna lähedased tulemused, vt jn 9.32, 9.33 ja 9.34. Viimasel joonisel on *QUARTIMAX*-meetodil pööratud faktorid. Jooniste 9.31 ja 9.34 vahelised erinevused pole peaaegu märgatavad. Tõsiasi, et erinevate pööramismeetodite rakendamise tulemusena faktorite järjestus (3. ja 4. faktor) muutub, on iseloomulik ja ei oma erilist sisulist tähendust.

QUARTIMAX ROTATED FACTOR MATRIX

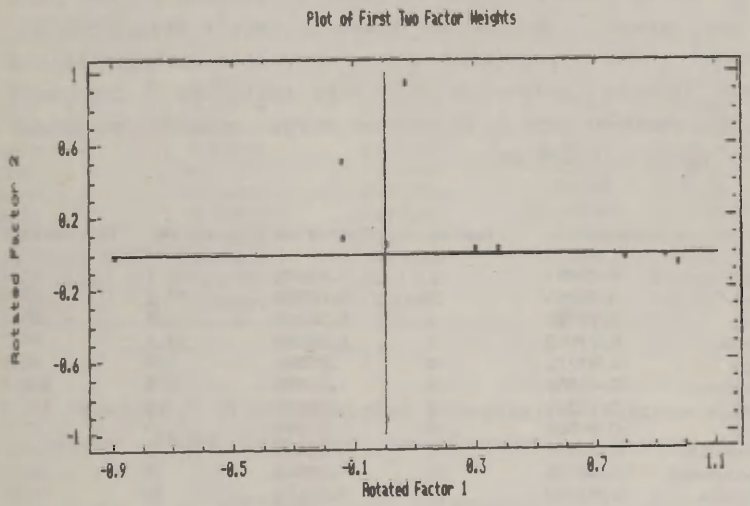
Variable/Factor	1	2	3	4
kaal	0.81973	-0.04507	-0.10925	0.27973
kasv	0.92437	-0.03167	0.03789	-0.16287
king	0.98316	-0.08854	-0.01629	0.05680
vanus	0.06190	0.04246	-0.21154	0.85272
sugu	-0.89651	0.00514	-0.03370	-0.01972
abielus	-0.10699	0.51539	0.29574	0.17200
pere	0.10588	0.94838	0.19017	0.03904
tulu	0.33766	0.02909	0.87560	0.09484
teadtoid	0.44129	0.01305	0.32017	0.72895
arvutoskus	-0.12475	0.09724	0.56052	-0.12151
eritulu	-0.00266	-0.91766	0.29242	0.10575

Jn 9.32

EQUIMAX ROTATED FACTOR MATRIX

Variable/Factor	1	2	3	4
kaal	0.78160	-0.02779	0.38131	-0.08430
kasv	0.93585	-0.01705	-0.04252	0.07424
king	0.96918	-0.07119	0.18239	0.01720
vanus	-0.04008	0.05012	0.85095	-0.22198
sugu	-0.88463	-0.00930	-0.13486	-0.06670
abielus	-0.14862	0.50555	0.15880	0.30207
pere	0.07564	0.94464	0.05170	0.21813
tulu	0.28807	0.01143	0.14751	0.88637
teadtoid	0.33160	0.01279	0.78315	0.32456
arvutoskus	-0.13154	0.07987	-0.13020	0.55976
eritulu	-0.01108	-0.92498	0.11069	0.26616

Jn 9.33



Jn 9.34

9.2.7. Klassikaline faktoranalüüs. Valides esimesele tellimuspaneelile järgneval paneelil küsimusele korrelatsioonimaatriksi diagonaalielementide asendamise kohta (vt jn 9.22) jaatava vastuse *Yes*, oleme sellega käivitanud nn klassikalise faktoranalüüsi protseduuri. Järgneb uus küsimus - kas tellija soovib ise valida kommunaliteetide jaoks väärtusi (lähtuvalt mingist eelnevast teabest) või määratakse need programmiliselt (vaikimisi) *defaults*, vt jn 9.35.

```
Do you wish to replace diagonal elements with communalities? (N/Y): Y
Enter communality estimates (defaults):
```

Jn 9.35

Kommunaliteetide alghinnangud võib anda m arvust koosneva jadana h_1^2, ..., h_m^2 (eraldajaks tühik), kusjuures peab olema täidetud tingimus

$$0 \leq h_i^2 \leq 1.$$

Vaikimisi arvutatakse kommunaliteedid mitmeste korrelatsioonikordajatena (vt p 10.1.3).

Tellimuse esimene tulemus ilmub kaksiktabelina, vt jn 9.36, mis oma struktuurilt langeb kokku tabeliga joonisel 9.23. Ka on täpselt identsed tabelite vasakud pooled (kaks esimest veergu). Erinevused ilmnevad tabeli teises pooles. Kuivõrd siin on esitatud mitte korrelatsioonimaatriksi R (nagu tabelis joonisel 9.23), vaid maatriksi $\bar{R}$ (vt valem (9.6)) omaväärtused, siis on nad varem vaadeldutest erinevad, üldiselt väiksemad.

Variable	Communality	Factor	Eigenvalue	Percent Var	Cum Percent
kaal	0.79871	1	3.82692	42.7	42.7
kasv	0.90215	2	2.07358	23.1	65.8
king	0.95777	3	1.38049	15.4	81.2
vanus	0.77802	4	1.28188	14.3	95.5
sugu	0.87771	5	.29866	3.3	98.8
abielus	0.41556	6	.10392	1.2	100.0
pere	0.93331	7	-.00304	.0	100.0
tulu	0.86867	8	-.01294	.0	100.0
teadtoid	0.82694	9	-.03060	.0	100.0
arvutoskus	0.43918	10	-.09012	.0	100.0
eritulu	0.92918	11	-.10156	.0	100.0

Jn 9.36

See tõsiasi, et maatriksi $\bar{R}$ viimased viis omaväärtust on negatiivsed, näitab seda, et $\bar{R}$ ei ole mittenegatiivselt määratud, piltlikult öeldes - maatriksi R diagonaalielemente on liiga palju vähendatud, kommunaliteedid on liiga väikesed. Kokkuleppeliselt aga loetakse edaspidi negatiivsed omaväärtused nullideks, seda arvestades leitakse ka faktorite kirjeldusmäärad *Percent Var* ja summaarsed kirjeldusprotsendid *Cum Percent*.

Näeme tabelist, et maksimaalne võimalik faktorite arv on praegu 6, mõistlik faktorite arv (s.o ühest suuremate omaväärtuste arv) on 4. Veendume, et käesoleval juhul faktorid kirjeldavad faktorite poolt kirjeldatavast koguvarieeruvusest üle 95 %. Ometi on tunnuste üldhajuvuse kirjeldatuse protsent praegusel juhul madalam kui peakomponentide meetodi kasutamisel (võrdleme omaväärtuste suurusi tabelites joonistel 9.23 ja 9.36).

Faktormaatriks, mis on arvutatud klassikalisel meetodil, on esitatud joonisel 9.37. Võrreldes seda tabelit joonisel 9.25 esitatuga, on näha, et käesoleva andmestiku puhul on erinevus kahe maatriksi vahel tühine.

Factor Matrix

Variable/Factor	1	2	3	4
kaal	0.84157	-0.05522	-0.17735	0.14638
kasv	0.85905	-0.05689	-0.17987	-0.33149
king	0.96144	-0.10224	-0.17030	-0.11878
vanus	0.22700	0.08103	0.00144	0.84818
sugu	-0.87279	0.01220	0.13846	0.15560
abielus	-0.04427	0.58794	0.20308	0.07249
pere	0.09955	0.96282	-0.07916	-0.07147
tulu	0.46123	0.23834	0.74916	-0.24442
teadtoid	0.63498	0.15756	0.38803	0.50020
arvutoskus	-0.07822	0.21560	0.47985	-0.26667
eritulu	0.09561	-0.80580	0.52814	0.03775

Jn 9.37

Joonisel 9.38 on kujutatud klassikalise faktoranalüüsi tulemuste põhjal arvutatud üksiktunnuste kommunaliteedid.

Variable	Est Communality
kaal	0.76417
kasv	0.88343
king	0.97794
vanus	0.77751
sugu	0.80529
abielus	0.39413
pere	0.94831
tulu	0.89053
teadtoid	0.82879
arvutoskus	0.35397
eritulu	0.93881

Jn 9.38

Võrreldes neid joonisel 9.30 toodud kommunaliteetidega, näeme jällegi küllalt suurt sarnasust, kusjuures uuesti veendume selles, et klassikalise faktoranalüüsi korral on üksiktunnuste kirjeldatus keskmiselt madalam - 8 tunnuse puhul on kommunaliteet suurem komponentanalüüsi korral, 3 tunnuse puhul aga klassikalise faktoranalüüsi korral.

Faktorite pööramine toimub klassikalise faktoranalüüsi korral samuti, nagu peakomponentide faktoranalüüsi korralgi.

Märgime veel, et faktoranalüüsi protseduur võimaldab kõiki faktoranalüüsi vahetulemusi salvestada, kasutades aknas joonisel 9.27 esitatud täiendavate võimaluste loetelu.

9.2.8. Faktoranalüüs kovariatsioonimaatriksi põhjal. Faktoranalüüsi on kovariatsioonimaatriksi põhjal mõtet teha üksnes sel juhul, kui kõik uuritavad tunnused on mõõdetud samal skaalal (näiteks psühholoogilised testid). Kui aga erinevate tunnuste dispersioonid on väga erinevad, siis "neelavad" suurte väärtustega, tugevasti hajuvad tunnused ülejäänud alla ning viimaste analüüs jääb puudulikuks.

Demonstreerime seda olukorda, kusjuures lähteks kasutame kovariatsioonimaatriksit, millest osa on esitatud joonisel 9.2. Tunnuste loetelu on sama, mis ülejäänud näideteski.

Esimeses tulemuste tabelis (vt jn 9.39) on kommunaliteetide hinnangud mõttetud. Sama väljundi teises tabelis on aga kasulik märkida omaväärtuste jada järsku kahanemist: kaks esimest omaväärtust kirjeldavad üle 99 % tunnuste ühisest hajuvusest. Saadud faktormaatriksit kujutab joonis 9.40.

Variable	Communality	Factor	Eigenvalue	Percent Var	Cum Percent
kaal	-16.8697	1	32328.65986	74.2	74.2
kasv	-4.75949	2	10895.74651	25.0	99.3
king	0.51695	3	193.89012	.4	99.7
vanus	-11.0041	4	93.79406	.2	99.9
sugu	0.96894	5	18.78205	.0	100.0
abielus	0.86781	6	9.32619	.0	100.0
pere	0.89273	7	.94210	.0	100.0
tulu	-4100.34	8	.85972	.0	100.0
teadtoid	-27.7156	9	.17587	.0	100.0
arvutoskus	0.20847	10	.11233	.0	100.0
eritulu	-843.978	11	.02911	.0	100.0

Jn 9.39

Factor Matrix

Variable/Factor	1	2	3	4
kaal	1.83886	-0.53555	7.49913	4.80948
kasv	2.46182	-0.80977	3.48686	5.88924
king	1.05886	-0.13396	2.18519	2.04529
vanus	-0.81437	0.35400	5.52480	-3.48065
sugu	-0.16234	0.03957	-0.26666	-0.25397
abielus	0.08577	-0.21573	-0.02299	-0.14026
pere	0.15573	-1.18751	0.01618	-0.14397
tulu	175.220	-22.9368	-0.49846	0.00466
teadtoid	7.09810	-0.67795	9.47284	-4.42420
arvutoskus	0.50555	-0.10486	-0.37608	-0.00562
eritulu	39.5562	101.817	0.00148	0.05664

Jn 9.40

Näeme, et kaks esimest faktorit kirjeldavad põhiliselt tunnuseid **tulu** ja **eritulu**, millele vastavad ülejäänutest märksa suuremad dispersioonid. Alles kolmandasse ja neljandasse faktorisse ilmuvad tunnused **teadtoid, kaal, kasv, king ja vanus**. Tunnuste kirjeldatuse taseme id (absoluutarvudes) iseloomustab tabel jooniselt 9.41. Kõrvutades hinnatud kommunaliteete vastavate tunnuste dispersioonidega näeme, et tunnuste **tulu** ja **eritulu** dispersioonist on kirjeldatud 100 %, tunnuse **teadtoid** dispersioonist 96,5 %, **kaalust** 93,6 %, **kasvust** 91,0 %. Teiste tunnuste kirjeldatuse protsent ulatub 32,2 % ja 28,9 % vastavalt tunnuste **abielu** ja **arvutoskus** puhul.

Variable	Est Communality
kaal	83.0362
kasv	53.5576
king	10.0974
vanus	43.4268
sugu	0.16353
abielus	0.07410
pere	1.45542
tulu	31228.4
teadtoid	160.151
arvutoskus	0.40805
eritulu	11931.3

Jn 9.41

Järeldus - kuigi 4-faktoriline faktoranalüüs kirjeldab tunnuste koguhajuvusest 99,93 %, on tunnuste kirjeldatuse protsentide keskmine siiski vaid 78,8 %. Tunnuste tugevasti erinevate skaalade tõttu pole tulemus sisuliselt hästi interpreteeritav.

10. LINEAARSED MUDELID

§ 10.1. Mitmene regressioonanalüüs

10.1.1. Mitmese regressioonanalüüsi teoreetiline mudel. Paragrahvis 7.5 vaadeldud lineaarse regressioonanalüüsi mudel, mille puhul funktsioontunnuse Y prognoosimiseks kasutatakse argumenttunnuse X lineaarfunktsiooni

$$Y = a + bX + \varepsilon,$$

on lihtsalt üldistatav k argumenttunnuse juhule. Sellisel korral on prognoosivaks mudeliks avaldis

$$Y = b_0 + \sum_{i=1}^{k} b_i X_i + \varepsilon. \qquad (10.1)$$

Siin X_1, X_2, ..., X_k on argumenttunnused ja b_0, b_1, ..., b_k mudeli parameetrid, mis tuleb valimi põhjal määrata, ε aga on prognoosijääk (juhuslik viga), mille kohta eeldatakse, et ta on keskmiselt null: $E\varepsilon = 0$. Prognoosiks nimetatakse juhuslikku suurust $\hat{Y}$,

$$\hat{Y} = b_0 + \sum_{i=1}^{k} b_i X_i.$$

Mitmese regressioonimudeli parameetrid b_i ($i = 1, \ldots, k$) hinnatakse vähimruutude printsiibil valemist

$$b = (X'X)^{-1}X'Y, \qquad (10.2)$$

kus X on tsentreeritud argumenttunnustele vastav objekt-tunnus-maatriksi alammaatriks, Y - funktsioontunnuse väärtusi sisaldav veeruvektor, $b = (b_1, \ldots, b_k)$ on regressioonikordajate vektor ja sümbol ' tähistab maatriksi transponeerimist. Vabaliige b_0 määratakse seosest

$$b_0 = \mathbf{EY} - b'\mathbf{EX},$$

kus **EY** on funktsioontunnuse keskmine, **EX** argumenttunnuste keskmiste vektor ja b valemi (10.2) põhjal leitud regressioonikordajate vektor.

Metoodikat regressiooniparameetrite b_i hindamiseks (kõige sagedamini vähimruutude meetodil), nende ja ühtlasi mudeli (10.1) olulisuse kontrollimiseks, mudeli tõhususe (kirjeldatuse protsendi) kindlakstegemiseks ja veel mitmesuguse täiendava teabe saamiseks mudeli (10.1) kohta nimetatakse mitmeseks regressioonanalüüsiks. Mitmese regressioonanalüüsi (samuti nagu punktis 7.5.2 vaadeldud lihtsa regressioonanalüüsi) mudel on parameetrite b_i suhtes lineaarne. Kujul (10.1) esitatuna on ta lineaarne ka argumenttunnuste suhtes. Tuleb aga silmas pidada, et argumente on võimalik teisendada, näiteks võtta mudelisse lisaks liikmetele $\mathbf{X}_i$ ka liikmed $\mathbf{X}_i\mathbf{X}_j$ ja $\mathbf{X}_i^2$, mida formaalselt võib käsitleda kui muutujaid $\mathbf{X}_{k+1}, \ldots, \mathbf{X}_p$. Sellise mõttekäigu tagajärjel saame valemit (10.1) kasutada ka teatavate argumentide suhtes mittelineaarsete (nn lineariseeruvate) mudelite konstrueerimiseks.

10.1.2. Protseduur *Multiple Regression*. Eeldused ja tellimus. Mitmene regressioonanalüüs on süsteemis SG realiseeritud protseduurina *Multiple Regression* (kolmas allmenüüs *K. Regression Analysis*).

Protseduuri tellimuspaneeli kujutab joonis 10.1. Sellele tuleb trükkida kõigepealt funktsioontunnuse (*Dep. var.*) nimetus, seejärel (suvalises järjekorras) argumenttunnused (*Ind. vars.*) üksteise alla või järjest, kasutades eraldajat &. Nende arv on piiratud vabade ridadede arvuga tellimuspaneelil (ca 15), praktiliseks tööks on see arv täiesti küllaldane.

Meenutagem, et kõik tunnused - niihästi funktsioon- kui argumenttunnused - peavad olema arvulised. Nõutav on, et kõigi tunnuste pikkus oleks sama, kuid tühikud on võimalikud. Mõningase reservatsiooniga on kasutatavad ka järjestustunnused, kuid sel juhul tuleks ka mudelit tõlgendada eeskätt kvalitatiivselt. Eeldust kõigi tunnuste ühise mitmemõõtmelise normaaljaotuse kohta kasutatakse hüpoteeside

Multiple Regression

Dep. var.: tulu

Ind. vars.: pere
teadtoid
vanus

Weights:

Constant: Yes Vertical bars: No Conf. level: 95

Jn 10.1

kontrollimisel ja vahemikhindamisel, mitte aga punkthindamisel. Seetõttu on kirjeldav regressioonimudel lubatav ja korrektne igasuguse lähtejaotuse korral.

Mis puutub hüpoteeside kontrollimisse, siis võivad tulemused mõnevõrra küsitavaks osutuda tugevasti mittenormaalsete tunnuste korral (vt § 6.5). Kuid isegi niisuguse juhu jaoks ei ole SG-süsteemis olemas alternatiivseid võimalusi (selle põhjuseks on üldkasutatavate teoreetiliste lahenduste puudumine).

Tellimuspaneeli allservas on regressioonimudeli jaoks täiendavaid võimalusi lubavad tellimusparameetrid.

Kaalud *Weights* lubavad üksikvaatlustele omistada erinevaid osatähtsusi; kaalud sisestatakse tunnustega sama pikkusega vektorina.

On võimalik arvutada ka homogeense lineaarse mudeli parameetrid; see mudel erineb mudelist (10.1) vabaliikme b_0 puudumise poolest. Selline mudeli kuju saavutatakse, trükkides tellimuspaneeli aknasse *Constant: No* (vaikimisi arvutatakse mudel 10.1, millele vastab *Constant: Yes*). Kui töötluse esimesel sammul on osutunud, et b_0 on statistiliselt mitteoluline, siis on mõttekas mudeli täpsustamise käigus arvutada ka homogeenne mudel.

Parameeter *Vertical bars* on kasutatav graafikute kujundamisel ja *Conf. level* määrab kasutatava usaldus- ja olulisuse nivoo (vaikimisi väärtusteks on vastavalt 0,95 ja 0,05). Enamasti võib tellimuspaneeli allservas paiknevatele parameetritele jättagi vaikimisi väärtused.

10.1.3. Regressioonimudeli parameetrite hinnangud. Pärast regressioonanalüüsi protseduuri käivitamist ilmub ekraanile tabel regressiooniparameetrite hinnangutega.

Tabeli päises on märgitud ka funktsioontunnuse nimi (vt jn 10.2); päise täpne tõlge oleks "mudeli sobitamise tulemused".

Model fitting results for: tulu

Independent variable	coefficient	std. error	t-value	sig.level
CONSTANT	750.591162	120.751396	6.2160	0.0000
pere	18.979373	18.167201	1.0447	0.3066
teadtoid	12.30119	2.181498	5.6389	0.0000
vanus	-15.34866	3.780633	-4.0598	0.0005

R-SQ. (ADJ.) = 0.5528 SE= 118.169508 MAE= 97.665278 DurbWat= 1.231
Previously: 0.0000 0.000000 0.000000 0.000
28 observations fitted,forecast(s) computed for0 missing val.of dep.var.

Jn 10.2

Tabelis on viis veergu: argumenttunnuse nimi *Independent variable*, regressioonikordaja *coefficient*, regressioonikordaja hinnangu standardhälve *std. error*, t-statistiku väärtus selle regressioonikordaja olulisuse kontrollimiseks *t-value* ning selle t-testi olulisuse tõenäosus *sig. level*.

Tabeli esimene rida *CONSTANT* vastab vabaliikmele, järgmised read argumenttunnustele tellimusega määratud järjestuses.

Tabeli teise veeru põhjal on võimalik kirjutada välja mudel, näiteks käesoleval juhul on see järgmine:

tulu = 750,59 + 18,98×pere + 12,30×teadtoid - 15,35×vanus.

Prognoosimudelist järeldub näiteks, et 30-aastasel aspirandil, kellel on 3-liikmeline pere ja kirjas 5 teadustööd, on (meie andmetel) prognoositav kuusissetuleku suurus

$$750{,}59 + 18{,}98 \cdot 3 + 12{,}30 \cdot 5 - 15{,}35 \cdot 30 = 408{,}53.$$

Kasutasime siin j-nda objekti jaoks prognoosi $\tilde{y}_j$ arvutamiseks valemist (10.1) järelduvat valemit

$$\tilde{y}_j = b_0 + \sum_{i=1}^{k} b_i x_{ij}, \qquad (10.1')$$

kus x_{ij} on j-nda objekti i-nda tunnuse väärtus ja $\tilde{y}_j$ vastav prognoos.

Tabeli viimasest veerust aga näeme, et argument pere ei ole statistiliselt oluline.

10.1.4. Regressioonimudelit iseloomustavad karakteristikud: mitmene korrelatsioonikordaja jt. Tabeli alla on trükitud mõned täiendavad suurused. Neist esimene on parandatud determinatsioonikordaja (mitmese korrelatsioonikordaja ruut) *R-SQ. (ADJ.)*.

Mitmese korrelatsioonikordaja ruut R^2 näitab mudeli poolt kirjeldatud hajuvuse SS_1 suhet funktsioontunnuse koguhajuvusse SS,

$$R^2 = \frac{SS_1}{SS}, \qquad (10.3)$$

kus

$$SS = \sum_{j=1}^{n} (y_j - \bar{y})^2,$$

$$SS_1 = \sum_{j=1}^{n} (\tilde{y}_j - \bar{y})^2.$$

Lineaarse regressiooni puhul kehtib võrdus

$$SS = SS_0 + SS_1,$$

kus

$$SS_0 = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

on jääkhajuvus, mida regressioonimudel ei selgita.

Valemiga (10.3) esitatud determinatsioonikordaja on aga nihkega hinnanguks vastavale üldkogumi determinatsioonikordajale, kusjuures nihe on seda suurem, mida väiksem on n ja

suurem k. Nihke parandamise tulemusena saadakse nn parandatud determinatsiooonikordaja $\hat{R}^2$,

$$\hat{R}^2 = 1 - (1 - R^2)\,\frac{n-1}{n-k-1}\,. \qquad (10.4)$$

Olgu märgitud, et ka parandatud determinatsioonikordaja väärtus peab alati mittenegatiivne olema, ning kui see arvutusprotsessi tulemusena osutub negatiivseks, tuleb teha otsustus $\hat{R}^2 = 0$.

Nagu järeldub definitsioonist, iseloomustab determinatsioonikordaja mudeli kirjeldavat toimet, st seda, kui suure osa funktsioontunnuse koguhajuvusest mudel kirjeldab; alati kehtib seos $0 \le R^2 \le 1$.

Ruutjuurt determinatsioonikordajast nimetatakse mitmeseks korrelatsioonikordajaks R, see on lineaarne korrelatsioonikordaja prognoosi $\tilde{Y} = b_0 + \sum_{i=1}^{k} b_i X_i$ ja funktsioontunnuse Y vahel:

$$R = r(Y,\tilde{Y}).$$

Järgmine suurus tabeli all on tähistatud sümboliga *SE* (*standard error of estimate*, ka *residual standard error*), see on mudeli poolt kirjeldamata hajuvuse osa (jääkhajuvuse) põhjal arvutatud standardhälve:

$$SE = \sqrt{\frac{SS_0}{n-k-1}}\,. \qquad (10.5)$$

Järgneb keskmine absoluutviga *MAE* (*mean absolute error*),

$$MAE = \frac{\sum_{j=1}^{n} |y_j - \tilde{y}_j|}{n}\,.$$

Järgmiseks on tabeli alla trükitud Durbin-Watsoni statistik d (*DurbWat*), mida kasutatakse kontrollimaks hüpoteesi jääkide jada korreleerituse kohta.

Tähistame j-nda objekti prognoosimisel tekkinud prognoosijääki sümboliga ε_j ($j = 1, \ldots, n$),

$$\varepsilon_j = y_j - \tilde{y}_j, \qquad (10.6)$$

ja formuleerime hüpoteesi jada sõltuvuse kohta mudeli

$$\varepsilon_j = \rho\varepsilon_{j-1} + \delta_j$$

kohaselt. Siis saame

$$H_0: \rho = 0,$$
$$H_1: \rho \neq 0,$$

kus nullhüpoteesi puhul prognoosijääkidel puudub nn seriaalne sõltuvus (mida mõõdab seriaalse sõltuvuse kordaja ρ).

Hüpoteesipaari kontrollimiseks tuletatud statistiku

$$d = \frac{\sum_{j=2}^{n} (\varepsilon_j - \varepsilon_{j-1})^2}{\sum_{j=2}^{n} \varepsilon_j^2}$$

jaoks on kriitilised väärtused küll tabuleeritud (vt nt [2], lk 159-178), kuid tabelid ei ole üldlevinud.

Durbin-Watsoni statistiku väikesed väärtused (märgatavalt alla 1) kõnelevad seriaalse sõltuvuse poolt. Väärtused üle 1,5 lubavad oletada H_0 kehtivust, kuid eriti suurte väärtuste korral tuleb oletada spetsiaalset tüüpi võnkumisi, seega vaatlused on jälle sõltuvad.

Eelviimane rida tabelis, mis algab sõnaga *Previously*, on käesoleva protseduuri korral tühi (sisaldab väärtusi 0, millel ei ole mingit sisulist tähendust).

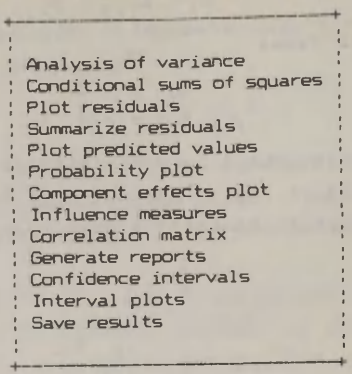
Viimases reas märgitakse, mitme punkti jaoks on prognoosid arvutatud (.. *observations fitted*).

Prognoosid arvutatakse nende punktide jaoks, millel on kõigi argumenttunnuste väärtused olemas, sõltumatult sellest, kas funktsioontunnuse väärtus on teada või mitte. Viimased märgitakse alumises reas eraldi (*forecast(s) computed for ... missing val. of dep. var.*).

10.1.5. Mudeli olulisus. Dispersioonanalüüsi tabel

Pärast põhitulemuste (jn 10.2) vaatamist on võimalik kutsuda võtme F5 abil ekraanile aken täiendavate tulemuste loeteluga (vt jn 10.3). Vaatleme need võimalused ükshaaval läbi. Mõned allprotseduurid nimetatud loetelust võimaldavad veel omakorda rea valikuid.

Esimesel real on dispersioonanalüüsi tabel *Analysis of variance*, mille valimise tulemusena saame ekraanile joonisel



Jn 10.3

10.4 kujutatud tabeli, mis vastab täielikult punktis 7.4.4 esitatud kirjeldusele. Tabeli all esitatud statistikud on kirjeldatud eelmises punktis, need on determinatsioonikordaja R^2 (*R-squared*), vt valem (10.3), parandatud determinatsioonikordaja $\hat{R}^2$ (*R-squared (Adj. for d.f.)*), vt valem (10.4), SE (*Stnd. error of est.*), vt valem (10.5) ja Durbin-Watsoni statistik (*Durbin-Watson statistic*).

Analysis of Variance for the Full Regression

Source	Sum of Squares	DF	Mean Square	F-Ratio	P-value
Model	508030.	3	169343.	12.1271	.0000
Error	335137.	24	13964.0		
Total (Corr.)	843167.	27			

R-squared = 0.602526 Stnd. error of est. = 118.17
R-squared (Adj. for d.f.) = 0.552842 Durbin-Watson statistic = 1.23122

Jn 10.4

Tabeli viimases veerus esitatud F-statistiku olulisuse tõenäosus (*P-value*) annab informatsiooni selle kohta, kas mudel tervikuna on oluline. Nimelt kontrollitakse hüpoteesipaari

$$H_0: b_1 = b_2 = \ldots = b_k = 0,$$
$$H_1: \text{leidub } i, \text{ nii et } b_i \neq 0.$$

Kui $p \leq \alpha$, siis võetakse vastu hüpotees H_1, mis tähendab mudeli olulisust. Joonisel 10.4 toodud näites on mudel tervikuna väga veenvalt oluline (hoolimata sellest, et üks liige oli joonisel 10.2 esitatud tabeli andmetel mitteoluline).

Mudeli kui terviku olulisuseks piisab sellest, kui oluline on vähemalt üks argument (välja arvatud vabaliige - kui ainult vabaliige on oluline, pole mudel oluline, sest statistilise mudeli mõistesse kuulub juhusest sõltuva argumendi sisaldumine selles).

Järgmine rida aknas joonisel 10.3 - tinglikud ruutude summad (*Conditional sums of squares*) - võimaldab üksikute tunnuste mõju detailsemalt analüüsida, vt jn 10.5. Tabeli päisesse on trükitud "Täiendav dispersioonanalüüs" (*Further ANOVA for Variables in the Order Fitted*, täpne tõlge oleks "Muutujate täiendav dispersioonanalüüs vastavalt nende mudelisse sobitamise järjekorrale). Tabeli struktuur vastab standardsele dispersioonanalüüsi tabelile.

Further ANOVA for Variables in the Order Fitted

Source	Sum of Squares	DF	Mean Sq.	F-Ratio	P-value
pere	49902.737	1	49902.74	3.57	.0708
teadtoid	227971.178	1	227971.18	16.33	.0005
vanus	230156.264	1	230156.26	16.48	.0005
Model	508030.179	3			

Jn 10.5

Tabeli igas reas on üks argument, argumendid on paigutatud tellimuse järjekorras. Ruutude summa (*Sum of Squares*) veerus on q-ndas reas märgitud suuru $SS^{(q)}$, mis arvutatakse järgmisel viisil:

$$SS^{(q)} = SS_1(1, \ldots, q) - SS_1(1, \ldots, q-1).$$

Seega on $SS^{(q)}$ q-nda argumenttunnuse poolt põhjustatud täiendav jääkhajuvuse kahanemine. Esimeses reas on see suurus $SS_1(1)$, st ainult esimest argumenti sisaldava mudeli poolt kirjeldatud osa funktsioontunnuse hajuvusest.

Kõik vabadusastmete arvud (*DF*) võrduvad ühega, kõik keskruudud (*Mean Sq.*) vastavate ruutudesummadega $SS^{(q)}$.

Veerus *F-Ratio* (F-suhe) paiknevad statistikud saame $SS^{(q)}$ jagamisel dispersioonanalüüsi tabelis (vt jn 10.4) reas *Error* ja veerus *Mean Square* paikneva suurusega - see on mudeliveale vastav dispersioonikomponent. Olulisuse tõenäosus on arvutatud F-jaotuse järgi, kasutades vabadusastmete arve 1 ja n-q-1.

Tabeli all on kogu mudelit tervikuna iseloomustav suurus SS_1 ($= \sum_{q=1}^{k} SS^{(q)}$) ning vabadusastmete arv k.

10.1.6. Argumentide valik regressioonimudelisse. Üheks põhilise tähtsusega küsimuseks regressioonanalüüsi juures on - kuidas valida mudelisse õige arv õigeid argumenttunnuseid?

Valides mudelisse liiga vähe argumente, võib juhtuda, et kirjeldatuse tase (mida mõõdab determinatsioonikordaja R^2) jääb põhjendamatult madalaks.

Valides mudelisse liiga palju argumente, võib juhtuda, et kuigi R^2 suureneb, väheneb mudeli olulisust mõõtev F-statistiku väärtus sedavõrd, et mudel osutub mitteoluliseks.

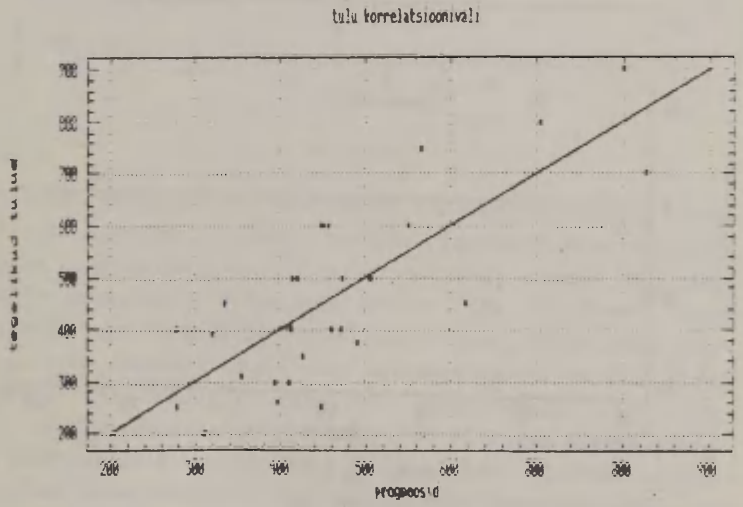
Teatava määrani võib argumentide mõju üle otsustada tunnustevahelise korrelatsioonimaatriksi põhjal. Need argumenttunnused, mis on funktsioontunnusega tugevasti korreleeritud, omavad kaheldamatult ka mudelis olulist kaalu. Asja komplitseerib aga tunnuste koosmõju. Võib juhtuda, et juhul kui mitu argumenttunnust on funktsioontunnusega tugevasti korreleeritud, kuid ka nende omavaheline korrelatsioon on kõrge, ei ole mitme argumendiga mudel märkimisväärselt parem ühe argumendiga mudelist, sest kõigi argumentide toime on üsna ühesugune ja iga argumendi täiendav toime üht argumenti sisaldavale mudelile on tühine. Sellises olukorras võib mudelit parandada

- juba sisestatud argumentidega nõrgalt, kuid funktsioontunnusega tugevasti seotud argumendi lisamine mudelile,
- mõnikord ka ülejäänud argumentidega üsna tugevasti, kuid funktsioontunnusega nõrgalt (või ülejäänud tunnustega vastassuunaliselt) seotud argumendi lisamine,
- vahel aga lihtsalt osa argumentide mudelist välja jätmine.

Üldiselt võib lugeda igati põhjendatuks ainult statistiliselt oluliste (vt nt jn 10.2, viimane veerg) argumentide mudelisse jätmise.

Argumentide valimist mudelisse hõlbustab joonistel 10.2 ja 10.5 esitatud tabelites väljastatav informatsioon, kuid selleks on välja töötatud ka spetsiaalseid nn samm-protseduure (*Step-wise*). Igal juhul tuleb meeles pidada, et vanast mudelist ei saa uut mitte lihtsalt argumentide mahatõmbamisega, vaid kogu arvutuskäik tuleb uue argumenttunnuste komplektiga uuesti teha.

Mudeli sobivust võimaldab visuaalselt hinnata prognoosi ja prognoositava tunnuse korrelatsiooniväli. See korrelatsiooniväli saadakse, valides joonisel 10.3 kujutatud aknast 5. rea. Väljale on hõlbustuseks kantud ka diagonaalsirge - "ideaalse vastavuse joon" (vt jn 10.6). Siin on x-teljel prognoosi $\hat{Y}$, y-teljel funktsioontunnuse Y vaadeldud/ /mõõdetud väärtused. Neil punktidel, mis paiknevad sirgel, langevad tunnuse väärtus ja prognoos ühte.



Jn 10.6

10.1.7. Prognoosijääkide analüüs. Kui olemasoleva regressioonimudeli ja selle üksikliikmete analüüs lubab teha

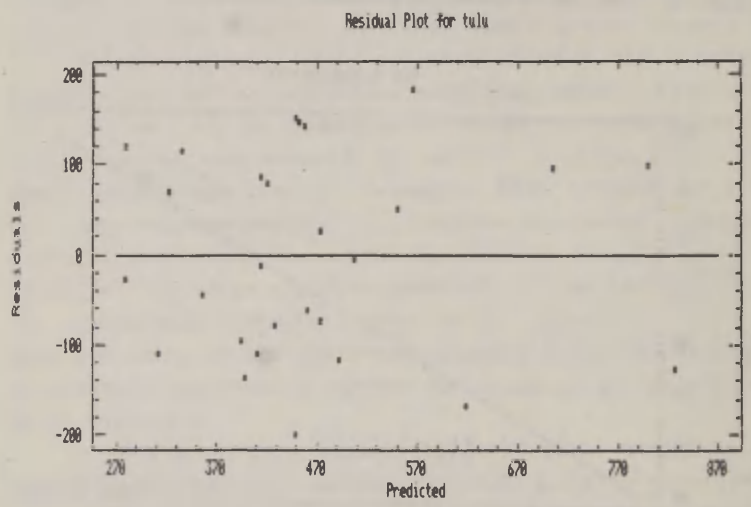
järeldusi võimalike liigsete argumentide kohta, siis prognoosijääkide $\tilde{y}_j$, vt (10.1'), analüüs on vajalik eeskätt selleks, et teha oletusi selle kohta, kas regressioonimudelisse tuleks veel täiendavaid liikmeid lisada.

Visuaalselt on võimalik hüpoteese genereerida vastavate graafikute (korrelatsiooniväljade) abil, millele vastab aknas joonisel 10.3 kolmas rida *Plot residuals*. Selle allprotseduuri valimise järel ilmub ekraanile aken täiendava küsimusega graafiku x-telje (argumendi) kohta (vt jn 10.7).

```
Plot residuals vs Predicted values, Index, or another Variable (P/I/V):
```

Jn 10.7

Esimene valik *P* - prognoositud väärtused (*Predicted values*) - annab joonisel 10.8 esitatud graafiku, kus x-teljel on prognoosid $\tilde{y}_j$, y-teljel - prognoosijäägid e_j, vt (10.8).

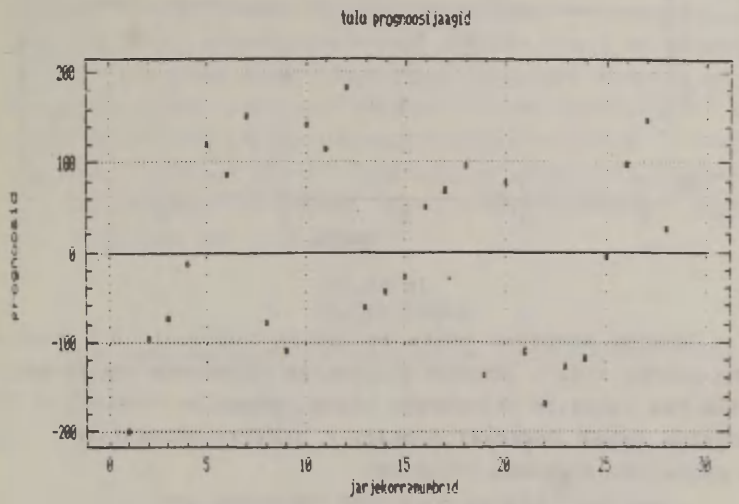


Jn 10.8

Vähimruutude meetod tagab prognooside ja prognoosijääkide mittekorreleerituse; juhul kui tunnuste ühisjaotuseks on normaaljaotus, järeldub sellest ka sõltumatus. Kui aga

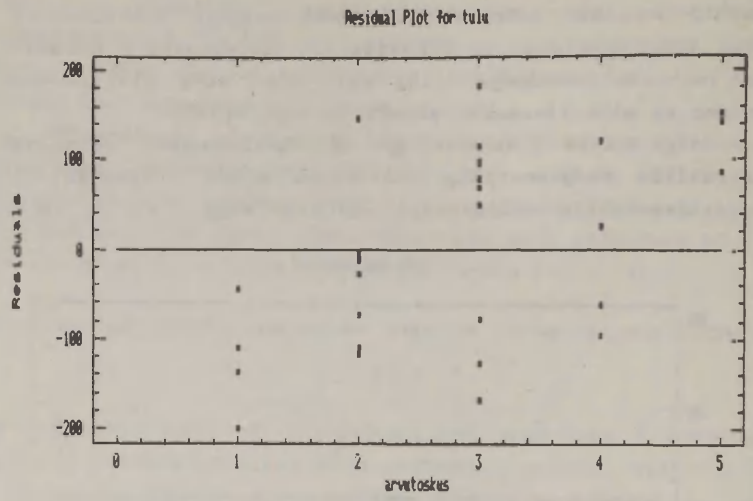
punktid joonisel tunduvad paiknevat selgelt mittejuhuslikult, tuleb oletada, et tegemist on lähtetunnuste normaalsest erineva jaotusega, ning sel juhul võib otstarbekaks osutuda ka mittelineaarse mudeli konstrueerimine.

Teine valik *I* (*Index*, st järjekorranumber) võimaldab kontrollida prognoosijäägi sõltuvust ajast (oletades, et järjekorranumbrid väljendavad mõõtmise aega), vt jn 10.9.



Jn 10.9

Kolmandat võimalust *V* - muu tunnus (*another Variable*) - saab kasutada suvalise täiendava tunnuse oletatava mõju analüüsimiseks: kui graafik näitab prognoosijäägi ilmset sõltuvust mingist tunnusest, tuleb see tunnus (kas lineaarsel kujul või mõni tema funktsioon) mudelisse lisada. Näite selle kohta esitame joonisel 10.10, kus argumendiks on valitud arvutioskus. See tunnus on diskreetne väärtustega 1...5 (hinded), kusjuures graafikult ilmneb, et väiksema arvutioskusega aspirantidele vastab üldiselt negatiivne, suurema arvutioskusega aspirantidele aga positiivne prognoosijääk. See tähelepanek lubab nimetatud tunnuse edaspidi mudelisse lisada.



Jn 10.10

Jääkide analüüsi hulka kuulub ka joonisel 10.3 toodud akna neljas rida - jääkide statistika (*Summarize residuals*). Selle rea valimise tulemusena ilmub ekraanile joonisel 10.11 esitatud tabel *Residual Summary* - jääkide statistikud, kus on esitatud järgmised näitajad:

- vaatluste arv ja puuduvate väärtuste arv;
- jääkide keskmine *Residual average*, mis vähimruutude meetodi puhul peab olema praktiliselt võrdne nulliga;
- jääkdispersioon *Residual variance*, mis on arvutatud eeskirjaga $SS_o/(n-k-1)$ (vt ka valem (10.5)) ning mida on kasutatud dispersioonanalüüsi tabelites joonistel 10.4 ja 10.5;

Residual Summary

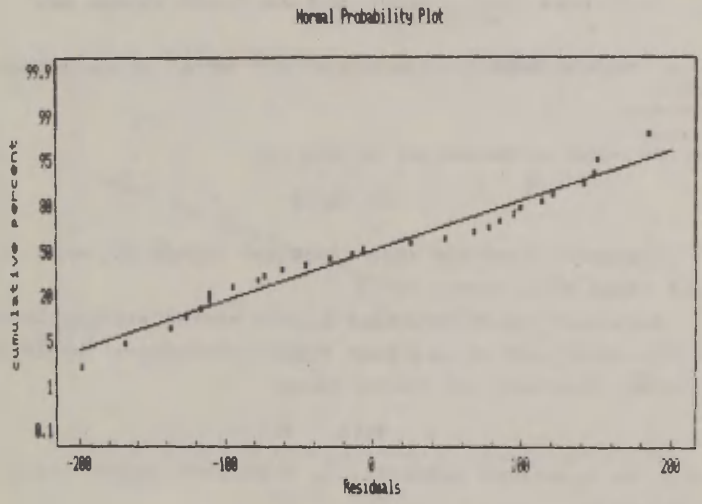
```
Number of observations = 28  (0 missing values excluded)
Residual average = 5.48133E-13
Residual variance = 13964
Residual standard error = 118.17

Coeff. of skewness = -0.024651   standardized value = -0.0532523
Coeff. of kurtosis = -1.30083    standardized value = -1.40506

Durbin-Watson statistic = 1.23122
```

Jn 10.11

- jääkstandardhälve *Residual standard error*, ruutjuur eelmisest statistikust, mis esineb ka joonisel 10.2 esitatud tabelis (suurus *SE*) ning joonisel 10.4 esitatud tabelis (*Std. error of est*$_i$).

Edasi on toodud jääkide e_j põhjal arvutatud asümmeetriakordaja (*Coeff. of skewness*) ja ekstsess (*Coeff. of kurtosis*) ning nende standardiseeritud väärtused (*standardized value*), vt ka § 6.2, mis võimaldavad kontrollida hüpoteesi jääkide jaotuse normaalsusest. Meenutame, et normaaljaotusest oluliselt erineva jaotusega jääkide puhul on tõenäoline, et leidub parimast lineaarsest mudelist sobivam mittelineaarne mudel, kusjuures jääkide korrelatsiooniväljal ilmnevad trendid (punktide süstemaatilised kuhjumid) võimaldavad teha hüpoteese selle mittelineaarse mudeli kuju kohta.

Kõige lõpuks on uuesti esitatud Durbin-Watsoni statistiku d väärtus (vt p 10.1.2).



Jn 10.12

Teine võimalus otsustada visuaalselt jääkide jaotuse normaalsuse üle on nn tõenäosuspaberi (*normal probability plot*) kasutamine (vt § 6.5), milleks tuleb valida aknast

joonisel 10.3 kuues rida (*Probability plot*). Joonisel 10.12 on esitatud näitena vaadeldud ülesande jääkide jaotusfunktsioon normaaljaotuse jaotusfunktsiooni järgi skaleeritud graafikul. Graafikult ilmneb, et empiirilise jaotusfunktsiooni alumine vasakpoolne ots on allpool, ülemine parempoolne ots ülalpool standardset sirget, kusjuures keskmises osas on S-kujuline laine. See tunnistab, et jääkide jaotus on pigem ühtlane, "lõigatud sabadega", suuri jääke esineb harvemini kui normaaljaotuse puhul. Muide, sama järelduseni viib ekstsessi negatiivsus. Siiski pole jääkide jaotus silmapaistvalt erinev normaalsest.

10.1.8. Üksikargumentide mõjude graafikud. Üksikargumentide mõjude graafikud saame, valides aknast joonisel 10.3 seitsmenda rea *Component effects plot*.

Pärast valikut ilmub ekraanile joonisel 10.13 kujutatud täiendava tellimuse paneel, kus on märgitud argumentide loetelu (tellimuse järjestuses), millest tuleb valida üks.

```
You may create a component-plus-residual plot for any of the following:
 1.pere
 2.teadtoid
 3.vanus
Enter the number of the desired variable (1):
```

Jn 10.13

Argumendi **teadtoid** mõju kirjeldab joonis 10.14 ja argumendi **vanus** mõju joonis 10.15.

Joonistel on horisontaalteljeks vastav argumenttunnus, vertikaaltelg aga on esitatud funktsioontunnuse (prognoosi) ühikutes. Joonisele on kantud sirge

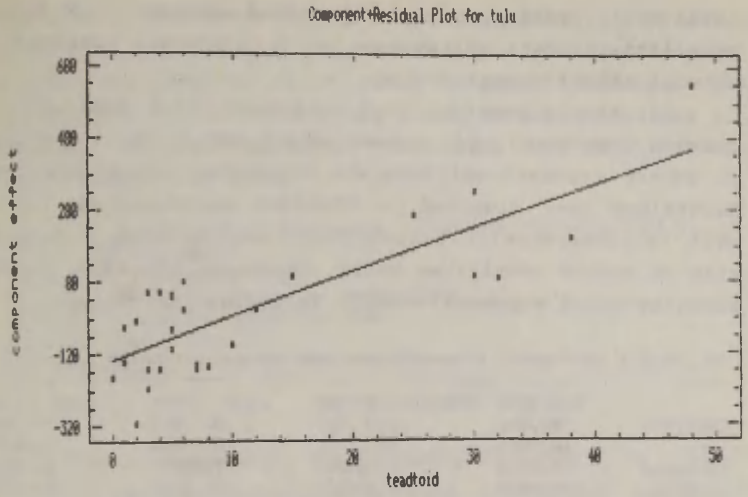
$$y = b_i(x_i - \bar{x}_i),$$

kus x_i on vaadeldav argument, b_i - sellele vastav regressioonikordaja.

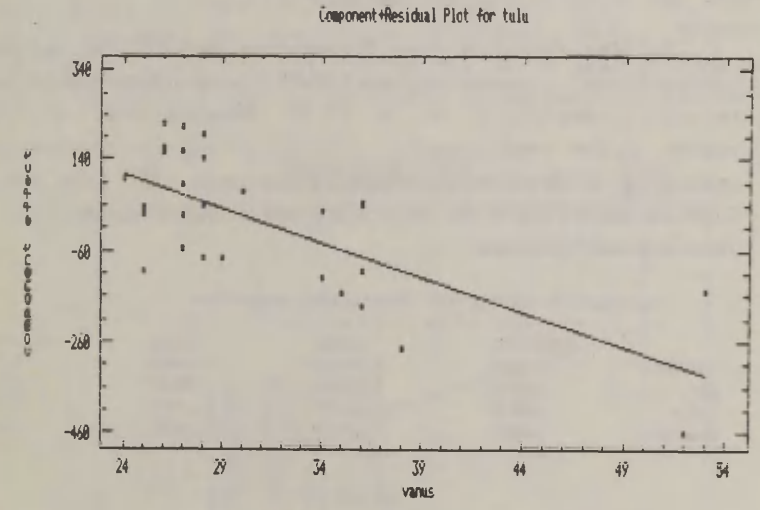
Graafikul on punktide ordinaatide väärtusteks suurus

$$e_j + b_i(x_{ij} - \bar{x}_i),$$

see graafik abistab üksikute argumentide suhtelise toime selgitamisel (vt [5]).



Jn 10.14



Jn 10.15

10.1.9. Regressiooniparameetrite vahemikhinnangud ja korrelatsioonimaatriks. Kuivõrd regressioonikordajad on hin-

natud valimi põhjal, on nad juhuslikud suurused ning tulenevalt lähtetunnuste sõltuvusest on ka hinnangud suuremal või vähemal määral korreleeritud.

Kui aknast joonisel 10.3 valida 11. rida *Confidence intervals* (usalduspiirid), saame tabeli (vt jn 10.16), millesse on kantud regressioonikordajate hinnangud, hinnangute standardhälbed ning alumised ja ülemised usalduspiirid (vastavalt tellimuspaneelil fikseeritud usaldusnivoole). Esimeses reas on andmed vabaliikme kohta, järgnevad ülejäänud regressioonikordajad argumenttunnuste tellimise järjekorras.

95 percent confidence intervals for coefficient estimates

	Estimate	Standard error	Lower Limit	Upper Limit
CONSTANT	750.591	120.751	501.313	999.869
pere	18.9794	18.1672	-18.5248	56.4836
teadtoid	12.3012	2.18150	7.79773	16.8047
vanus	-15.3487	3.78063	-23.1534	-7.54396

Jn 10.16

Valides aknast 9. rea *Correlation matrix* - korrelatsioonimaatriks -, saame regressioonikordajate hinnangute korrelatsioonimaatriksi, vt jn 10.17. Esimene rida ja veerg vastab jälle vabaliikmele, ülejäänud regressioonikordajad paiknevad tellimuses esitatud järjestuses. NB! Seda korrelatsioonimaatriksit ei tohi segi ajada lähtetunnuste korrelatsioonimaatriksiga!

Correlation matrix for coefficient estimates

	CONSTANT	pere	vanus	teadtoid
CONSTANT	1.0000	-.4478	-.8806	.4219
pere	-.4478	1.0000	.0507	-.1538
vanus	-.8806	.0507	1.0000	-.5740
teadtoid	.4219	-.1538	-.5740	1.0000

Jn 10.17

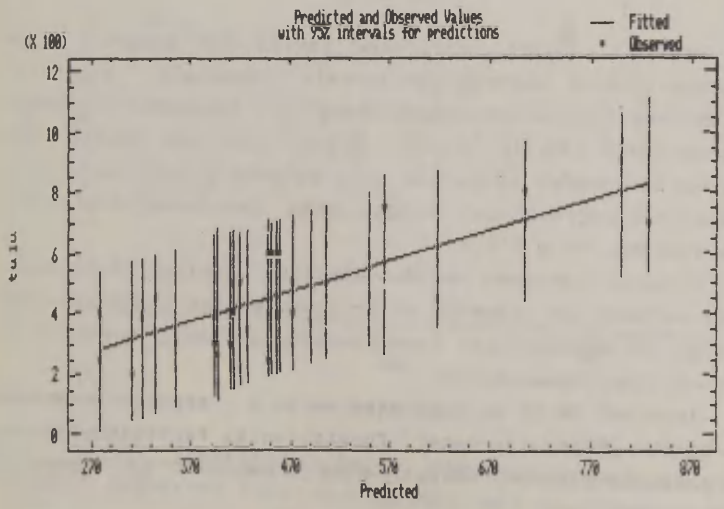
10.1.10. Üksikväärtuste prognooside ja regressioonijoone usalduspiirkond. Et ka üksikväärtuste prognoosid $\tilde{y}_j$ (vt 10.1'), mis sõltuvad valimi põhjal määratud regressioonikordajatest b_i, on juhuslikud, on ka nende jaoks võimalik arvu-

tada hajuvuse karakteristikud - standardhälbed $s_{\tilde{y}_j}$ ja samuti ka usalduspiirid. Valides aknast joonisel 10.3 usalduspiiride graafikud *Interval plots*, saame kõigepealt täiendava tellimuspaneeli (vt jn 10.18), mis annab jällegi võimaluse mitmesuguste hajuvusdiagrammide saamiseks niihästi usaldusintervallidega kui ilma.

```
Plot data vs Predicted values, Index, or another Variable (P/I/V): V
You may plot the data:
  1. without confidence limits
  2. with confidence limits for predictions
  3. with confidence limits for means
Enter your choice (1): 2
Enter the desired variable: arvutoskus
```

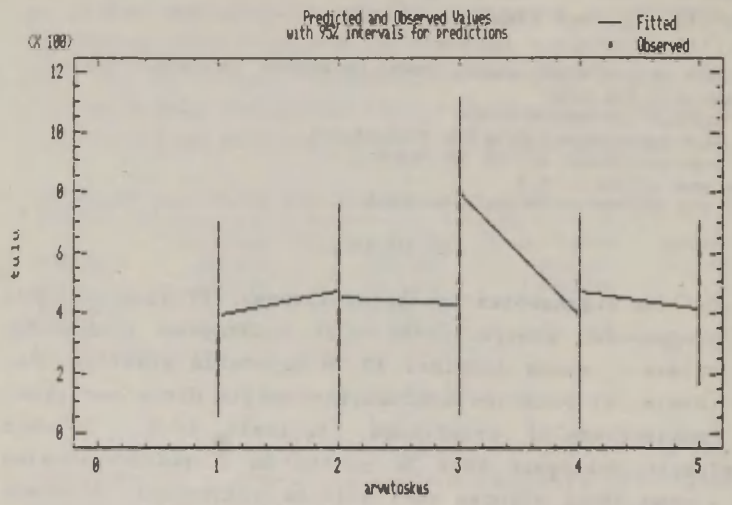
Jn 10.18

Valides argumendiks funktsioontunnuse (*P*) ning variandi 2 - prognooside usalduspiirid *with confidence limits for predictions* -, saame joonisel 10.19 kujutatud graafiku. Paneme tähele, et punktide konfiguratsioonilt ühtib see graafik (ootuspäraselt) graafikuga jooniselt 10.6. Täiesti juhuslikult paiknevad kõik 28 punkti 95 % usaldusvahemiku sees - sama hästi võinuks mõni olla ka väljaspool. Joonisel



Jn 10.19

10.20 on argumendiks valitud tunnus **arvutoskus**. Kuna sel tunnusel on ainult 5 väärtust, on ka usaldusintervalle kujutavaid lõike viis, igaühel neist aga paikneb terve hulk vaatlusi. Prognoose ühendab murdjoon.

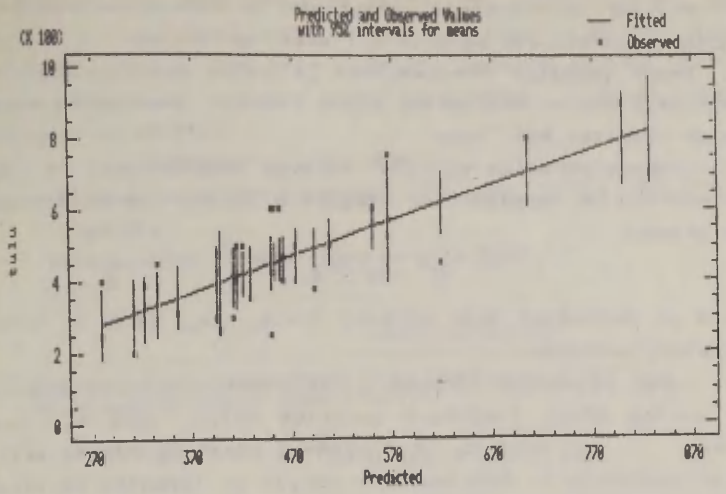


Jn 10.20

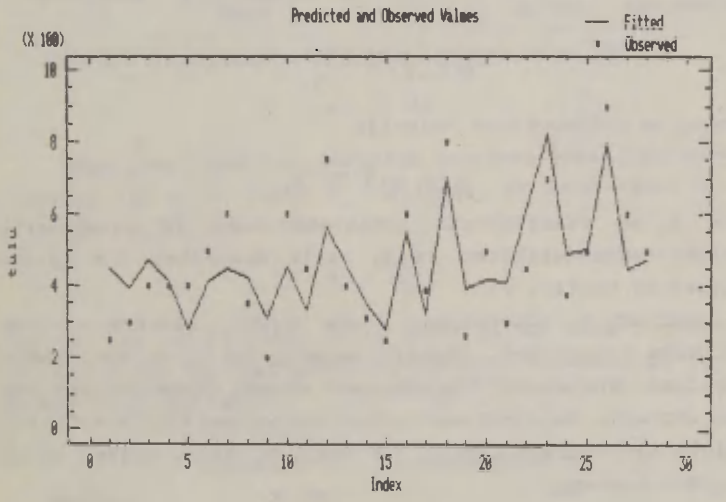
Valik 3 - *with confidence limits for means* - annab usalduspiirkonna regressioonijoonele (täpsemalt - pinnale), mis kujutab funktsioontunnuse (tinglike) keskmiste sõltuvust argumentidest (vt jn 10.21). Asjaolu, et osa punkte jääb usalduspiirkonnast välja, on siin täiesti ootuspärane (meenutame lihtregressiooni sirge jaoks konstrueeritud usalduspiirkonda, vt p 7.5.4).

Erineva ulatusega usaldusvahemikud joonisel 10.21 tulenevad sellest, et tegemist on projektsiooniga 4-mõõtmelisest ruumist (3 argumenti ja funktsioontunnus) ning seetõttu on pilt oluliselt moonutatud.

Joonisel 10.22 on kujutatud valik 1 - argumenttunnuseks on objekti järjekorranumber, funktsiooniks funktsioontunnus, usaldusintervalli ei märgita ning prognoosid on ühendatud murdjoonega.



Jn 10.21



Jn 10.22

10.1.11. Iseärased punktid. Regressioonikordajate hinnanguid mõjutavad tugevasti punktid, mis vaadeldavas (k+1)-

-mõõtmelises tunnusruumis (argument- ja funktsioontunnused) ülejäänud punktidest märgatavalt eemal paiknevad.

Nende punktide avastamiseks ja välja selgitamiseks on kasutusele võetud statistik, mille nimetus *leverage* on eesti keelde tõlgitav kui "kang".

Rühmade/punktide vahelisi kaugusi tunnusruumis on loomulik mõõta nn Mahalanobise kauguse D järgi, mis on arvutatav seosest

$$D^2 = d'S^{-1}d,$$

kus d on keskmiste vahe vektor, $d = \bar{x}_1 - \bar{x}_2$, ja S on kovariatsioonimaatriks.

Loeme esimeseks rühmaks j-nda punkti, teiseks rühmaks aga valimi kõigi ülejäänud punktide hulga, siis $\bar{x}_1 = x_j = (x_{1j}, \ldots, x_{kj})$ ning $\bar{x}_2$ on ülejäänud punktide põhjal arvutatud keskpunkt $\bar{\bar{x}}$. Mahalanobise kaugus on tunnuste hajuvuse suhtes normeeritud.

Punkti x_j Mahalanobise kaugus keskpunktini $\bar{\bar{x}}$ on esitatav seosega

$$D(x_j, \bar{\bar{x}}) = \frac{nh_j}{1-h_j}.$$

Siin h_j on defineeritud valemiga

$$h_j = x_i(X'X)^{-1}x_j,$$

kus X on vaadeldavaid tunnuseid sisaldav tsentreeritud objekt-tunnus-maatriks ja x_j selle maatriksi j-s (j-ndale objektile vastav) rida.

Suurust h_j nimetatakse j-nda objekti omapära karakteristikuks (*leverage*). Omapära karakteristik on suur nendel punktidel, mis asuvad ülejäänutest kaugel (Mahalanobise kauguse mõttes). Empiiriliselt (kuid ka teoreetiliste arutelude põhjal) on kindlaks tehtud, et punktid, mille korral kehtib üks võrratustest

$$h_j > 0{,}2; \quad h_j \geq \frac{2k}{n};$$

pakuvad huvi iseärastena.

Iseäraste punktide analüüsi saab tellida, valides aknast joonisel 10.3 rea *Influence measures* (mõju mõõdud).

Esimene vastus tellimusele on paneel (vt jn 10.23). Kui soovitakse ka iseäraste punktide loetelu, siis tuleb teha valik *Y* (*Yes*). Saadud iseäraste punktide tabelit kujutab joonis 10.24, millel on esitatud punktide loetelu ja neid iseloomustavad parameetrid:

järjekorranumber j (*Obs. Number*),
standardiseeritud prognoosijääk e_j/s_{e_j} (*Stnd. Residual*),
Leverage,
Mahalanobise kaugus (*Mahalanobis Dist.*),
DFITS.

Influence Measures

Number of flagged observations (high leverage or DFITS) = 1
Do you want to display them? (Y/N):

Jn 10.23

Flagged Observations for tulu

Obs. Number	Stnd. Residual	Leverage	Mahalanobis Dist.	DFITS
12	2.34688	0.47444	22.5081	2.22982

Number of flagged observations (high leverage or DFITS) = 1

Jn 10.24

Käesoleval juhul on loetelus üksainus punkt (järjekorranumbriga 12), mille puhul $h_j = 0{,}47444$. Et meie näite korral $\frac{2k}{h} = \frac{8}{28} = 0{,}286$, kehtib võrratus

$$h_j > \frac{2k}{h} .$$

Tegemist on indiviidiga, kelle vanus on 53, pere 2, tulu 750 ja teadustöid 48. Erinevusi vastavatest keskmistest iseloomustab järgmine tabel:

	keskmine $\bar{x}$	standardhälve s	$(x_j - \bar{x})/s$
vanus	31,18	7.35	2,97
pere	2,86	1.27	0,68
teadtoid	11,3	12.9	2,84
tulu	465,5	176,7	1,61

Niisuguste ülejäänutest tunduvalt erinevate objektide puhul on kahtlemata suurem võimalus, et tegemist on veaga; kui mitte, tuleb kontrollida sobivust antud massiivi (ka käesoleva näite puhul on tegemist eriolukorraga - suhteliselt eakas kraaditaotleja noorte aspirantide seas) ning arvestada selle mõju mudelile (vrdl vastava protseduuriga lihtregressooni korral, p 7.5.5).

Kui sisuline analüüs näitab vastava punkti kuuluvust kogumisse, ei ole põhjust seda välja jätta.

Punktide iseärasuskordajate tabel on esitatud joonisel 10.25.

```
Variable: WORKAREA.LEVERAGE   (length = 28)
-------------------------------------------------------
( 1) 0.0665581    (19) 0.358773
( 2) 0.188919     (20) 0.0646322
( 3) 0.0633429    (21) 0.117081
( 4) 0.145956     (22) 0.131822
( 5) 0.160597     (23) 0.384801
( 6) 0.072648     (24) 0.0485354
( 7) 0.0904056    (25) 0.0669052
( 8) 0.0807814    (26) 0.294526
( 9) 0.15223      (27) 0.0451583
(10) 0.0567531    (28) 0.0840954
(11) 0.133715
(12) 0.47444
(13) 0.0646097
(14) 0.132221
(15) 0.160597
(16) 0.0712161
(17) 0.133971
(18) 0.15471
-------------------------------------------------------
```

Jn 10.25

10.1.12. Regressioonitulemuste trükkimine ja salvestamine. Mitmemõõtmelise regressioonanalüüsi protseduur pakub rea võimalusi tulemuste salvestamiseks ja trükkimiseks. Nende võimalusteni jõuame, kasutades aknas joonisel 10.3 ridu *Generate reports* (tulemustabelite moodustamine) ja *Save results* (tulemuste salvestamine). Kumbki neist võimalustest viib omakorda akna abil antava allmenüüni, mis sisaldab rea valikuid.

Joonis 10.26 esitab võimalike tulemustabelite loetelu, need on

1) funktsioontunnuse väärtused y_j (j = 1, ..., n) *observed values*,

2) prognooside väärtused $\tilde{y}_j$ (j = 1, ..., n) *fitted values*;

3) prognoosijäägid ε_j (j = 1, ..., n) *residuals*,

4) standardiseeritud prognoosijäägid $\varepsilon_j/s_{\varepsilon_j}$ *standardized residuals*.

You may display any of the following:
1. observed values
2. fitted values
3. residuals
4. standardized residuals

Enter list of desired choices (quit):

Jn 10.26

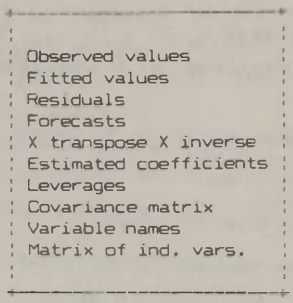
Regression results for tulu

Observation Number	Standardized Residuals
1	-1.81966
2	-0.88287
3	-0.63039
4	-0.12109
5	1.12838
6	0.74801
7	1.36362
8	-0.67775
9	-1.01975
10	1.25494
11	1.04330
12	2.34688
13	-0.52316
14	-0.40299
15	-0.25830
16	0.42661
17	0.61456
18	0.87682
19	-1.48850
20	0.68560
21	-0.99560
22	-1.57102
23	-1.38798
24	-1.01140
25	-0.05491
26	1.00669
27	1.30403
28	0.22878

0 residuals beyond 3 sigma.

Jn 10.27

Valiku tulemusena saadud standardiseeritud prognoosijääkide tabel on esitatud joonisel 10.27.



Jn 10.28

Tulemuste salvestamise võimaluste loetelu on veelgi ulatuslikum, vt akent joonisel 10.28. Selgitame esitatud võimaluste sisu.

1. Funktsioontunnuse väärtused y_j (j = 1, ..., n) *Observed values.*

2. Prognoositud väärtused $\tilde{y}_j$ (j = 1, ..., n) *Fitted values.*

3. Prognoosijäägid ε_j (j = 1, ..., n) *Residuals.*

6. Regressioonikordajate hinnangud b_i (i = 0, ..., k; b_0 on vabaliige) *Estimated coefficients.*

7. Punktide iseärasuskordajad (*Leverages*), vt p 10.1.9.

8. Parameetrite hinnangute kovariatsioonimaatriks *Covariance matrix.*

9. Tunnuste nimed *Variable names.*

10. Lähtemaatriks (objekt-tunnus-maatriks) *Matrix of ind. vars.*

Reale 4 (*Forecasts*) vastab tühi tabel, ka real 5 pakutavat maatriksit $(X'X)^{-1}$ ei ole autoritel õnnestunud saada.

Tulemuste salvestamiseks tuleb valida rida ja fikseerida see klahviga **ENTER**. Seejärel ekraanile ilmunud aknas tuleb omistada andmestikule ja tunnusele nimi, kusjuures võib jätta ka süsteemi poolt pakutud variandi, kui vastavat protseduuri kasutatakse ainult üks kord.

Salvestatud tulemusi on võimalik standardsel viisil välja trükkida, kasutades allmenüü *Data Management* operatsioone (vt [7], p 4.6.9). Näitena võib tuua joonise 10.24, kus on trükitud punktide iseärasuskordajad.

Regressioonanalüüsi protseduuris sisaldub veel võimalus faktorargumentide lisamiseks, sisulistest kaalutlustest lähtudes käsitleme seda aga koos dispersioonanalüüsiga (vt § 10.5).

§ 10.2. Sammregressioon

Nagu märkisime paragrahvis 10.1, sõltub regressioonanalüüsi tulemuse sisukus oluliselt argumenttunnuste valikust. Argumenttunnuste valiku automatiseerimiseks teatavate statistikakriteeriumide abil on loodud rida meetodeid, mille ühiseks nimetuseks on sammregressioon (*Stepwise regression*). Sammregressiooni puhul valitakse suurest potentsiaalsete regressorite hulgast samm-haaval (kas juurdelisamise või ärajätmise teel) teatava kriteeriumi kohaselt välja soodsaim mudel.

10.2.1. Sammregressioon (protseduur *Stepwise Variable Selection*). Sammregressioon on neljas protseduur allmenüüs *K. Regression Analysis*. Selles realiseeritakse vaikimisi nn kasvav sammregressioon, mille käigus argumente mudelisse lisatakse.

Sammregressiooni tellimuspaneel langeb ühte mitmese regressiooni tellimuspaneeliga (vt p 10.1.2), samad on ka selle täitmise reeglid, kusjuures oluline on see, et tellimusse lülitatud argumenttunnuste arv võib olla märksa suurem, kui seda lõppmudelis näha soovitakse. Argumenttunnuste hulka võib lisada ka lähtetunnuste mitmesuguseid funktsioone (SG avaldisi).

Sammregressiooni tellimuse täitmisel ilmub kõigepealt esmatulemuste tabel joonisel 10.29 näidatud kujul, mis vastab 0-sammule. Ühtlasi on see tabel täiendavaks tellimuspaneeliks.

Stepwise Selection for tulu

Selection: Forward	Maximum steps: 500	F-to-enter: 4.00
Control: Manual	Step: 0	F-to-remove: 4.00

R-squared: .0000 Adjusted: .00000 MSE: 31228.4 d.f.: 27

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
			1. vanus	.1179	.3667
			2. teadtoid	.5510	11.3350
			3. sugu	.3280	3.1334
			4. pere	.2433	1.6356
			5. abielus	.2379	1.5596

Jn 10.29

Tabeli päisesse on märgitud, et tegemist on argumentide valikuga sammhaaval (*Stepwise Selection for*), ning näidatud on ka funktsioontunnuse nimi.

Tabeli kahes ülemises reas on kirjas meetodit iseloomustavad parameetrid:

- kasvav e ettepoole valik (*Forward*) tähistab lähtumist ainult vabaliiget sisaldavast mudelist ja argumentide lisamist sammukaupa, kahanev e tahapoole (*Backward*) valik tähistab lähtumist kõiki argumente sisaldavast mudelist ja nende sammukaupa väljajätmist;

- sammude maksimaalarv (*Maximum steps*); siin vaikimisi pakutud väärtus 500 on praktiliste ülesannete jaoks liigagi suur;

- käsitsijuhtimine - *Control: Manual* - tähistab seda, et kõik sammud on ekraanil näha ning järgmine samm tehakse **ENTER**-klahvi abil, automaatjuhtimise korral väljastatakse ainult lõppmudel.

Juhtimisparameetriteks on lisamise ja eemaldamise F-statistikute *F-to-enter* ja *F-to-remove* kriitilised väärtused; nende statistikute jaotuseks (nullhüpoteesile vastaval eeldusel) on q-ndal sammul vastavalt $F(1, n-q-1)$ ja $F(1, n-q)$. Et aga juhul $m \geq 30$ on jaotuse $F(1, m)$ 0,05-kriitiline punkt ligikaudselt võrdne neljaga, siis sobib vaikimisi pakutud kriitiline väärtus 4,00 mõlema parameetri jaoks küllalt hästi.

Viimaste parameetrite tähendus selgub sammprotseduuri kirjeldusest.

10.2.2. Kasvav sammregressioon. Olgu $SS_1^{(q)}$ q-argumendilise mudeli poolt kirjeldatud funktsioontunnuse hajuvus ning $SS_0^{(q)} = SS - SS_1^{(q)}$ vastav jääkhajuvus. Siis iga nimetatud q-elemendilisse komplekti mittekuuluv potentsiaalne argument X_{i_1}, X_{i_2}, ..., X_{i_p} (p = k - q) võiks olla (q+1)-ks argumendiks, mille puhul leitakse ühtlasi uued hajuvuskarakteristikud $SS_1^{(q+1)}$ ja $SS_0^{(q+1)}$. Lisanduva argumendi X_{i_h} arvele tulev hajuvuse osa $SS(i_h)$ arvutatakse lihtsalt,

$$SS(i_h) = SS_1^{(q+1)} - SS_1^{(q)},$$

ning vastav lisamise F-statistik arvutatakse standardselt,

$$F(i_h) = \frac{SS(j_h)(n-q-2)}{SS_0^{(q+1)}}.$$

Sõnastame nüüd hüpoteesipaari

H_0: (q+1)-se argumendi X_{i_h} lisamine ei paranda mudelit oluliselt;

H_1: (q+1)-se argumendi X_{i_h} lisamise tagajärjel mudel paraneb oluliselt.

Nullhüpoteesile vastaval juhul on statistik $F(j_h)$ F-jaotusega vabadusastmete arvudega 1 ja n-q-1. Kui mingi h korral F-statistiku väärtus on suurem selle kriitilisest väärtusest, siis võetakse hüpotees H_1 vastu.

Vastavalt kirjeldatud protseduurile lisatakse mudelisse argument siis, kui antud sammul suurim $F(j_h)$ on suurem või võrdne kui 4 (s.o F-statistiku ligikaudne kriitiline väärtus olulisuse nivool 0,05).

Üsna samal viisil saab juba mudelisse lülitatud q argumendi X_{i_1}, ..., X_{i_q} seast ükshaaval igaühe kohta arvutada, kui suur on funktsioontunnuse kirjeldamisel tema osa mudelis.

Olgu X_{i_g} kontrollitav tunnus. Siis

$$SS^*(i_g) = SS_1^{(q)} - SS_1^{(q-1)},$$

kus $SS_1^{(q-1)}$ on kirjeldatus, mis on saadud, arvates esialgsest q-argumendilisest komplektist välja argumenttunnuse X_{i_g}, $SS_1^{(q)}$ aga on q argumendile vastav kirjeldatus. Tunnuse

x_{i_g} mõju iseloomustab eemaldamise F-statistik

$$F^*(i_g) = \frac{SS^*(i_g)(n-q-1)}{SS_o^{(q)}} .$$

Paneme tähele, et kui mingile q-argumendilisele mudelile lisatakse (q+1)-ne argument, kusjuures lisamise F-statistiku väärtus on F, siis sama argumendi (q+1)-argumendilisest komplektist eemaldamisel on eemaldamise statistiku väärtuseks täpselt samuti F.

Peale lisamise- ja eemaldamise F-statistikute kriitiliste väärtuste on tabelisse märgitud veel parajasti lõpetatud järjekordse sammu number *Step* (joonisel 10.29 on see 0) ning selle sammu tulemusena saadud mudelit iseloomustavad parameetrid: determinatsioonikordaja (*R-squared*), selle parandatud väärtus (*Adjusted*), funktsioontunnuse jääkdispersioon (ruutkeskmine viga *MSE*), vabadusastmete arv n-q-1 (*d.f.*), kus q on juba mudelisse lülitatud argumentide arv.

Edasi järgneb tabeliosa, kus joone all on argumendid - vasakul mudelisse lülitatud q argumenti mudelisse lülitamise järjekorras ja neid iseloomustavad eemaldamise F-statistiku väärtused, paremal - mudelisse veel lülitamata k-q argumenttunnust, neid iseloomustavad lisamise F-statistiku väärtused ning ka iga argumendi ja funktsioontunnuse osakorrelatsioonikordajad (täpsemalt, nende absoluutväärtused), kusjuures juba mudelisse lülitatud argumentide mõju on elimineeritud.

Joonisel 10.29 esitatud tabelis on vasak pool tühi, osakorrelatsioonide asemel on tavalised lineaarsed korrelatsioonikordajad.

ENTER-klahvile vajutamisel tehakse <u>üks samm</u>:

kui mudelisse lülitamata argumentide hulgas leidub vähemalt üks selline, millele vastav lisamise F-statistik on tellimuspaneelil märgitud kriitilisest tasemest suurem, siis lisatakse see mudelisse;

kui sellist argumenti ei leidu, ilmub ekraanile kiri *Final model is fit* - lõplik mudel on leitud, sellega on ka automaatselt juhitav sammprotseduur lõppenud.

Käesoleval juhul kirjeldavad kaht järgmist sammu joonised 10.30 ja 10.31. Viimasega on ühtlasi mudeli automaatne valik lõpetatud.

Stepwise Selection for tulu

Selection: Forward Maximum steps: 500 F-to-enter: 4.00
Control: Manual Step: 1 F-to-remove: 4.00

R-squared: .30360 Adjusted: .27682 MSE: 22583.8 d.f.: 26

Variables in Model	Coeff	F-Remove	Variables Not in Model	P.Corr.	F-Enter
2. teadtoid	7.55901	11.3350	1. vanus	.6350	16.8962
			3. sugu	.1102	.3072
			4. pere	.1931	.9679
			5. abielus	.1762	.8015

Jn 10.30

Stepwise Selection for tulu

Selection: Forward Maximum steps: 500 F-to-enter: 4.00
Control: Manual Step: 2 F-to-remove: 4.00

R-squared: .58445 Adjusted: .55121 MSE: 14015.1 d.f.: 25

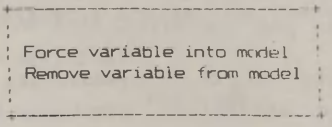
Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
1. vanus	-15.5487	16.8962	3. sugu	.1304	.4152
2. teadtoid	12.6518	34.3247	4. pere	.2086	1.0914
			5. abielus	.2441	1.5203

Jn 10.31

Kuivõrd mudeli kahanev valik (*Backward*) on ülalkirjeldatud protseduuriga põhimõtteliselt sarnane, ei hakka me seda lähemalt kirjeldama.

10.2.3. Tellija poolt soovitud tunnuste mudelile lisamine ja mudelist eemaldamine. Pärast automaatse valiku alusel teatud mõttes optimaalse mudeli leidmist on tellijal võimalik seda vastavalt oma soovile mõnevõrra muuta (tuginedes olemasolevale eelinformatsioonile).

Klahvi **F5** abil saadakse ekraanile aken (vt jn 10.32), mis lubab mudelisse argumente lisada - *Force variable into*



Jn 10.32

model - või eemaldada - *Remove variable from model*. Pärast vastava valiku tegemist ilmub ekraanile aken (vt joon 10.33), mis annab võimaluse lisatava/eemaldatava tunnuse nime märkimiseks.

Specify variable to be entered into model:

Jn 10.33

Järgneb standardne samm.

Märgime veel, et sammregressiooni protseduur võimaldab saada üldiselt samu täiendavaid tulemusi, mis mitmene regressioongi.

§ 10.3. Kantregressioon (*ridge regression*)

10.3.1. "Halvasti määratud" (*ill-conditioned*) argumenttunnuste maatriks ja sellega seotud probleemid. Mõnikord võib juhtuda, et prognoosiülesandes on argumendid omavahel tugevasti korreleeritud, sellise olukorra ekstremaalseks näiteks on olukord, kus üks argument on ülejäänute kaudu lineaarselt avalduv,

$$\mathbf{X}_{i_1} = \sum_{l=2}^{q} \gamma_l \mathbf{X}_{i_l} , \quad q \leq k,$$

millega on samaväärne seos

$$\sum_{l=1}^{q} \gamma'_l \mathbf{X}_{i_l} = 0 , \qquad (10.7)$$

kus kõik kordajad γ'_l ei võrdu nulliga.

Põhimõtteliselt on võimalik, et argumenttunnuste seas leidub mitu (omavahel sõltumatut) lineaarset sõltuvust kujul (10.7).

Juba ühe sõltuvuse (10.7) olemasolust järeldub, et regressioonikordajate hindamisel kasutataval maatriksil $\mathbf{X}'\mathbf{X}$ puudub pöördmaatriks, seega valem (10.2) regressioonikordajate vektori b hindamiseks ei ole kasutatav.

Probleemi lahendamiseks on mitu võimalust:

- kasutada $(\mathbf{X}'\mathbf{X})^{-1}$ asemel üldistatud pöördmaatriksit,

- teisendada maatriksit $\mathbf{X}'\mathbf{X}$ nõnda, et ta küllalt vähe muutuks, kuid teisendatud maatriksil eksisteeriks kindlasti pöördmaatriks.

Vajalike teisenduste seas on üks kõige lihtsamaid maatriksile $\mathbf{X}'\mathbf{X}$ positiivselt määratud diagonaalmaatriksi D liitmine, mis kindlasti tagab maatriksi $(\mathbf{X}'\mathbf{X} + D)^{-1}$ olemasolu. Sellele ideele ongi rajatud meetod, mille ingliskeelseks nimetuseks on *ridge regression* (vastava eestikeelse terminina soovitame nimetust kantregressioon).

Märgime ühtlasi, et toodud lahendus on efektiivne ka sellistel juhtudel, kus seos (10.7) ei kehti täpselt, vaid üksnes ligikaudselt: mõni argumenttunnus on ülejäänute kaudu väga hästi prognoositav (nt $R^2 > 0{,}9$). Nimelt niisuguste juhtude kohta öeldaksegi, et tegemist on "halvasti määratud" argumenttunnuste maatriksiga. Argumenttunnuste maatriksi $\mathbf{X}$ ja seetõttu ka maatriksi $\mathbf{X}'\mathbf{X}$ halvast määratusest tuleneb, et

- maatriksi $(\mathbf{X}'\mathbf{X})^{-1}$ elemendid on suured, järelikult ka regressioonikordajate hajuvuskarakteristikud on suured, need kordajad on ebatäpsed;

- kuivõrd $\det(\mathbf{X}'\mathbf{X})$ väärtus on nullilähedane, on regressioonikordajate arvutustäpsus suhteliselt madal.

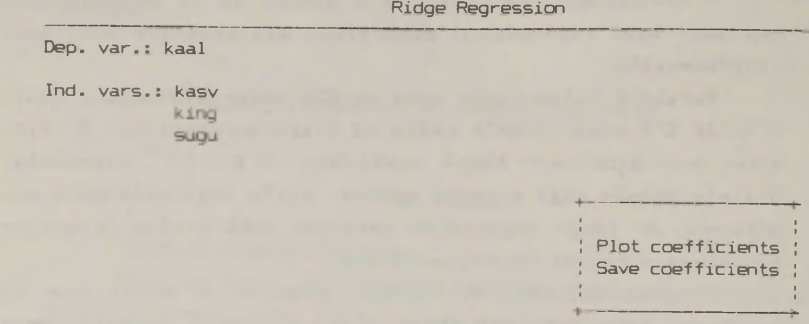
Just sellises olukorras soovitataksegi kasutada kantregressiooni, kuigi selle tulemusena saadud regressiooniparameetrite hinnangud on nihkega. Kui aga parameetri hajuvus märgatavalt väheneb, on mõnikord väike nihe eelistatav.

10.3.2. Protseduur *Ridge Regression*. Protseduur *Ridge Regression* on allmenüüs *K. Regression Analysis* viies. Tellimuspaneel on põhiosas sarnane mitmese regressiooni tellimuspaneeliga, kuid sisaldab allreas nimetatud protseduurile eriomaseid parameetreid, vt jn 10.34. Nimelt tehakse protseduuris läbi terve rida arvutusi, kus parandus maatriksile $\mathbf{X}'\mathbf{X}$ määratakse eeskirjaga

$$\mathbf{X}'\mathbf{X} + \vartheta\mathbf{I} \qquad (10.8)$$

ning parameetrit ϑ muudetakse samm-sammult ette antud miini-

mum- ja maksimumväärtuste vahel; ette antav on ka sammude arv vahemikus 2...20.



Jn 10.34

Vaikimisi on miinimumväärtuseks antud 0 (vastab vähimruutude mudelile) ja maksimumväärtuseks 1 (mis rakendustes end harva õigustab).

Joonisel 10.34 on ära trükitud ka tulemuste loetelu aken, mis saadakse ekraanile tegelikult alles klahvile F5 vajutamisel.

Näeme, et siin on ainult kaks valikut - regressioonikordajate graafiline esitamine ja salvestamine.

10.3.3. Standardiseeritud regressioonimudel. Kantregressioon arvutab nn standardiseeritud regressioonimudeli kordajad β:

$$\hat{Y} = \beta_1\hat{X}_1 + \beta_2\hat{X}_2 + \ldots + \beta_k\hat{X}_k, \qquad (10.8)$$

kus $\hat{Y}$ ja $\hat{X}_i$ on standardiseeritud tunnused,

$$\hat{Y} = \frac{Y - \bar{y}}{s_Y}, \qquad \hat{X}_i = \frac{X_i - \bar{x}_i}{s_i},$$

$\bar{y}$ ja $\bar{x}_i$ on vastavad keskmised ning s_Y, s_i - standardhälbed.

Tähistame tunnustele **kaal**, **kasv**, **king** ja **sugu** vastavad standardiseeritud tunnused **stankaal**, **stankasv**, **stanking** ja **stansugu**. Siis on vastav standardiseeritud regressioonimudel (mis on leitud protseduuri *Multiple Regression* abil, vt jn 10.35) järgmine:

$$\textbf{stankaal} = -0{,}109415\textbf{stankasv} + 1{,}18355\textbf{stanking} + \\ + 0{,}28347\textbf{stansugu}$$

Model fitting results for: stankaal

Independent variable	coefficient	std. error	t-value	sig.level
CONSTANT	-0.000011	0.108773	-0.0001	0.9999
stankasv	-0.109415	0.273968	-0.3994	0.6932
stanking	1.18355	0.38574	3.0683	0.0053
stansugu	0.28347	0.252667	1.1219	0.2730

R-SQ. (ADJ.) = 0.6687 SE= 0.575570 MAE=0.434431 DurbWat= 1.695
Previously: 0.0000 0.000000 0.000000 0.000
28 observations fitted, forecast(s) computed for 0 missing val.of dep.var.

Jn 10.35

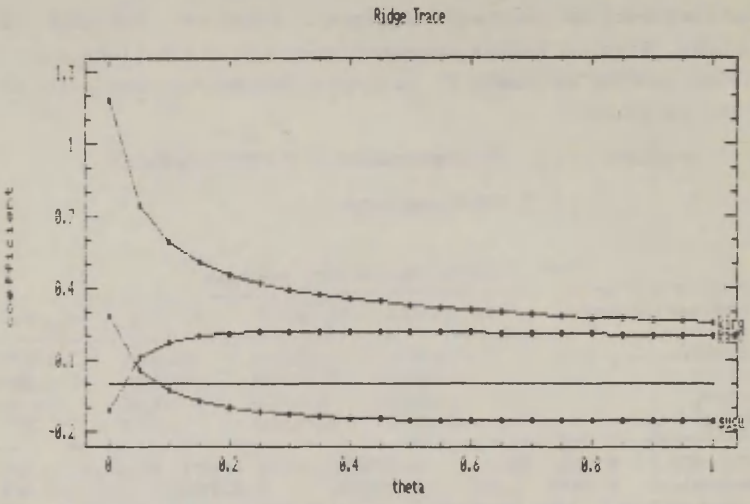
10.3.4. Kantregressiooni tulemus. Kantregressiooni käigus arvutatakse mudeli kordajatele β_1, β_2, ja β_3 rida hinnanguid, kasutades valemit

$$\beta = (\mathbf{X}'\mathbf{X} + \vartheta\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}$$

rea erinevate ϑ väärtuste korral. Joonisel 10.34 esitatud tellimusele vastab ϑ väärtuste hulk

$$\vartheta = i \cdot 0{,}05, \quad i = 0, \ldots, 20,$$

vastavad β väärtused on esitatud graafikuna joonisel 10.36 ja tabelina joonisel 10.37. Joonisel 10.38 on esitatud kordajate hinnangute muutumine sõltuvalt ϑ väärtusest, kui ϑ muutub vahemikus 0...0,1. Parameetri ϑ väärtuste 0,1, 0,5 ja 1 jaoks on välja arvutatud ka prognoosid vastavalt valemile (10.9), need on nimetatud tunnusteks **KANTO1**, **KANTO5** ja **KANT1**; ϑ väärtus 0 annab, nagu näha tabeli (jn 10.37) esimesest reast, tavalise regressioonanalüüsi tulemuse (vrdl jn 10.34). Vastava prognoosi nimetuseks on **FITKANT**.

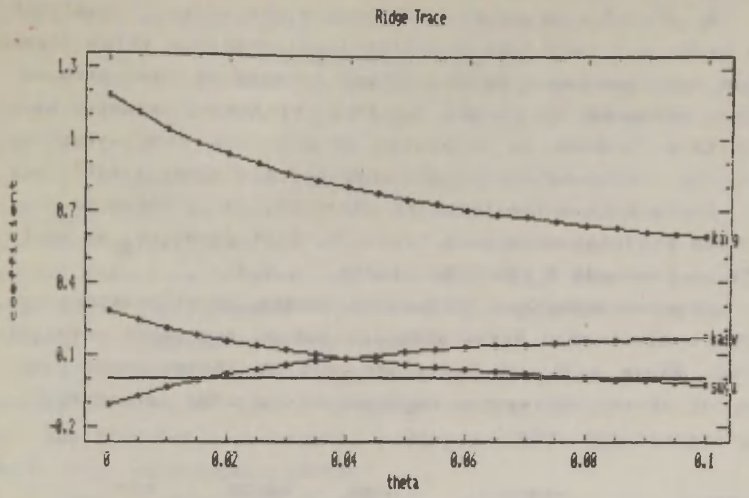


Jn 10.36

Variable: WORKAREA.COEFFS (length = 21 3)

(1,1)	0.109416	(1,2)	1.18356	(1,3)	0.283472
(2,1)	0.108348	(2,2)	0.748365	(2,3)	0.0603526
(3,1)	0.173577	(3,2)	0.592671	(3,3)	-0.0229176
(4,1)	0.201078	(4,2)	0.510917	(4,3)	-0.0669221
(5,1)	0.214247	(5,2)	0.45954	(5,3)	-0.0938947
(6,1)	0.220717	(6,2)	0.42364	(6,3)	-0.111799
(7,1)	0.223642	(7,2)	0.396726	(7,3)	-0.124262
(8,1)	0.224532	(8,2)	0.375518	(8,3)	-0.133193
(9,1)	0.224186	(9,2)	0.358175	(9,3)	-0.139703
(10,1)	0.223061	(10,2)	0.343582	(10,3)	-0.144485
(11,1)	0.221433	(11,2)	0.331026	(11,3)	-0.147994
(12,1)	0.219476	(12,2)	0.320027	(12,3)	-0.150544
(13,1)	0.217303	(13,2)	0.31025	(13,3)	-0.152356
(14,1)	0.214997	(14,2)	0.301455	(14,3)	-0.153592
(15,1)	0.212592	(15,2)	0.293463	(15,3)	-0.154372
(16,1)	0.210145	(16,2)	0.28614	(16,3)	-0.154788
(17,1)	0.207674	(17,2)	0.279381	(17,3)	-0.154911
(18,1)	0.205179	(18,2)	0.273104	(18,3)	-0.154795
(19,1)	0.202734	(19,2)	0.267244	(19,3)	-0.154485
(20,1)	0.200289	(20,2)	0.261748	(20,3)	-0.154017
(21,1)	0.19787	(21,2)	0.256574	(21,3)	-0.153418

Jn 10.37



Jn 10.38

Sample Correlations

	kaal	FITKANT	KANTO1	KANTO5	KANT1
kaal	1.0000	.8400	.7959	.7836	.7783
	(28)	(28)	(28)	(28)	(28)
	.0000	.0000	.0000	.0000	.0000
FITKANT	.8400	1.0000	.9476	.9329	.9265
	(28)	(28)	(28)	(28)	(28)
	.0000	.0000	.0000	.0000	.0000
KANTO1	.7959	.9476	1.0000	.9990	.9980
	(28)	(28)	(28)	(28)	(28)
	.0000	.0000	.0000	.0000	.0000
KANTO5	.7836	.9329	.9990	1.0000	.9998
	(28)	(28)	(28)	(28)	(28)
	.0000	.0000	.0000	.0000	.0000
KANT1	.7783	.9265	.9980	.9998	1.0000
	(28)	(28)	(28)	(28)	(28)
	.0000	.0000	.0000	.0000	.0000

Coefficient (sample size) significance level

Jn 10.39

Et võrrelda omavahel erinevate ϑ väärtuste korral saadud prognoose, arvutame korrelatsioonimaatriksi kõigi nimetatud viie tunnuse - prognoositava tunnuse ja tema prognooside - vahel (vt jn 10.39). Märgime, et korrelatsioonid prognoositava tunnuse ja erinevate prognooside vahel võrduvad vastavalt definitsioonile mitmeste korrelatsioonikordajatega.

Korrelatsioonimaatriksist järeldub, et ϑ väärtuse suurenedes prognoos mõnevõrra halveneb. Siit järeldub, et sobivaim on kasutada ϑ väärtusi piires 0...0,1.

Standardiseeritud funktsioontunnuse ja erinevate prognooside olulisemad arvkarakteristikud on kujutatud joonisel 10.40. Näeme siit, et kantprognoosid on vähimruutude prognoosist märksa väiksemate hajuvuse näitajatega (standardhälve, kvartiilide vahe).

Variable:	stankaal	FITKANT	KANT05	KANT1
Sample size	28	28	28	28
Average	-1.51618E-6	-1.51618E-6	6.49964E-7	4.50951E-7
Median	-0.0492773	-0.0901326	-0.0829538	-0.06784
Variance	0.999986	0.705514	0.0892338	0.063185
Standard deviation	0.999993	0.839949	0.29872	0.251366
Interquartile range	1.59199	1.48392	0.47988	0.396723

Jn 10.40

§ 10.4. Mittelineaarne regressioon

Mittelineaarne regressioon on ilmselt üks intrigeerivamaid ja ahvatlevamaid mudeleid, mida paljud andmeanalüüsijad soovivad oma andmete uurimisel rakendada, lähtudes (põhimõtteliselt õigest) eeldusest, et lineaarne mudel on teatavaks esialgseks lähendiks, mittelineaarne mudel aga annab otsitava seose edasi märksa adekvaatsemalt ja täiuslikumalt.

Samal ajal aga põhjustab mittelineaarse regressioonanalüüsi rakendamine (sõltumatult programmist, mida kasutatakse) pahatihti pettumusi ja arusaamatusi. Need on tingitud järgmistest asjaoludest.

1. Ei ole olemas üldist "mittelineaarset mudelit" erinevalt üldisest lineaarsest mudelist, mis tõepoolest eksis-

teerib. Soovides saada mittelineaarset mudelit, peab tellija täpsustama ka selle kuju. Statistikaprotseduur on võimeline ainult selle mudeli parameetreid hindama.

2. Mittelineaarse protseduuri eripärast tingituna tuleb hinnatavatele parameetritele ette määrata ka nende ligikaudsed väärtused.

3. Mittelineaarse regressioonanalüüsi protseduuris toimub arvutamine tulemuste järk-järgulise täpsustamise teel nn iteratsiooniprotsessis. Tuleb aga arvestada võimalust, et kas mudeli oletatava kuju või parameetrite alglähendite ebasobivuse tõttu iga järgmine lähend ei ole eelmisest täpsem ja lahendi koonduvuse asemel see hajub.

4. Mittelineaarne regressioonanalüüs on lineaarsest märksa töömahukam, samuti nõuab protseduur rohkem arvuti mälu. Seetõttu võtab arvutamine üldiselt palju aega, kuid võib ka katkeda mälunappuse tõttu.

Kõigi nende asjaolude tõttu on mittelineaarse regressioonanalüüsi protseduuride kasutamine ülejäänud protseduuridega võrreldes märksa keerulisem. Sellepärast alustamegi nn lineariseeruvate mudelite tutvustamisest. Neid saab leida lineaarse regressiooni protseduuride abil.

10.4.1. Mittelineaarse mudeli konstrueerimise protsess. Lineariseeruv mudel. Kui konstrueeritava mudeli kohta ei ole olemas olulist eelinformatsiooni, tuleb õigeks pidada mudeli konstrueerimist samm-sammult, uurides eraldi iga oletatava argumendi lineaarset ja ka mittelineaarset mõju (protseduur *Simple Regression*, vt § 7.5) kasutades ka mittelineaarseid mudeleid .

Jälgime seda mõttekäiku konkreetsel näitel; oletame, et meie eesmärgiks on leida tulu sõltuvust teadustööde arvust (teadtoid), arvutioskusest (arvutoskus) ja vastaja soost (sugu).

Esimesel sammul vaatame läbi funktsioontunnuse lineaarse (ja ka teiste mudelitega antud) sõltuvuse igast argumenttunnusest eraldi, vt näiteks jooniseid 10.41 ja 10.42. Seejärel lülitame kõik kolm argumenti ühisesse lineaarsesse mudelisse, vt jn 10.43. Saadav kirjeldatuse protsent 48,8 %

(parandatud determinatsioonikordaja järgi), mis on saadud mitmese lineaarse regressiooni mudelist, on märksa kõrgem üksiktunnuste mõjudest, mille hulgast suurim oli 36,26 %. See näitaja on orientiiriks, millest lähtume mudeli edasisel täpsustamisel.

Regression Analysis - Linear model: Y = a+bX

Dependent variable: tulu — Independent variable: teadtoid

Parameter	Estimate	Standard Error	T Value	Prob. Level
Intercept	379.957	38.1141	9.96894	.00000
Slope	7.55901	2.2452	3.36675	.00238

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	Prob. Level
Model	255987.25	1	255987.25	11.3	.00238
Error	587179.71	26	22583.84		
Total (Corr.)	843166.96	27			

Correlation Coefficient = 0.551001 R-squared = 30.36 percent
Stnd. Error of Est. = 150.279

Jn 10.41

Regression Analysis - Multiplicative model: Y = aX^b

Dependent variable: tulu — Independent variable: arvutoskus

Parameter	Estimate	Standard Error	T Value	Prob. Level
Intercept*	5.63289	0.12883	43.7233	.00000
Slope	0.471732	0.122654	3.84602	.00070

* NOTE: The Intercept is equal to Log a.

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	Prob. Level
Model	1.413797	1	1.413797	14.79190	.00070
Error	2.485059	26	.095579		
Total (Corr.)	3.898856	27			

Correlation Coefficient = 0.602178 R-squared = 36.26 percent
Stnd. Error of Est. = 0.309159

Jn 10.42

Model fitting results for: tulu

Independent variable	coefficient	std. error	t-value	sig.level
CONSTANT	293.799372	109.297576	2.6881	0.0129
teadtoid	6.642439	2.110898	3.1467	0.0044
arvutoskus	73.060464	20.815134	3.5100	0.0018
sugu	-69.745732	54.764737	-1.2736	0.2150

R-SQ. (ADJ.) = 0.4886 SE= 126.375046 MAE= 95.913915 DurbWat= 2.376
Previously: 0.0000 0.000000 0.000000 0.000
28 observations fitted, forecast(s) computed for 0 missing val.ofdep.var.

Jn 10.43

Järgmiseks mõeldavaks sammuks on nn lineariseeruva mudeli kasutamine - see on mudel, mis parameetrite suhtes on lineaarne, kuid sisaldab argumentide mittelineaarseid funktsioone.

Et üks kõige lihtsamaid võimalikke mittelineaarseid funktsioone on ruutpolünoom, siis võtamegi selle aluseks ning defineerime lisaks lähteargumendile veel argumendid

sugu×teadtoid,
sugu×arvutoskus,
teadtoid×arvutoskus,
teadtoid×teadtoid,
arvutoskus×arvutoskus.

Stepwise Selection for tulu

Selection: Forward Maximum steps: 500 F-to-enter: 4.00
Control: Manual Step: 0 F-to-remove: 4.00

R-squared: .00000 Adjusted: .00000 MSE: 31228.4 d.f.: 27

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
			1. sugu	.3280	3.1334
			2. teadtoid	.5510	11.3350
			3. arvutoskus	.4345	6.0495
			4. sugu*teadtoid	.4791	7.7463
			5. sugu*arvutoskus	.1534	.6269
			6. teadtoid*arvuto	.6694	21.1100
			7. teadtoid*teadto	.5275	10.0220
			8. arvutoskus*arvu	.3481	3.5843

Jn 10.44

Kuivõrd **sugu** on kaheväärtuseline tunnus, ei lisa selle ruut täiendavat informatsiooni. Seega oleks meie järgmiseks sammuks teha lineaarne regressioon, kasutades 8 argumenti (lisaks esialgsetele viis loetletud uut tunnust). Et nende hulgast ainult olulised kombinatsioonid välja valida, rakendame sammregressiooni, vt jn 10.44, mille tulemuseks saame esialgsest lineaarsest mudelist veidi parema (kirjeldatus 53 % parandatud determinatsioonikordaja järgi) kolmeparameetrilise mudeli:

tulu = 228,28 + 51,47×**arvutoskus** + 2,94×**teadtoid**×**arvutoskus**,

vt jn 10.45. Tunnus **sugu** mudelis ei sisaldu, kuid on näha (vt jn 10.46), et kõik liikmed, mis sisaldavad tunnust **sugu** tegurina, on suhteliselt suure lisamise F-statistiku väärtusega. See tekitab mõtte kasutada tunnust **sugu** ka lõpp-mudelis tegurina.

Model fitting results for: tulu

Independent variable	coefficient	std. error	t-value	sig.level
CONSTANT	228.2842	61.030666	3.7405	0.0010
arvutoskus	51.468129	19.815289	2.5974	0.0155
teadtoid*arvutoskus	2.94185	0.63204	4.6545	0.0001

R-SQ. (ADJ.) = 0.5306 SE= 121.070721 MAE= 90.617714 DurbWat= 1.819
Previously: 0.0000 0.000000 0.000000 0.000
28 observations fitted, forecast(s) computed for 0 missing val.of dep.var.

Jn 10.45

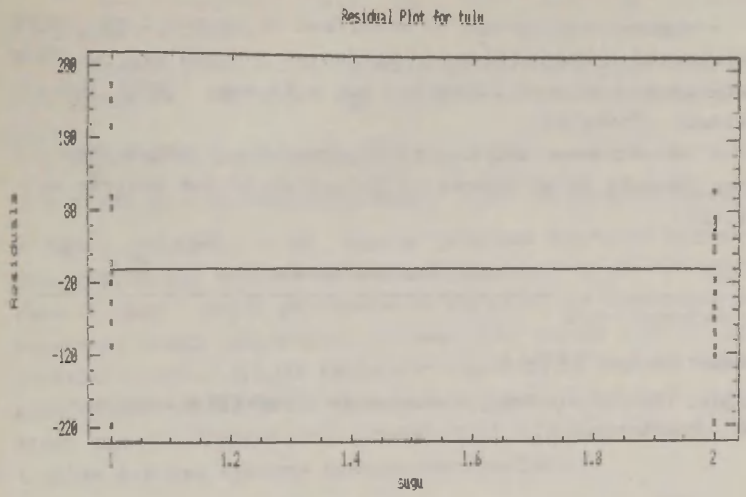
Stepwise Selection for tulu

Selection: Forward Maximum steps: 500 F-to-enter: 4.00
Control: Manual Step: 2 F-to-remove: 4.00

R-squared: .56539 Adjusted: .53062 MSE: 14658.1 d.f.: 25

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
3. arvutoskus	51.4681	6.7465	1. sugu	.2853	2.1261
6. teadtoid*arvuto	2.94185	21.6647	2. teadtoid	.1301	.4134
			4. sugu*teadtoid	.2918	2.2331
			5. sugu*arvutoskus	.3009	2.3899
			7. teadtoid*teadto	.2131	1.1420
			8. arvutoskus*arvu	.1084	.2854

Jn 10.46



Jn 10.47

Jooniselt 10.47 näeme, et üldiselt vastavad suuremad prognoosijäägid tunnuse sugu väiksematele väärtustele, siit tuleneb mõte kasutada lõppmudelis tunnuse sugu pöördväärtust. Selleks tuleb meil kasutada mittelineaarset regressioonanalüüsi, mida asumegi nüüd vaatlema.

10.4.2. Mittelineaarne regressioonanalüüs *Nonlinear Regression*. Protseduur *Nonlinear Regression* paikneb allmenüüs *K. Regression Analysis* viimasel real. Protseduuri käivitamisel ilmub ekraanile tellimuspaneel, vt jn 10.48.

Selle esimesele reale tuleb märkida funktsioontunnus (*Dep. variable*). Teisele reale *Parameter vector* on tarvis kirjutada parameetrite algväärtused nende indeksite järjekorras. See rida on kasulik täita pärast mudeli väljakirjutamist. Mudel tuleb kirjutada aknasse *Function*, jälgides järgmisi näpunäiteid:

- mudeliks on parameetrite ja argumenttunnuste mingi funktsioon, mis on kirjutatud SG avaldisena;
- mudel ei sisalda funktsioontunnust ega võrdusmärki;
- parameetrid kirjutatakse kujul PARM[i], kus i on parameetri indeks (järjekorranumber);

- argumenttunnusteks on suvalised arvtunnused või nende funktsioonid (**SG**-avaldised), kusjuures on nõutav, et kõik kasutatavad tunnused oleksid sama pikkusega (kuid tohivad sisaldada tühikuid);

- **SG**-süsteemi eripära tõttu (avaldised väärtustatakse alates lõpust) tuleb tehete järjekord avaldises määrata sulgude abil.

```
                    Nonlinear Regression
-------------------------------------------------------------------------

Dep. variable:  tulu

Parameter vector: 200 50 3

Function: PARM[1] + (PARM[2] * arvutoskus) + (PARM[3] * teadtoid *
          arvutoskus)



Maximum iterations: 25                    Initial Marquardt parameter: 0.01
Maximum function calls: 200               Initial scaling factor: 20
Stopping cond. on res. ss: 1E-4           Max. value of Marquardt parm.: 120
Stopping cond. on estimates: 1E-3
```

Jn 10.48

Joonisel 10.48 on funktsiooniks kirjutatud joonisel 10.43 leitud mudel, kusjuures parameetrite algväärtusteks on valitud vastavalt 200, 50 ja 3.

Tellimuspaneeli allservas on iteratsiooniprotsessi reguleerivad parameetrid; nende väärtusteks sobivad hästi programmi poolt vaikimisi antud väärtused.

Mittelineaarne regresioonanalüüs on aegavõttev protseduur; arvutamise käigus ilmub arvutiekraanile teave selle kohta, mitmes iteratsioon on parajasti käsil, samuti see, kui suur on jääkdispersioon arvutataval sammul.

Arvutamisprotsess lõpeb, kui on täidetud üks tingimustest:

on saadud vajaliku täpsusega lahend;

- on tehtud teatav arv samme, kuid protsess ei koondu ja tulemusi ei saada.

Viimasel juhul aitab vahel olukorda parandada alglähendite muutmine, mõnikord aga osutub mudel tervikuna sobimatuks.

Käesoleval juhul saime pärast teist iteratsiooni mudeli

tulu = 226,72 + 51,87×**arvutoskus** + 2,95×**teadtoid**×**arvutoskus**,

vt tabel jooniselt 10.49, kus on trükitud hinnatud parameetrite väärtused. Märkame erinevusi joonisel 10.45 leitud parameetritega - selle põhjuseks on asjaolu, et lineaarse regressiooni puhul määratakse parameetrid valemi (10.1) järgi põhimõtteliselt täpselt (kuigi iteratiivsete sammude kasutamine arvutuskäigus pole välistatud), mittelineaarne regressioon aga on olemuslikult iteratiivne ning arvutused lõpetatakse teatava täpsuse taseme saavutamisel.

Model Fitting Results

	estimate	stnd.error	ratio
Coefficient 1	226.716179	61.0314942	3.71474
Coefficient 2	51.870046	19.8155585	2.61764
Coefficient 3	2.947184	.6320486	4.66291

Total iterations = 2 Total function evaluations = 9

Jn 10.49

Mudeli detailsemaid analüüsivõimalusi pakub aken joonisel 10.50.

Analysis of variance
Plot fitted model
Plot residuals
Summarize residuals
Plot predicted values
Probability plot
Save residuals
Save covariance matrix
Save parameter estimates

Jn 10.50

Muudame nüüd oma mudeli mittelineaarseks, lisades tunnuse **sugu**:

$$\text{tulu} = \frac{b_1 + b_2\text{arvutioskus} + b_3\text{teadtoid}\times\text{arvutoskus}}{(\text{sugu})^{b_4}}$$

Vastav tellimus on esitatud joonisel 10.51, parameetrite hinnangud on kirjas tabelis joonisel 10.52 ning mudeli olulisust iseloomustav dispersioonanalüüsi tabel on esitatud joonisel 10.53.

Nonlinear Regression

Dep. variable: tulu

Parameter vector: 200 50 3 0.5

Function: (PARM[1] + (PARM[2]*arvutoskus) + (PARM[3]*teadtoid*arvutoskus))/
(sugu^PARM[4])

Maximum iterations: 25
Maximum function calls: 200
Stopping cond. on res. ss: 1E-4
Stopping cond. on estimates: 1E-3

Initial Marquardt parameter: 0.01
Initial scaling factor: 20
Max. value of Marquardt parm.: 120

Jn 10.51

Model Fitting Results

	estimate	stnd.error	ratio
Coefficient 1	231.569142	66.2074311	3.49763
Coefficient 2	67.235022	24.4647211	2.74824
Coefficient 3	2.748369	.6855750	4.00885
Coefficient 4	.231966	.1506052	1.54023

Total iterations = 3 Total function evaluations = 16

Jn 10.52

Tabelist joonisel 10.52 näeme, et ka selles mudelis ei ole tunnus sugu oluline, sest sellele tunnusele vastava parameetri b_4 ja tema standardhälbe suhe on 1,57 < 2. Teiselt

Analysis of Variance for the Full Regression

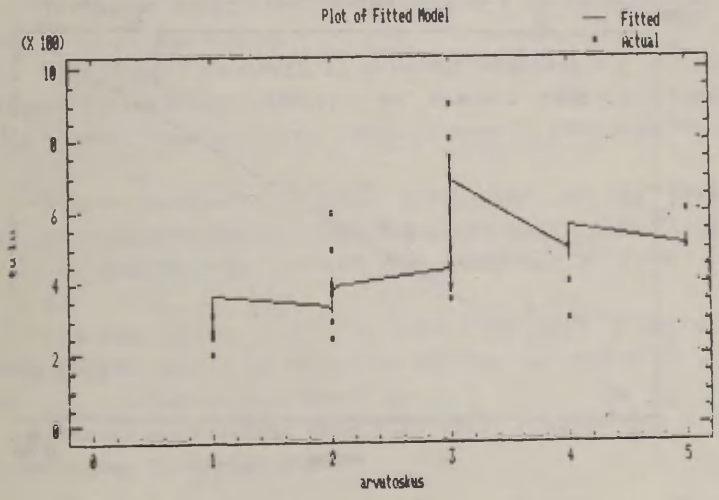
source	sum of squares	df	mean square	ratio
Model	6578881.2	4	1644720.3	118.7
Error	332543.78	24	13855.99	
Total	6911425.0	28		
Total (corr.)	843166.96	27		

R-squared = 0.605602

Jn 10.53

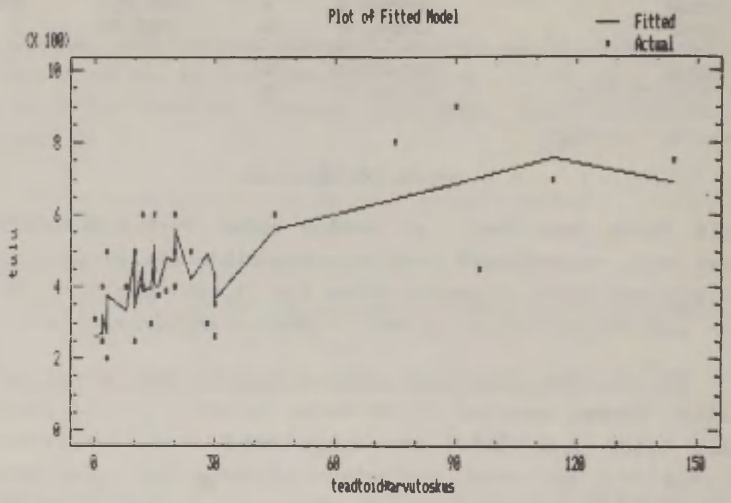
poolt tuleb tunnistada, et saadud mudel kirjeldab siiski (tõsi küll, parandamata determinatsioonikordaja alusel, mis kirjeldatust mõne protsendi võrra üle hindab) üle 60 % tunnuse tulu hajuvusest ja on seni vaadeldud mudelitest parim.

10.4.3. Mittelineaarse regressioonanalüüsi graafikud. Valides aknast joonisel 10.50 teise rea *Plot fitted model*, saame leida graafikuid, mis kirjeldavad funktsioontunnuse ja prognoosi sõltuvust üksikutest argumentidest. Joonistel 10.54 ja 10.55 on esitatud tulu sõltuvus tunnusest arvutoskus ning tunnuste teadtoid ja arvutoskus korrutisest, kusjuures

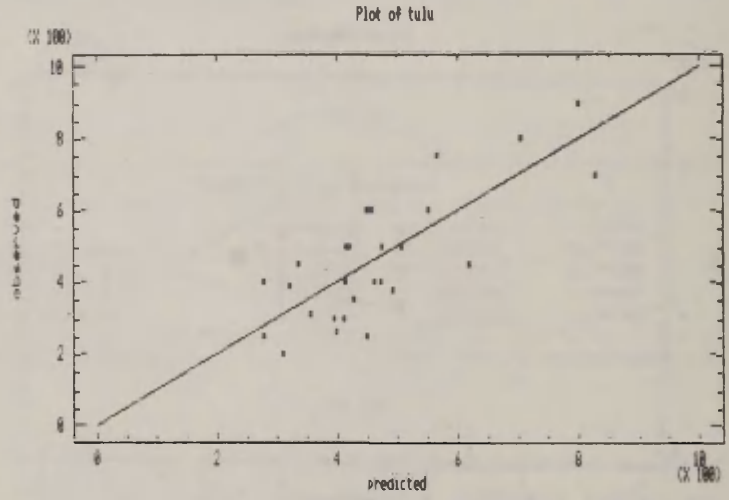


Jn 10.54

funktsioontunnuse mõõdetud väärtusi esitavad punktid, prog- - noositud väärtusi aga murdjoon.



Jn 10.55



Jn 10.56

Murdjoonekujulise mudeli kohta tuleb selgituseks märkida, et see ühendab prognoosipunkte, mis on arvutatud vastavalt olemasolevatele argumendi väärtustele. Kuivõrd aga prognoos sõltub mitmest argumendist, siis on graafikul esitatud argumendi sama (või lähedase) väärtuse korral võimalikud mitmed küllaltki erinevad prognoosi väärtused. Sellest tulenebki graafiku esmapilgul mõistetamatu sarviline kuju.

Seevastu joonis 10.56 esitab traditsioonilise graafiku, mis kujutab funktsioontunnuse ja prognoosi vahekorda. Et siit erilisi ebareeglipärasusi ei paista, võib leitud mudeli lugeda lõplikuks.

§ 10.5. Mitmefaktoriline dispersioonanalüüs

10.5.1. Mitmefaktorilise dispersioonanalüüsi teoreetiline mudel. Dispersioonanalüüs on laialdaselt kasutatav mudeli konstrueerimise meetod, mille puhul

- prognoositav e funktsioontunnus on arvuline;
- argumendid e faktorid on diskreetsed tunnused (arvulised või mittearvulised).

Tähistame funktsioontunnuse tähega Y ja faktorid vastavalt $X_1, \ldots, X_k$, $k \geq 1$.

Olgu i-ndal faktoril s_i erinevat väärtust $x_1^i, \ldots, x_{s_i}^i$. Dispersioonanalüüsis nimetatakse faktori väärtusi tasemeteks. Lause "faktor X_i on 1-ndal tasemel" tähendab võrdust $X_i = x_1^i$.

Dispersioonanalüüsi puhul oletatakse, et iga faktori iga tase avaldab funktsioontunnusele teatavat mõju. Faktori X_i 1-nda taseme mõju tähistatakse sümboliga α_1^i ($1 = 1, \ldots, s_i$; $i = 1, \ldots, k$).

Vaatleme valimi j-ndat objekti (vaatlust). Oletame, et selle objekti puhul on faktor X_1 tasemel 1_1, faktor X_2 tasemel 1_2, ..., faktor X_k tasemel 1_k.

Dispersioonanalüüsi mudeli kohaselt avaldub siis funktsioontunnuse Y väärtus kujul

$$y_j = \mu + \alpha_{1_1}^1 + \alpha_{1_2}^2 + \ldots + \alpha_{1_k}^k + \varepsilon_j, \qquad (10.10)$$

kus μ on konstantne vabaliige, ε_j - juhuslik prognoosijääk (ka viga) ja $\alpha_{i_l}^{i}$ - faktorite mõjud.

Mudelit (10.10) nimetatakse peamõjude mudeliks. Peamõjude mudel sisaldab üldiselt

$$1 + s_1 + \ldots + s_k$$

parameetrit, need on μ, α_1^1, ..., $\alpha_{s_1}^1$, α_1^2, ..., $\alpha_{s_k}^k$.

Sageli on loomulik oletada, et funktsioontunnusele mõjuvad faktorite kõrval ka nende kombinatsioonid. Illustreerime seda järgmise näite varal.

Olgu funktsioontunnus Y mingi vilja saagikus, X_1 - sademete hulk ja X_2 - päikesepaiste hulk suve jooksul (kuigi viimased tunnused on oma olemuselt pidevad, on neid võimalik sobivalt diskretiseerida).

Võib arvata, et funktsioontunnust mõjutab X_1 ja X_2 kõrval ka nende koosmõju, sest erineva päikesepaiste hulga korral on vilja saagikusele soodus erinev sademete hulk - kui päikest on palju, on ka sadeneid rohken tarvis, pilviste ilmade puhul on optimaalne hoopis väiksem sademete kogus.

Et seda arvestada, tuleks dispersioonanalüüsi mudelisse lisada täiendava faktorina faktorite koosmõju. Koosmõju erinevate tasemete arv võrdub kummagi faktori tasemete arvude korrutisega.

Üldiselt tähistatakse faktorite X_i ja X_h koosmõjusid sümbolitega γ_{lg}^{ih}, $l = 1, \ldots, s_i$, $g = 1, \ldots, s_h$, ning teist järku (kahe faktori) koosmõjudega mudeli üldkuju on järgmine:

$$y_j = \mu + \sum_{i=1}^{k} \alpha_{i_l}^{i} + \sum_{i=1}^{k-1} \sum_{h=i+1}^{k} \gamma_{i_l i_h}^{ih} + \varepsilon_j \qquad (10.11)$$

Analoogiliselt võib mudelisse lisada ka kõrgemat järku koosmõjud.

Dispersioonanalüüsi ülesandeks on:

1. Kontrollida mudeli olulisust, st kontrollida hüpoteesipaari

H_0: kõik mudeli parameetrid peale vabaliikme μ võrduvad üldkogumis nulliga;

H_1: mudelis on vähemalt ühel faktoril oluline mõju.

Hüpoteeside kontrollimiseks on tarvis eeldada, et funktsioontunnus rahuldab järgmisi tingimusi:

1° Y on normaaljaotusega;

2° Y dispersioon ei sõltu faktorite mõjudest, st kõigi faktorite kõigi tasemete korral on DY sama.

2. Hinnata mudeli parameetreid.

Viimase ülesande lahendamiseks ei ole eeldusi tunnuse Y jaotuse kohta tarvis teha.

Lisaks nimetatud põhiülesannetele lahendab dispersioonanalüüs veel rea täiendavaid ülesandeid, nende hulka kuuluvad näiteks

- erinevate mudelite võrdlemine,
- funktsioontunnuse keskväärtuste võrdlemine faktorite erinevate tasemete mõjude korral,
- funktsioontunnuse dispersiooni jaotamine osadeks (komponentideks) vastavalt faktori mõjudele jne.

Kui mudelis võetakse arvesse faktorite X_i kõik võimalikud (huvipakkuvad) tasemed, siis nimetatakse dispersioonanalüüsi mudelit fikseerituks e determineerituks (e I tüüpi mudeliks).

Kui aga mingi faktori puhul võimalikust tasemete hulgast tehakse juhuslik valik (näiteks linna 60 koolist valitakse uurimiseks juhuslikult viis kooli), siis nimetatakse dispersioonanalüüsi mudelit juhuslikuks (e II tüüpi mudeliks). Juhusliku mudeli korral ei paku huvi faktorite mõjude α_i, γ_{ig}^{ih} hindamine, kuid hinnatakse funktsioontunnuse dispersioonikomponente vastavalt faktorite toimele.

Dispersioonanalüüsi segamudeli puhul on osa faktoreid juhuslikud (st vastavad juhuslikule mudelile), osa determineeritud.

Kõigepealt asume vaatlema fikseeritud mudelit.

10.5.2. Mitmefaktoriline dispersioonanalüüs *Multifactor Analysis of Variance*. Tellimus. SG-süsteemis kuulub protseduur *Multifactor Analysis of Variance* mahukate protseduuride hulka, mille võimaluste ja tulemuste arv on suhteliselt suur.

See protseduur on allmenüüs *Analysis of Variance* järjestuses teine.

Märgime, et süsteemis SG realiseeritud mitmefaktorilise dispersioonanalüüsi protseduur võimaldab leida mudelist (10.11) mõnevõrra üldisema, nn kovariatsioonanalüüsi mudeli, mis sisaldab lisaks k (diskreetsele) faktorile $X_1, \ldots, X_k$ veel m arvulist argumenttunnust (kovarianti) $Z_1, \ldots, Z_m$:

$$y_j = \mu + \sum_{g=1}^{m} \beta_g z_{gj} + \sum_{i=1}^{k} \alpha_{i_i}^{i} + \sum_{i=1}^{k-1} \sum_{h=i+1}^{k} \gamma_{i_i i_h}^{ih} + \varepsilon_j. \quad (10.12)$$

See mudel sisaldab erijuhuna mitmese regressiooni (siis $\alpha_{i_i}^{i} = 0$, $\gamma_{i_i i_h}^{ih} = 0$ kõigi indeksite korral), samuti ka tavalise dispersioonanalüüsi (siis $\beta_g = 0$, $g = 1, \ldots, m$).

Protseduuri käivitamisel ilmub ekraanile tellimuspaneel, vt jn 10.57.

Multifactor Analysis of Variance

Data: erituiu *funktsioontunnus*

Covariates: pere *pidevad argumenttunnused*

kategoorilised nimed (diskr. arg.)

Factors: A: abielus

B:

C:

D:

E:

F:

G:

H:

I:

J:

K:

L:

Interactions
ABCDEFGHIJKL
*
**
***
****
*****
******
****
****
****
****
****
ABCDEFGHIJKL

Range test: Conf. Int. Confidence level: 95

Jn 10.57

Paneeli pealkirjaks on protseduuri nimi - Mitmene dispersioonanalüüs *Multifactor Analysis of Variance*.

Kõigepealt sisestatakse funktsioontunnuse nimi.

Sellele järgneb võimalus kahe argumenttunnuste rühma sisestamiseks. Esimesena sisestatakse (nimede abil) pidevad argumenttunnused *Covariates*. Neid võib olla null kuni viis, siin võib kasutada põhimõtteliselt kõiki arv- (aga tinglikult ka järjestus-) tunnuseid.

Järgneb faktorite nimede sisestamine. Faktoriteks (diskreetseteks argumentideks) võivad olla suvalised arv- või sümboltunnused, kusjuures oluline on see, et nende tasemete arv oleks küllalt väike. Näiteks meie poolt analüüsitud tunnus teadtoid ei sobi faktoriks, sest sellel on 17 erinevat väärtust (taset) vahemikus 0...48, mis nõuaks 17 parameetri hindamist, seda aga on 28 vaatluse jaoks liiga palju.

Käesolevas näites on sisestatud funktsioontunnus eritulu, pidev argument pere ja üks faktortunnus abielus. Edaspidi tähistatake faktortunnused tähestiku algustähtedega - esimne faktor on A, teine B jne.

Edasi tuleb märkida aknasse *Range Test*, millist meetodit soovitakse kasutada keskmiste võrdlemiseks. Valida on järgmiste võimaluste vahel (kasutades tühikuklahvi):

- usalduspiirid *Conf Int.* (vaikimisi);
- vähima olulise intervalli test *LSD*;
- Scheffe test *Scheffe*;
- Tukey test *Tukey*.

Mitmetes otsustustes, muuhulgas ka keskmiste võrdlemise korral, kasutatakse usaldusnivood. Ka see on üks protseduuri muudetavaid parameetreid (*Confidence level*), vaikimisi väärtuseks on siin, nagu mujalgi, 95 %.

Tellimuspaneeli parempoolses servas on tärnikestest moodustatud kolmnurk (püramiid). Selle abil on võimalik teha valik teist järku koosmõjude seas. Märgime, et kõrgemat järku koosmõjude arvestamist käesolev programm ei võimalda.

Paneme tähele, et tärnikestest moodustatud kolmnurgas on iga tärn määratud kahe koordinaadiga:

- veeru täht (A...K; tähele L ei vasta ühtki tärni);
- rea täht, mis tuleb lugeda paneeli vasakult servalt (B...L).

Tärn reas C ja veerus B tähistab faktori C (kolmas faktor tellimuses) ja faktori B (teine faktor tellimuses) koosmõju.

Kui tellija soovib seda koosmõju hinnata, tuleb tellimuspaneelile sellesse kohta tärnike alles jätta. Kui tellijat nimetatud koosmõju hindamine ei huvita, tuleb tärnikene (tühikuga ületrükkimise teel või klahvi Del abil) kustutada.

Paneme siinjuures tähele veel järgmisi asjaolusid:

- kahe faktori koosmõju saab hinnata üksnes sel juhul, kui nende faktorite tasemete iga kombinatsiooni korral on tehtud vähemalt üks vaatlus;

- koosmõjusid, mis vastavad defineerimata faktoritele, ei arvutata, seetõttu pole vajadust neid paneelil kustutada;

- dispersioonanalüüsi jaoks vajalik mälu maht ja ka arvutamise aeg sõltub oluliselt summaarsest faktorite tasemete arvust, seetõttu võib paljude teist järku koosmõjude lisamisel tekkida mudeli arvutamisel tõrge mälu vähesuse tõttu.

Käesolevas näites jäid kõigile ülejäänud parameetritele süsteemi poolt pakutud vaikimisi väärtused.

10.5.3. Mitmefaktorilise dispersioonanalüüsi tulemused. Üks pidev argument ja üks faktor. Esimene dispersioonanalüüsi tulemus sisaldub standardses dispersioonanalüüsi tabelis, mis on ka esitatud joonisel 10.58.

Analysis of Variance for eritulu

Source of variation	Sum of Squares	d.f.	Mean square	F-ratio	Sig. level
COVARIATES	221045.02	1	221045.02	56.507	.0000
pere	221045.02	1	221045.02	56.507	.0000
MAIN EFFECTS	3305.9626	1	3305.9626	.845	.3764
abielus	3305.9626	1	3305.9626	.845	.3764
RESIDUAL jääk	97794.851	25	3911.7940		
TOTAL (CORR.)	322145.83	27			

0 missing values have been excluded.

Jn 10.58

Esimene veerg - varieeruvuse allikas - on siin organiseeritud hierarhiliselt. Suurtähtedega on trükitud teatava hajuvuse allika (faktorite) tüübi nimetus ja samas reas on esitatud ka kõigi karakteristikute koondnäitajad selle tüübi jaoks.

Iga tüübi nimetuse alla on väiketähtedega trükitud üksikud faktorid, mis sellesse tüüpi kuuluvad, ning ühtlasi neile vastavad näitajad. Viimaste summa annab tüübi vastavad näitajad.

Faktoriteks liigendamist ei toimu ainult kahe viimase tüübi - jääägi (*RESIDUAL*) ja kogusumma *TOTAL (CORR.)* - puhul. Joonisel 10.58 toodud tabelis on mõlemas faktorite tüübis - pidevad argumendid (*COVARIATES*) ja peamõjud (*MAIN EFFECTS*) - ainult üks esindaja, mis tingib lihtsalt arvtulemuste ridade kordumise.

Tabelist joonisel 10.58 näeme, et pere suurus omab eritulule (tulu pereliikme kohta) olulist mõju, st hüpoteesipaarist

$$H_0: \beta = 0,$$
$$H_1: \beta \neq 0,$$

kus β on tunnusele pere vastav parameeter eritulu kirjeldavas mudelis, kummutame nullhüpoteesi ja loeme H_1 tõestatuks. Abielulisusel (faktor A) aga ei ole eritulule mõju, tuleb vastu võtta hüpotees H_0: $\alpha^1 = 0$, kus α^1 on abielulisuse mõju. Seda näeme tabeli viimasest veerust, kus on trükitud olulisuse tõenäosused (*Sig. level*). Esimesel juhul on olulisuse tõenäosus olulisuse nivoost 0,05 väiksem, teisel juhul aga suurem.

Joonisel 10.59 on toodud pideva argumendi toimet iseloomustava regressioonikordaja väärtus

$$b_1 = -76,87,$$

st perekonna suurenedes ühe liikme võrra tulu ühe perekonnaliikme kohta väheneb keskmiselt 76,87 rbl võrra niihästi abielus kui ka vallaliste aspirantide puhul.

COVARIATES	coefficient
pere	-76.869497

Jn 10.59

Märgime siinkohas, et kõigi aspirantide jaoks koos on eritulu lineaarne regressioonimudel kujuga

$$\text{eritulu} = 405.50 - 71,34 \times \text{pere},$$

kusjuures $r = -0,828$, $r^2 = 68,62$ %, vt jn 10.60.

Tabelis jooniselt 10.61 on esitatud keskmised eritulud sõltuvalt vastaja abielulisest staatusest. Vallalistel

Regression Analysis - Linear model: Y = a+bX

Dependent variable: eritulu Independent variable: pere

Parameter	Estimate	Standard Error	T Value	Prob. Level
Intercept	405.504	29.4923	13.7495	.00000
Slope	-71.3432	9.46243	-7.53962	.00000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	Prob. Level
Model	221045.02	1	221045.02	56.8	.00000
Error	101100.81	26	3888.49		
Total (Corr.)	322145.83	27			

Correlation Coefficient = -0.82835 R-squared = 68.62 percent
Stnd. Error of Est. = 62.3578

Jn 10.60

Table of means for eritulu

Level	Count	Average	Stnd. Error (internal)	Stnd. Error (pooled s)	95 Percent Confidence for mean	
abielus						
0	9	257.40741	33.778407	20.848112	214.45972	300.35510
1	19	175.26316	24.109047	14.348655	145.70453	204.82178
Total	28	201.66667	11.819769	11.819769	177.31762	226.01572

Jn 10.61

(neid on 9) on erituluks 257,41 rbl, abieluinimestel aga (neid on 19) 175,26 rbl. On lihtne kontrollida, et tulemus on hästi kooskõlas eritulu keskmise tasemega 201,67:

$$(19 \cdot 175{,}26 + 9 \cdot 257{,}41)/28 = 201{,}66.$$

Selleks, et leida mudelit, mis sisaldaks mõlemat muutujat, peame arvestama, et abieluinimeste keskmine pere suurus on 3,32, vallalistel aga 1,69.

Kasutades leitud b_1 väärtust (vt jn 10.59), tunnuste eritulu ja pere keskmisi vallaliste ja abieluliste jaoks ning standardset valemit

$$a = \bar{y} - b\bar{x},$$

saame:

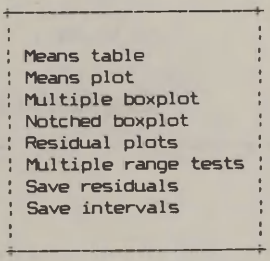
$$a_v = 257.41 - (-76,87)\cdot 1.89 = 402,69,$$
$$a_a = 175,26 - (-76,87)\cdot 3,32 = 430,47,$$

ning järelikult otsitav mudel on järgmine:

vallalistel **eritulu** = 402,69 - 76,87×**pere**,
abielus **eritulu** = 430,47 - 76,87×**pere**.

Siit selgub ka esmapilgul paradoksaalne tulemus tabelis joonisel 10.58, et hoolimata eritulu keskmiste suurest erinevusest vallalistel ja abielus aspirantidel ei ole pereseisu (tunnus **abielus**) mõju oluline: pereseisu mõju on suurel määral tingitud pereliikmete arvu erinevusest, mis aga tõepoolest väga tugevasti mõjutab eritulu.

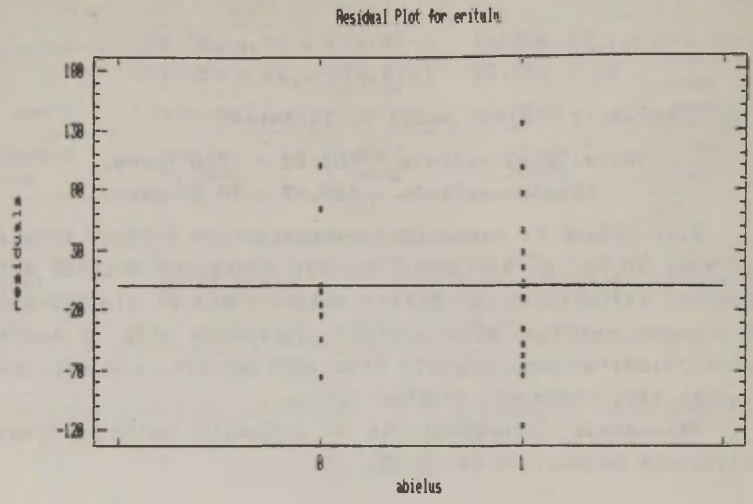
Täiendavat informatsiooni on võimalik saada valikute tulemusena aknast, vt jn 10.62.



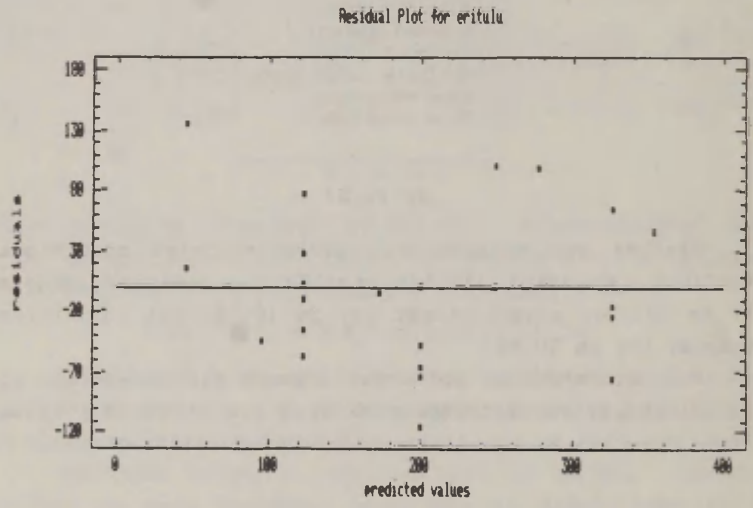
Jn 10.62

Olulise osa täiendavast informatsioonist moodustavad graafikud, sh eriti jääkide analüüs sõltuvalt argumendist (vt jn 10.63), prognoosidest (vt jn 10.64) või järjekorranumbrist (vt jn 10.65).

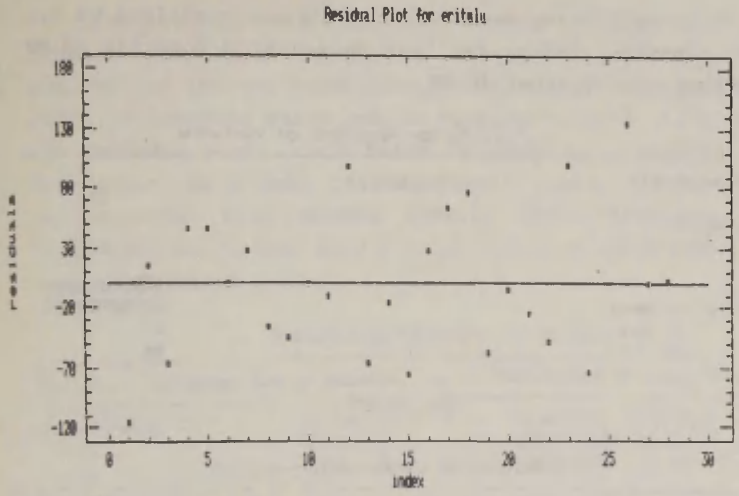
Funktsioontunnuse sõltuvust argumentide tasemetest illustreeritakse ka usalduspiiride abil (vt jn 10.66), võimalik on kasutada ka karpdiagrammi mitmesuguseid variante.



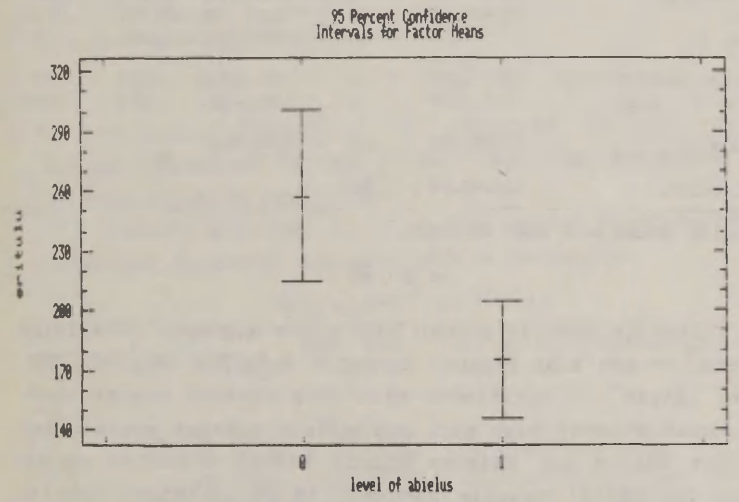
Jn 10.63



Jn 10.64



Jn 10.65



Jn 10.66

10.5.4. Dispersioonanalüüsi tulemused mitme faktori korral. Vaatleme nüüd ülesannet, mille puhul ka pere suurust käsitleme faktorina (st arvestame iga taset eraldi), teiseks

faktoriks aga võtame vastaja soo. Lülitame mudelisse ka tunnuste koosmõju. Vastav tellimus on esitatud joonisel 10.67, tulemused aga joonisel 10.68.

Multifactor Analysis of Variance

Data: eritulu

Covariates:

Factors:		Interactions
A:	pere	ABCDEFGHIJKL
B:	sugu	*
C:		**

Jn 10.67

Analysis of Variance for eritulu

Source of variation	Sum of Squares	d.f.	Mean square	F-ratio	Sig. level
MAIN EFFECTS	257044.46	5	51408.891	15.873	.0000
pere	256428.31	4	64107.076	19.794	.0000
sugu	3457.30	1	3457.305	1.067	.3152
2-FACTOR INTERACTIONS	6804.1537	4	1701.0384	.525	.7186
pere sugu	6804.1537	4	1701.0384	.525	.7186
RESIDUAL	58297.222	18	3238.7346		
TOTAL (CORR.)	322145.83	27			

0 missing values have been excluded.

Jn 10.68

Tulemuste tabelis puudub nüüd pidev argument *COVARIATES*, seevastu on aga kahe faktori koosmõju *2-FACTOR INTERACTIONS*. Paneme tähele, et käsitledes igat pere suurust eraldi tasemena, saab tunnuse pere abil kirjeldada tunnuse **eritulu** hajuvusest rohkem kui eelmise mudeli korral - vastav suurus SS_1 on 256 428,31 tabelis joonisel 10.68, eelmises mudelis (jn 10.58) oli see 221 045,02. Kuid kuna hinnata tuleb nüüd 5 parameetrit (iga üksiku pere suuruse mõju), on vabadusastmete arv 5 - 1 = 4 ning seetõttu keskruut ja sellest tulenevalt ka F-statistiku väärtus peaaegu 4 korda väiksem, ent mõju on siiski tugevasti oluline (P = 0).

Tunnuse sugu mõju ning kahe tunnuse koosmõju on tühine. Viimast tõsiasja arvestades on otstarbekas teha uus tellimus, milles loobuda koosmõjust (kustutada tärn püramiidi tipust). Tulemuseks saame tabeli jooniselt 10.69. Siin on varem koosmõju poolt kirjeldatud hajuvuse osa liidetud jääkhajuvusele, kuid kuna jääkhajuvusele vastav vabadusastmete arv suureneb, siis väheneb jäägile vastav keskruut. Selle tagajärjel suurenevad kõik F-statistikud ja vähenevad olulisuse tõenäosused.

Analysis of Variance for eritulu

Source of variation	Sum of Squares	d.f.	Mean square	F-ratio	Sig. level
MAIN EFFECTS	257044.46	5	51408.891	17.373	.0000
pere	256428.31	4	64107.078	21.664	.0000
sugu	3457.30	1	3457.305	1.168	.2914
RESIDUAL	65101.376	22	2959.1535		
TOTAL (CORR.)	322145.83	27			

0 missing values have been excluded.

Jn 10.69

Viimane tähelepanek on edasiseks oluline: liigsete liikmete mudelist kõrvale jätmine üldiselt suurendab allesjäävate liikmete olulisust.

Tabelis joonisel 10.70 on ära toodud funktsioontunnuse keskmised vastavalt kummagi faktori tasemetele.

Table of means for eritulu

Level	Count	Average	Stnd. Error (internal)	Stnd. Error (pooled s)	95 Percent Confidence for mean	
pere						
1	6	341.66667	26.002137	22.207932	295.59918	387.73415
2	4	306.25000	32.874445	27.199051	249.82908	362.67092
3	8	148.95833	13.947668	19.232633	109.06272	188.85395
4	8	117.50000	17.531859	19.232633	77.60439	157.39561
5	2	120.00000	60.000000	38.465266	40.20877	199.79123
sugu						
1	12	207.08333	28.253843	15.703379	174.50870	239.65796
2	16	197.60417	30.010885	13.599525	169.39371	225.81463
Total	28	201.66667	10.280275	10.280275	180.34156	222.99177

Jn 10.70

Kaks standardvea väärtust (4. ja 5. veerg, vastavalt *Stnd. Error (internal)* ja *Stnd. Error (pooled s)*) on arvutatud erinevalt leitud dispersioonihinnangute järgi:

- esimesel juhul iga üksikrühma jaoks eraldi vastavalt valemile

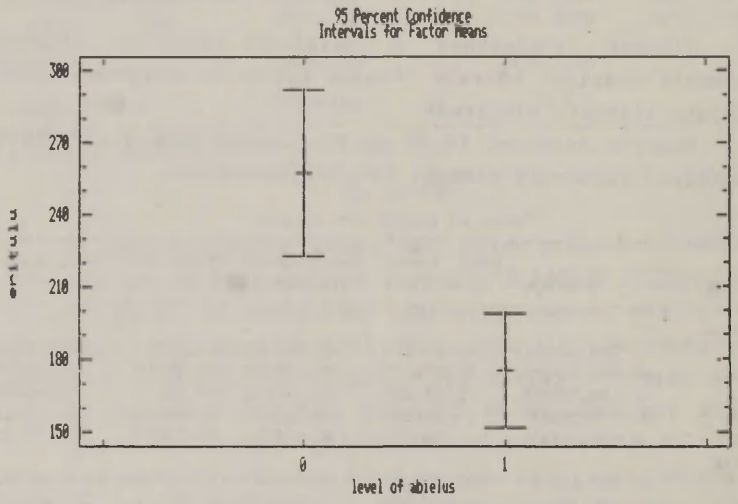
$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2;$$

- teisel juhul üksikrühmade dispersioone ühendades,

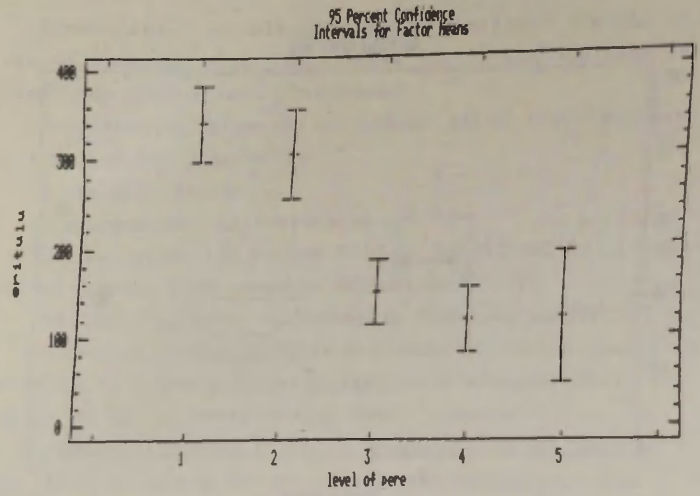
$$s^2 = \frac{1}{n-k} \sum_{i=1}^{k} s_i^2 (n_i - 1).$$

Usalduspiiride arvutamisel on kasutatud teist võimalust (*pooled s*).

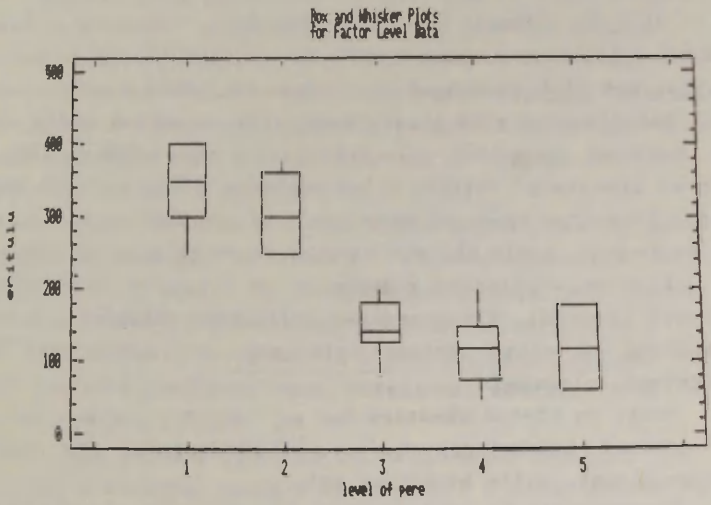
Joonistel 10.71 ja 10.72 on esitatud leitud keskmiste graafikud koos usalduspiiridega, joonistel 10.73 ja 10.74 vastavad mitmesed karpdiagrammid (viimasel juhul koos usalduspiiridega).



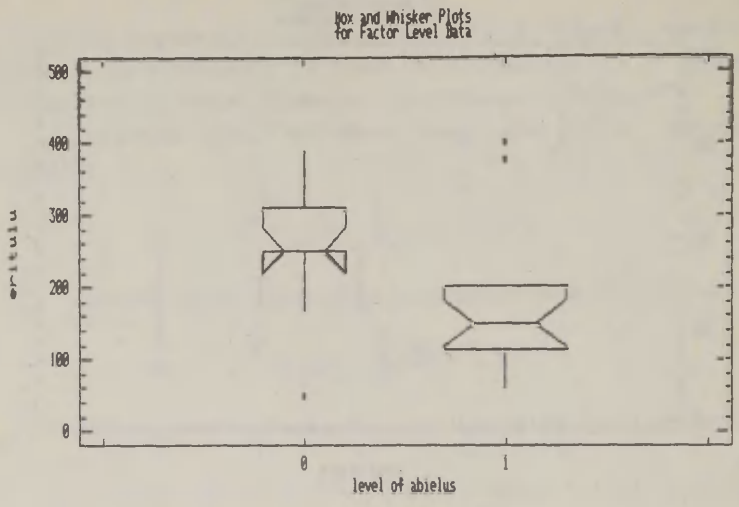
Joon 10.71



Joon 10.72



Joon 10.73



Joon 10.74

10.5.5. Rühmade keskmiste võrdlemine. Dispersioonanalüüsi tulemusena saadakse küll kontrollida hüpoteese selle kohta, kas kõik keskmised on võrdsed või mõned neist erinevad üksteisest/ülejäänutest, kuid seda, millised keskmised on teistest erinevad, ei saa vahetult dispersioonanalüüsi põhjal otsustada. Selleks tuleb keskmisi täiendavalt analüüsida. Nimetatud eesmärgi saavutamiseks on välja töötatud rida meetodeid, mille ühiseks nimetajaks on mitmene võrdlemine.

Mitmese võrdlemise eesmärgiks on leida nn homogeensed klassid (grupid), mis koosnevad sellistest rühmadest, mille keskmised ei erine omavahel oluliselt, st samaväärsed on järgmised tulemused:

vastu on võetud hüpotees H_0: $\mu_{i_1} = \mu_{i_2} = \ldots = \mu_{i_q}$;

rühmad indeksitega $i_1, \ldots, i_q$ moodustavad ühe homogeense klassi, mille tähistame $G(i_1, \ldots, i_q)$.

Homogeensete klasside uurimisel pakuvad alati huvi kõik nn maksimaalsed homogeensed klassid, st sellised rühmade hulgad, millele ei saa enam ühtki rühma nii juurde liita, et tulemus oleks homogeenne.

Arusaadavalt ei ole üldjuhul homogeensed klassid üksteist välistavad; võib isegi leida rühmi, mis kuuluvad kõikidesse homogeensetesse klassidesse.

Homogeensete klasside struktuur antud dispersioonanalüüsi andmestiku puhul sõltub

- usaldusnivoost;
- homogeensuse kriteeriumist.

Mõlemad nimetatud parameetrid on tellitavad tellimuspaneeli allservas (homogeensuse kriteeriume on 4).

Vaatleme järgnevas homogeensete klasside moodustamist.

Järjestame rühmad keskmiste kasvamise järjekorras ning moodustame nn intsidentsusmaatriksi, s.o k×k-maatriksi, mille iga rida ja iga veerg vastab ühele rühmale.

Maatriksi i-ndasse ritta ja g-ndasse veergu märgime +, kui μ_i ja μ_g ei ole erinevad (kehtib nullhüpotees); kui $\mu_i \neq \neq \mu_g$, jääb lahter tühjaks.

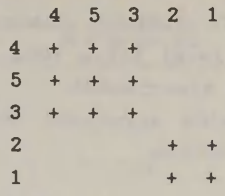
On selge, et maatriks on sümmeetriline, seetõttu vaatleme edaspidi intsidentsusmaatriksist ainult tema alumist kolmnurka.

Iga selle maatriksi veerg kirjeldab üht homogeenset klassi: esimeses veerus on kõik esimesena nimetatud rühmaga statistiliselt võrdsete keskmistega rühmad, teises vastavalt teisena nimetatud rühmaga jne. Üldjuhul ei ole aga kõik selliselt saadud klassid maksimaalsed.

Kui mingis veerus paiknevad +-märgid moodustavad mõnes talle eelnevas veerus paiknevate +-märkide alamhulga, siis ei kirjelda vaadeldav veerg maksimaalset homogeenset klassi. Tõmbame intsidentsusmaatriksist maha kõik veerud, mis ei vasta maksimaalsele homogeensele klassile. Maatriksi järelejäänud veerud kirjeldavad kõiki võimalikke maksimaalseid homogeenseid klasse.

Vaatleme näitena joonisel 10.70 esitatud keskmisi. Järjestamise tulemusena saame jada 4 5 3 2 1 ning moodustame 95 % usalduspiiride järgi intsidentsusmaatriksi, vt jn 10.75.

Pärast ülemise kolmnurga mahatõmbamist näeme, et ristide hulk teises ja kolmandas veerus on esimese veeru ristidehulga alamhulk, mistõttu võime teise ja kolmanda, aga samuti ka viienda veeru maha tõmmata. Jääb kaks homogeenset klassi,



Jn 10.75

mis sisaldavad vastavalt 4., 5. ja 3. ning 2. ja 1. rühma. Täiesti juhuslikult on need klassid teineteist välistavad.

Mitmese võrdlemise allprotseduuri saab tellida, valides aknast joonisel 10.62 rea *Multiple range analysis*.

Tulemus on esitatud joonisel 10.76. Kuivõrd on kasutatud 95 %-usalduspiire, siis on tulemus seesama, mille saime joonisel 10.76 esitatud tabeli analüüsimisel. Intsidentsusmaatriksi kaks veergu, mis vastavad rühmadele 3-4-5 ja 1-2, paiknevad tabelis pealkirja *Homogeneous Groups* all.

Multiple range analysis for eritulu by pere

Method: 95 Percent Confidence Intervals

Level	Count	Average	Homogeneous Groups
4	8	117.50000	*
5	2	120.00000	*
3	8	148.95833	*
2	4	306.25000	*
1	6	341.66667	*

Jn 10.76

Seevastu aga argumendi sugu toimel moodustub üksainus homogeenne klass, vt jn 10.77. Selle põhjuseks on asjaolu, et selle faktori toime kohta võeti dispersioonanalüüsiski vastu nullhüpotees.

Multiple range analysis for eritulu by sugu

Method: 95 Percent Confidence Intervals

Level	Count	Average	Homogeneous Groups
2	16	197.60417	*
1	12	207.08333	*

Jn 10.77

Vaatleme nüüd homogeensete klasside moodustamist, rakendades ka teisi eeskirju ja usaldusnivoosid.

Valime esimesel tellimuspaneelil usaldusnivoo väärtuseks 0,70. Usalduspiiride alusel saame sel juhul samast andmestikust kolm homogeenset klassi (vt jn 10.78).

Multiple range analysis for tulu by pere

Method: 70 Percent Confidence Intervals

Level	Count	Average	Homogeneous Groups
1	6	341.66667	*
3	8	446.87500	**
4	8	470.00000	***
5	2	600.00000	**
2	4	612.50000	*

Jn 10.78

Järgmiseks võtame kasutusele Fisheri vähima olulise erinevuse (*LSD*) meetodi, kasutades jälle 70 % usaldusnivood.

Arvutame vähima olulise erinevuse 2 rühma keskmiste $\bar{y}_i$ ja $\bar{y}_g$ vahel eeskirjaga

$$LSD_{i,g} = t_{p,1-\alpha/2} \sqrt{s^2(n_i^{-1} + n_g^{-1})},$$

$$p = n_i + n_g - 2,$$

kus $t_{p,1-\alpha/2}$ on t-jaotuse $(1-\alpha/2)$-kvantiil vabadusastmete arvul p. Hüpotees H_0: $\mu_i = \mu_g$ võetakse vastu, kui $|\bar{y}_i - \bar{y}_g| < LSD_{i,g}$, see tulemus vastab iga rühmade paari puhul standardsele t-testile, vt p 7.3.1. *LSD* meetodil saame joonisel 10.79 näidatud tulemuse, st tekib 4 homogeenset klassi, kusjuures esimene klass sisaldab ainult esimest rühma.

Multiple range analysis for tulu by pere

Method: 70 Percent LSD Intervals

Level	Count	Average	Homogeneous Groups
1	6	341.66667	*
3	8	446.87500	*
4	8	470.00000	**
5	2	600.00000	**
2	4	612.50000	*

Jn 10.79

Valides Tukey meetodi, saame nn *Tukey HSD* (*honestly significant difference*), mis tähendab 'ausalt olulist vahet'.

Tukey meetodi puhul võetakse lähteks normaaljaotusega N(0, 1) kogumis moodustunud k juhusliku rühma keskpunktide jaotus. Oluline vahe w_k kasutab selle jaotuse (1-α)-kvantiili $q_{k,\nu,1-\alpha}$

$$w_k = q_{k,\nu,1-\alpha} \cdot \sqrt{s^2 (n_i^{-1} + n_g^{-1})},$$

hüpotees H_0: $\mu_i = \mu_g$ võetakse vastu juhul, kui $|\bar{y}_i - \bar{y}_g| < w_k$.

Tukey ja *LSD* meetod annavad sama tulemuse kahe rühma puhul (eeldades loomulikult sama olulisuse nivood). Kui k > 2, siis võib juhtuda, et sama olulisuse nivoo korral *LSD* võimaldab tõestada hüpoteesi H_1: $\mu_i \neq \mu_g$, Tukey meetod aga mitte. Teoreetiliselt on viimane otsustus eelistatav.

Neljas esitatud meetod on Scheffe meetod, mille puhul kasutatakse F-jaotust. Nimelt arvutatakse statistik S,

$$S = \sqrt{(k-1) F_{k-1,\nu,1-\alpha} \cdot s^2 (n_i^{-1} + n_g^{-1})},$$

kus $F_{k-1,\nu,1-\alpha}$ on F-jaotuse (1-α)-kvantiil ja s^2 tähistab ühist dispersioonihinnangut (*pooled*). Hüpotees H_0: $\mu_i = \mu_g$ võetakse vastu juhul, kui $|\bar{y}_i - \bar{y}_g| < S$.

Nagu näha joonistelt 10.80 ja 10.81, moodustub nende meetodite korral ühtviisi 2 homogeenset rühma, mis teineteisest erinevad ainult äärmiste elementide poolest.

Multiple range analysis for tulu by pere

Method: 70 Percent Tukey HSD Intervals

Level	Count	Average	Homogeneous Groups
1	6	341.66667	*
3	8	446.87500	**
4	8	470.00000	**
5	2	600.00000	**
2	4	612.50000	*

Jn 10.80

Multiple range analysis for tulu by pere

Method: 70 Percent Scheffe

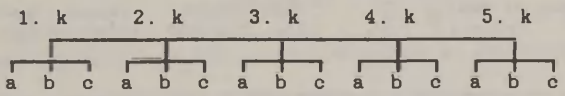
Level	Count	Average	Homogeneous Groups
1	6	341.66667	*
3	8	446.87500	**
4	8	470.00000	**
5	2	600.00000	**
2	4	612.50000	*

Jn 10.81

10.5.6. Dispersioonanalüüsi hierarhiline mudel juhuslike faktorite korral. Hierarhilise mudeli puhul ei ole faktorid samaväärsed, vaid üks neist on allutatud teisele. Toome selle kohta lihtsa näite, kusjuures kasutame juhuslikku mudelit.

Olgu vaadeldud linna 5 kooli 6. klasside õpilaste matemaatikateadmisi. Oletame, et igas koolis on mitu (vähemalt 3) paralleelklassi ning koolide kaupa tahetakse kontrollida ka paralleelklasside erinevusi. Sealjuures oletatakse, et koolide ja klasside hulgast on tehtud juhuslik valik, igast koolist on valimisse võetud 3 paralleelklassi.

Saame järgmise hierarhilise struktuuri.



Jn 10.82

Meid huvitab

- kooli mõju, see on faktor A viie tasemega,
- klassi mõju, see on faktor B viieteistkümne tasemega.

Kuivõrd kasutatakse juhuslikku mudelit, siis hinnatakse mitte mudeli (10.10) parameetreid, vaid funktsioontunnuse dispersiooni komponente.

Loomulikult ei paku erilist huvi võrrelda eraldi kõigi koolide a-klasside õpilasi kõigi koolide b-klasside õpilastega jne, sest erinevates koolides on õpilaste paralleelklasside vahel jaotamise alus erinev. Seega faktor B tavalises

mõttes (millel oleks 3 taset) jääb arvestamata, faktorina B käsitleme me praegu tegelikult faktori B peamõju ning faktorite A ja B koosmõju summat.

Olgu 1. faktoril k taset. Kui igal 1. faktori tasemel on vaadeldud 1 2. faktori taset ning igal tasemete kombinatsioonil on vaatluste arv n (igast klassist valiti juhuslikult 10 õpilast), siis on meil tegemist tasakaalus (*balanced*) hierarhilise mudeliga ning me saame funktsioontunnuse hajuvust iseloomustava ruutude summa $SS = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - n\bar{y}^2$ jaoks järgmise jaotuse:

$$SS = SS_A + SS_{AB} + SS_O,$$

kus n = klm ning kus SS_A iseloomustab esimese faktori (kooli) mõju, SS_{AB} teise, esimesele faktorile allutatud faktori (klassi) mõju ja SS_O on jääkhajuvus. Ruutliikmete vabadusastmete arvud on sel juhul järgmised: SS - n-1, SS_A - k-1, SS_{AB} - k(l-1), SS_O - kl(m-1).

10.5.7. Hierarhilise dispersioonanalüüsi protseduur *Analysis of Nested Designs*. SG-süsteemis on hierarhiline dispersioonanalüüs realiseeritud 2-faktorilise juhusliku mudelina, kusjuures eeldatud on materjali tasakaalustatust.

Cursor at Row: 13 Data Editor Maximum Rows: 12
Column: 3 File: DISPERSI Number of Cols: 3

Row	A	B	Y
1	1	1	1
2	1	1	2
3	1	2	2
4	1	2	3
5	2	1	5
6	2	1	6
7	2	2	6
8	2	2	7
9	3	1	10
10	3	1	11
11	3	2	8
12	3	2	7
13			
14			
Length	12	12	12
Typ/Wth	N/ 7	N/ 7	N/ 7

Jn 10.83

Vastav protseduur *Analysis of Nested Designs* paikneb allmenüüs *J. Analysis of Variance* kolmandal kohal.

Selle protseduuriga tutvumiseks on meil tarvis kasutusele võtta nõuetele vastav andmestik, sest seni uuritud andmetes ei ole tasakaalustatuse nõue täidetud. Moodustame selle 12-objektilise ja 3-tunnuselisena, vt jn 10.83.

Võrdluseks toome kõigepealt selle andmestiku tavalise 2-faktorilise dispersioonanalüüsi, arvestades ka koosmõjusid, vt jn 10.84.

Analysis of Variance for Y

Source of variation	Sum of Squares	d.f.	Mean square	F-ratio	Sig. level
MAIN EFFECTS	99.000000	3	33.000000	66.000	.0001
A	98.666667	2	49.333333	98.667	.0000
B	.333333	1	.333333	.667	.4538
2-FACTOR INTERACTIONS	10.666667	2	5.3333333	10.667	.0106
A B	10.666667	2	5.3333333	10.667	.0106
RESIDUAL	3.0000000	6	.5000000		
TOTAL (CORR.)	112.66667	11			

0 missing values have been excluded.

Jn 10.84

Saame üldhajuvuse SS = 112,667 järgmise jaotuse:

$$SS = SS_A + SS_B + SS_{AB} + SS_O,$$

kus

SS_A = 98.67, 2 vabadusastet, mõju on oluline;

SS_B = 0,33, 1 vabadusaste, mõju ei ole oluline;

SS_{AB} = 10,67, 2 vabadusastet, mõju on oluline;

SS_O = 3,00, 6 vabadusastet.

Teeme nüüd ka analüüsi hierarhilise mudeli alusel, kasutades protseduuri *Analysis of Nested Designs*. Saame järgmise tulemuse, vt jn 10.85. Näeme siin hajuvuskomponente SS, SS_A, SS_B+SS_{AB} ja SS_O (tabeli II veerus), samuti vabadusastmeid (tabeli III veerus) ja keskruute, mis on saadud hajuvuskomponendi jagamisel vabadusastmete arvuga (tabeli IV veerus).

Analysis of Variance - Nested Designs

Response variable: Y

Source of Variation	Sum of Squares	D.F.	Mean Square	Var.Comp.	Percent
Total (Corr.)	112.66667	11			
A	98.666667	2	49.333333	11.417	84.57
B	11.000000	3	3.666667	1.583	11.73
ERROR	3.000000	6	.500000	.500	3.70

Jn 10.85

Hierarhilise skeemi puhul on võimalik leida ka dispersiooni jaotus üksikuteks dispersioonikomponentideks, antud juhul

$$s^2 = s_A^2 + s_{AB}^2 + s_O^2, \qquad (10.13)$$

sest keskruudud (tabelis veerg *Mean Square*) avalduvad nimetatud dispersioonikomponentide lineaarkombinatsioonidena:

$$\frac{SS_O}{n-mk} = s_O^2,$$

$$\frac{SS_{AB}}{k(l-1)} = a_{11}s_O^2 + a_{12}s_{AB}^2,$$

$$\frac{SS_A}{k-1} = a_{21}s_O^2 + a_{22}s_{AB}^2 + a_{23}s_A^2,$$

kus kordajad a_{ij} avalduvad vabadusastmete kaudu. Käesoleval juhul saame võrrandisüsteemi

$$\begin{cases} s_O^2 = 0{,}5 \\ s_O^2 + 2s_{AB}^2 = 3{,}6607 \\ s_O^2 + 2s_{AB}^2 + 4s_A^2 = 49{,}3333 \end{cases},$$

mille lahendiks on veerus *Var.Comp.* märgitud dispersioonikomponentide väärtused $s_A^2 = 11{,}417$, $s_B^2 = 1{,}583$ ja $s_O^2 = 0{,}500$.

Viimases veerus on märgitud variatsioonikomponentide suhtelised osakaalud protsentides $(s_A^2/s^2)\cdot 100\%$, $(s_B^2/s^2)\cdot 100\%$ ja $(s_O^2/s^2)\cdot 100\%$, kus s^2 on summa (10.13).

10.5.8. Dispersioonanalüüs mitmese regressioonanalüüsi protseduuri abil. Dispersioonanalüüsi on võimalik teha ka regressioonanalüüsi protseduuri *Multiple Regression* abil,

kasutades viimases sisalduvaid tunnuseteisendusi **IND**, **CROSS**, **NEST**.

Vaatleme kõigepealt tunnuse teisendusi, mis võimaldavad diskreetse arvtunnuse teisendada faktortunnuste komplektiks.

Joonistel 10.86 ja 10.87 on esitatud 5-väärtuseline tunnus **pere** ja sellest operatsiooni **IND** toimel saadud (5-1)-veeruline maatriks **IND pere**. Näeme, et tunnuse **pere** väikseimale väärtusele 1 vastab maatriksi **IND pere** ainult 0-dest koosnev rida, kõigile ülejäänud väärtustele aga vastab maatriksi rida, kus on üks 1 ja ülejäänud väärtused nullid; väärtus 1 paikneb veerus p-1, kui tunnuse väärtus on suuruselt p-ndal kohal.

```
Variable: ASPIRANT.pere   (length = 28)
-------------------------------------------
( 1) 3   (19) 4
( 2) 5   (20) 3
( 3) 3   (21) 1
( 4) 1   (22) 3
( 5) 1   (23) 2
( 6) 2   (24) 3
( 7) 4   (25) 2
( 8) 4   (26) 5
( 9) 4   (27) 3
(10) 3   (28) 4
(11) 4
(12) 2
(13) 3
(14) 1
(15) 1
(16) 4
(17) 1
(18) 4
-------------------------------------------
```

Jn 10.86

Teisendus **CROSS** seob kaht tunnust; kui ühel neist on l_1 väärtust ja teisel l_2 väärtust, siis avaldis

tunnus1 CROSS tunnus2

defineerib $(l_1-1)\times(l_2-1)$-veerulise maatriksi, mille veergudeks on vastavalt maatriksite **IND tunnus1** ja **IND tunnus2** veergude otsekorrutistena moodustunud veerud. See loob võimaluse koosmõju faktori defineerimiseks.

Variable: WORKAREA.INDPERE (length = 28 4)

(1,1) 0	(1,2) 1	(1,3) 0	(1,4) 0
(2,1) 0	(2,2) 0	(2,3) 0	(2,4) 1
(3,1) 0	(3,2) 1	(3,3) 0	(3,4) 0
(4,1) 0	(4,2) 0	(4,3) 0	(4,4) 0
(5,1) 0	(5,2) 0	(5,3) 0	(5,4) 0
(6,1) 1	(6,2) 0	(6,3) 0	(6,4) 0
(7,1) 0	(7,2) 0	(7,3) 1	(7,4) 0
(8,1) 0	(8,2) 0	(8,3) 1	(8,4) 0
(9,1) 0	(9,2) 0	(9,3) 1	(9,4) 0
(10,1) 0	(10,2) 1	(10,3) 0	(10,4) 0
(11,1) 0	(11,2) 0	(11,3) 1	(11,4) 0
(12,1) 1	(12,2) 0	(12,3) 0	(12,4) 0
(13,1) 0	(13,2) 1	(13,3) 0	(13,4) 0
(14,1) 0	(14,2) 0	(14,3) 0	(14,4) 0
(15,1) 0	(15,2) 0	(15,3) 0	(15,4) 0
(16,1) 0	(16,2) 0	(16,3) 1	(16,4) 0
(17,1) 0	(17,2) 0	(17,3) 0	(17,4) 0
(18,1) 0	(18,2) 0	(18,3) 1	(18,4) 0
(19,1) 0	(19,2) 0	(19,3) 1	(19,4) 0
(20,1) 0	(20,2) 1	(20,3) 0	(20,4) 0
(21,1) 0	(21,2) 0	(21,3) 0	(21,4) 0
(22,1) 0	(22,2) 1	(22,3) 0	(22,4) 0
(23,1) 1	(23,2) 0	(23,3) 0	(23,4) 0
(24,1) 0	(24,2) 1	(24,3) 0	(24,4) 0
(25,1) 1	(25,2) 0	(25,3) 0	(25,4) 0
(26,1) 0	(26,2) 0	(26,3) 0	(26,4) 1
(27,1) 0	(27,2) 1	(27,3) 0	(27,4) 0
(28,1) 0	(28,2) 0	(28,3) 1	(28,4) 0

Jn 10.87

Kaheväärtuseliste tunnuste **sugu** ja **abielus** põhjal moodustunud maatriks (mis on ,-operatsiooni abil jadaks teisendatud, vt [7], p 4.3.16) on esitatud joonisel 10.88. Näeme, et tunnusel **sugu** **CROSS** **abielus** on väärtus 1 parajasti nende objektide korral, millel mõlema lähtetunnuse väärtus on maksimaalne.

```
: ,sugu CROSS abielus
1 1 1 1 1 0 1 1 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
: sugu
2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 2 1 1 1 1 1 1 1 1 1
: abielus
1 1 1 1 1 0 1 1 0 1 1 1 1 0 0 1 0 1 1 0 0 1 0 1 0 1 1 1
:
```

Jn 10.88

Kolmas teisendus **NEST** peaks põhimõtteliselt andma võimaluse hierarhilise mudeli loomiseks, kuid autoritel on tõsiseid kahtlusi selle operatsiooni töö korrektsuses.

Vaatleme ka üht rakenduslikku näidet.

Sisestame regressioonanalüüsi paneelil funktsioontunnuse tulu ja argumenttunnused **IND** pere, teadtoid, vanus. Tulemused on esitatud joonisel 10.89.

Model fitting results for: tulu

Independent variable	coefficient	std. error	t-value	sig.level
CONSTANT	787.178472	136.386532	5.7717	0.0000
IND pere	71.847631	94.303023	0.7619	0.4546
IND pere	9.2569	70.023419	0.1322	0.8961
IND pere	79.532451	68.736254	1.1571	0.2602
IND pere	90.819495	107.909309	0.8416	0.4095
teadtoid	11.552678	2.709366	4.2640	0.0003
vanus	-15.862017	4.287838	-3.6993	0.0013

R-SQ. (ADJ.) = 0.5148 SE= 123.091699 MAE= 93.880125 DurbWat= 1.462
Previously: 0.1290 164.924912 119.464286 2.044
28 observations fitted,forecast(s) computed for 0 missing val.of dep var.

Jn 10.89

Näeme, et siin on hinnatud mudeli parameetreid faktori **IND** pere nelja taseme jaoks (miinimumtase välja arvatud), lisaks sellele on hinnatud veel vabaliiget ning kahe pideva argumendi teadtoid ja vanus kordajaid. Seega saame järgmised mudelid:

1-liikmelises peres

tulu = 787,18 + 11,55×teadtoid - 15,86×vanus,

2-liikmelises peres

tulu = 787,18 + 71,85 + 11,55×teadtoid - 15,86×vanus,

3-liikmelises peres

tulu = 787,18 + 9,26 + 11,55×teadtoid - 15,86×vanus,

4-liikmelises peres

tulu = 787,18 + 79,53 + 11,55×teadtoid - 15,86×vanus,

5-liikmelises peres

tulu = 787,18 + 90,82 + 11,55×teadtoid - 15,86×vanus.

Paneme tähele, et käesoleval juhul on mudelit kui tervikut iseloomustavad parameetrid (tabeli all joonisel 10.89) hinnatud kahes etapis: esimese etapi tulemused on trükisel alt teises reas (*Previously*), selleks esimeseks etapiks loetakse ainult **IND**-teisenduse abil moodustatud uute argumentide sobitamise tulemusi. Mudeli sobitamise lõpptulemused on tabeli all nagu ka ülejäänud puhkudel.

Dispersioonanalüüsi tabel saadud lahenduse kohta on esitatud joonisel 10.90.

Analysis of Variance for the Full Regression

Source	Sum of Squares	DF	Mean Square	F-Ratio	P-value
Model	524984.	6	87497.3	5.77481	.0011
Error	318183.	21	15151.6		
Total (Corr.)	843167.	27			

R-squared = 0.622634 , Stnd. error of est. = 123.092
R-squared (Adj. for d.f.) = 0.514815 Durbin-Watson statistic = 1.46239

Jn 10.90

§ 10.6. Kanooniline analüüs

10.6.1. Kanoonilise analüüsi eesmärk. Kanooniline analüüs on mitmemõõtmelise analüüsi protseduur, mis kasutab faktoranalüüsile lähedasi matemaatilisi meetodeid, sisuliselt on aga mõneti sarnane regressioonanalüüsiga.

Olgu antud kaks (arvuliste) tunnuste rühma $X_1, \ldots, X_k$ ja $Y_1, \ldots, Y_q$. Lihtsuse mõttes oletame, et $EX_i = EY_i = 0$, $DX_i = DY_j = 1$, $i = 1, \ldots, k$, $j = 1, \ldots, q$.

Me tahame leida tunnusrühmade vahelise võimalikult tugeva lineaarse seose järgmises mõttes. Olgu U esimese tunnusrühma lineaarteisenduse tulemusena saadud tunnus,

$$U = \sum_{i=1}^{k} a_i X_i,$$

ja V - teise tunnusrühma lineaarteisenduse tulemus,

$$V = \sum_{j=1}^{q} b_j Y_j.$$

Meie eesmärgiks on leida kordajad a_i ja b_j nii, et lineaarne korrelatsioonikordaja $\rho = r(U, V)$ saavutaks maksimaalse võimaliku väärtuse. Sel juhul võetakse kasutusele järgmised terminid:

ρ on kanooniline korrelatsioon,

U ja V (eeldusel, et nad rahuldavad tingimust $DU = DV = 1$) on kanoonilised komponendid.

Paneme tähele, et erijuhul, kui üks rühmadest (näiteks II) sisaldab ainult üht tunnust, siis $V = Y_1/\sqrt{DY_1}$ ning U on võrdeline Y_1 parima lineaarse (homogeense) prognoosiga, erinedes sellest ainult teguri $1/\sqrt{DY_1}$ poolest, seega sel juhul kanooniline analüüs taandub regressioonanalüüsiks.

Kui aga mõlemas rühmas on üle ühe tunnuse, siis me võime protseduuri jätkata. Tähistades leitud suurused vastavalt U_1, V_1 ja ρ_1, püstitame üldisema ülesande:

Olgu leitud h kanooniliste komponentide paari $U_1, V_1; \ldots; U_h, V_h$, mis määravad kanoonilised korrelatsioonid $\rho_1, \ldots, \rho_h$,

$$\rho_i = r(U_i, V_i),$$

kusjuures on täidetud järgmised tingimused:

1° kõik kanoonilised komponendid on normeeritud,

$$DU_i = DV_i = 1, \quad i = 1, \ldots, h;$$

2° sama tunnusrühma põhjal leitud kanoonilised komponendid on ortogonaalsed (mittekorreleeritud), st

$$EU_iU_j = 0, \quad i \neq j; \qquad EV_iV_j = 0, \quad i \neq j.$$

Tähistame

$$p = \min(k, q)$$

Kui $h < p$, siis seame enesele ülesandeks leida $(h+1)$-ne kanooniliste komponentide paar U_{h+1}, V_{h+1}, mis rahuldaks tingimusi 1° ja 2° ning mille puhul

$$\rho_{h+1} = r(U_{h+1}, V_{h+1})$$

oleks maksimaalne võimalik korrelatsioon.

Arusaadavalt on kanoonilisi komponente ja ka korrelatsioone võimalik leida ülimalt p tükki, kusjuures kehtib võrratuste jada

$$\rho_1 \geq \rho_2 \geq \ldots \geq \rho_p.$$

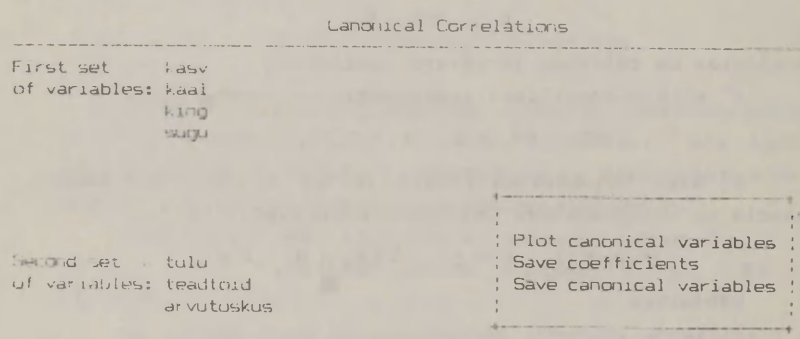
Märgime, et kanooniline analüüs (nii nagu peakomponentide ja faktoranalüüski) taandub omaväärtusülesande lahendamisele, nimelt on $\rho_i = \sqrt{\lambda_i}$, kus suurused λ_i ($i = 1, \ldots, p$) on maatriksi

$$R_1^{-1/2} R_{12} R_{22}^{-1} R_{21} R_1^{-1/2} \qquad (10.14)$$

omaväärtused (järjestatuna kahanevalt). Sama maatriksi omavektorite põhjal leitakse ka kanoonilised komponendid.

10.6.2. Kanooniline analüüs protseduuri *Canonical Correlations* abil. Protseduur *Canonical Correlations* paikneb allmenüüs *Q. Multivariate Methods* kaheksandal kohal.

Selle protseduuri tellimust kirjeldab joonis 10.91, kuhu on ruumi kokkuhoiu mõttes lisatud ka aken täiendavate võimaluste valimiseks (selle akna saamiseks tuleb pärast esmaste tulemuste leidmist vajutada klahvi F5).



Jn 10.91

Kanoonilise analüüsi tulemuste tabel on esitatud joonisel 10.92. Tabel koosneb kolmest osast, millest kõige ülemine sisaldab kanoonilist analüüsi tervikuna iseloomustavaid näitajaid. Esimeses veerus on siin järjekorranumber h, mille maksimaalväärtuseks on p. Esitatud näites p = min (3,4) = 3,

sest esimene vaadeldav tunnusrühm sisaldab 4. teine 3 tunnust !B! Esimeses rühmas ei või olla vähem tunnuseid kui teises!

Canonical Correlations

Number	Eigenvalue	Canonical Correlation	Wilks Lambda	Chi-Square	D.F.	Sign. Level
1	.4773	.6908	.4039	20.850	12	.0526
2	.2159	.4646	.7727	5.931	6	.4310
3	.0145	.1206	.9855	.337	2	.8450

Coefficients for Canonical Variables of the First Set

kasv	1.81412	0.86520	0.93131
kaal	-0.66559	-0.53821	0.26068
king	-1.81046	-0.57875	0.76952
sugu	-0.19591	-1.09233	1.59678

Coefficients for Canonical Variables of the Second Set

tulu	0.62791	1.28881	0.07233
teadtoid	-1.20941	-0.34177	0.30891
arvutoskus	-0.03235	-0.77102	0.91776

Jn 10.92

Järgmises kahes veerus on maatriksi (10.14) omaväärtused λ_h ja ruutjuured nendest - kanoonilised korrelatsioonikordajad ρ_h,

$$\rho_h = \sqrt{\lambda_h}.$$

Järgneb Wilksi Λ-statistik, $\Lambda = \prod_{i=h}^{p}(1-\rho_i^2)$, kontrollimaks järgmist hüpoteesipaari selle kohta, et kanoonilised korrelatsioonid on nullid:

$$H_0: \rho_h = \rho_{h+1} = \ldots = \rho_p = 0,$$
$$H_1: \max_{h \le i \le p} \rho_i > 0.$$

Kui mingi h väärtuse korral võetakse vastu nullhüpotees, siis ei ole mõtet kanoonilist analüüsi jätkata, sest kanoonilised komponendid U_i, V_i, i = h, ..., p, ei oma mõtet. Traditsiooniliselt kasutatakse Wilksi Λ jaotuse lähendamist χ^2-jaotuse abil:

$$\chi^2 \approx -m \ln \Lambda, \quad m = n - 0{,}5(f_1 - f_2 + 1),$$

kus f_1 ja f_2 avalduvad X- ja Y-tunnuste arvude k ja q ning järjekorranumbri h kaudu, $f_1 = k - h + 1$, $f_2 = q - h + 1$, ning vabadusastmete arv on $f_1 f_2$.

Kui vaatleme konkreetset näidet, siis osutub, et rangelt võttes ei olegi olulisuse nivool 0,05 tabelis joonisel 10.92 toodud kanooniline analüüs mõttekas, sest juba h = 1 korral p = 0,0526 > 0,05.

Järgmistes ridades esitatakse parameetrid, mis iseloomustavad järgmisi kanoonilisi kordajaid ja komponente. Näeme, et järjekorranumbri suurenedes

- kanooniline korrelatsioon väheneb;
- vabadusastmete arv väheneb (pärast 1 kanoonilise korrelatsiooni arvutamist on see (k-1)(q-1));
- olulisuse tõenäosus (arvutatakse vastavalt χ^2-jaotusele) suureneb.

Tabeli järgmine osa sisaldab kanooniliste kordajate a_{ij} maatriksit, kus i = 1, ..., k, j = 1, ..., p. Need kordajad määravad (veergude kaupa) kanoonilised komponendid U_j,

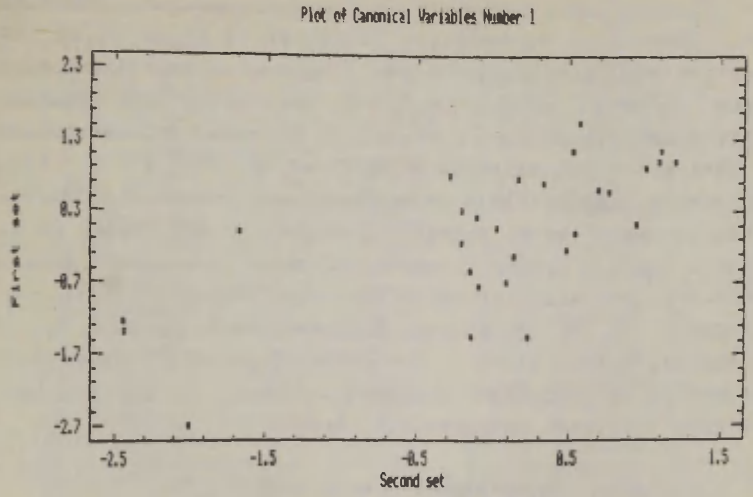
$$U_j = \sum_{i=1}^{k} a_{ij} X_i, \quad j = 1, \ldots, p.$$

Esimeses veerus on siin tunnuste nimetused, järgneb p veergu kanooniliste komponentide kordajatega, kusjuures kanoonilise komponendi number on ühe võrra väiksem veeru numbrist.

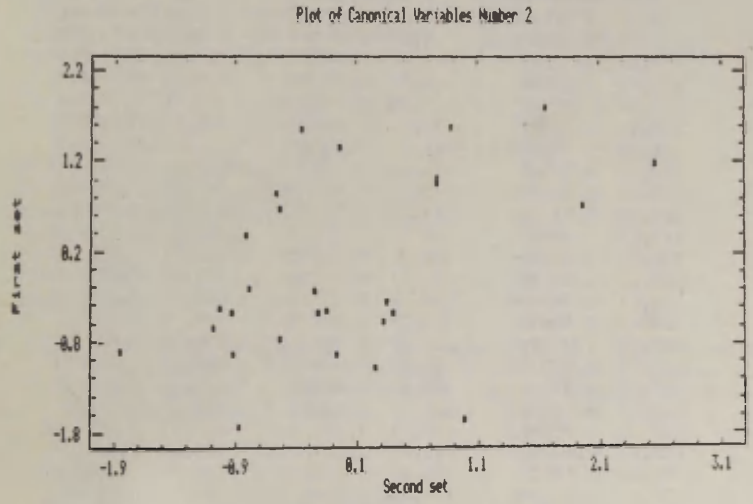
Samal viisil on tabeli järgmises osas toodud kordajate b_{ij} maatriks, i = 1, ..., q, j = 1, ..., p.

Valides aknast (vt jn 10.91) esimese rea, saame tellida meid huvitava kanooniliste komponentide paari jaoks korrelatsioonivälja. Joonisel 10.93 on kujutatud esimeste kanooniliste komponentide U_1 (*First set*)ja V_1 (*Second set*) korrelatsiooniväli. Kuivõrd mõlemad komponendid on standardiseeritud, muutuvad nende väärtused põhiliselt vahemikus -3...3. Jooniselt ilmneb, et arvestatava tugevusega korrelatsioon tunnuste vahel on tingitud eeskätt mõnest objektist, millel mõlemad kanoonilised komponendid on oluliselt negatiivsed.

U_2 ja V_2 korrelatsiooniväli on kujutatud joonisel 10.94 Et vastav kanooniline kordaja on mitteoluline, ei vääri ka need kanoonilised komponendid lähemat vaatlemist.



Jn 10.93



Jn 10.94

Kanooniliste komponentide (kui uute tunnuste) väärtused objektide kaupa on esitatud joonistel 10.95 ja 10.96. **SG**-süsteem väljastab kanoonilised komponendid maatriksmuutujatena. Selleks, et neid edaspidi üksiktunnustena kasutada, võib rakendada näiteks joonisel 10.97 toodud **SG** operatsioone (vastavate operatsioonide selgitused vt [7], § 4.3). Selle tulemusena saadud üksikute kanooniliste komponentide korrelatsioonimaatriks on esitatud joonisel 10.98. Näeme, et see vastab täpselt teoreetilistele ootustele - esimesele kolmele tunnusele ja teisele kolmele tunnusele vastavad blokid (peadiagonaalil) on tõepoolest ühikmaatriksid ja R(**U**, **V**), st diagonaaliväline blokk on omakorda diagonaalne, diagonaalil paiknevad kanoonilised korrelatsioonid ρ_1, ρ_2, ρ_3. Veendume, et kõik arvutused on piisavalt täpsed.

Variable: WORKAREA.VARS1 (length = 28 3)

(1,1) 0.771094	(1,2) -0.445597	(1,3) -1.60552
(2,1) 0.301141	(2,2) -0.904049	(2,3) -1.20746
(3,1) -0.234152	(3,2) -1.07517	(3,3) -0.97994
(4,1) 0.591967	(4,2) -0.564011	(4,3) -0.875496
(5,1) 0.711676	(5,2) -0.624079	(5,3) -0.494381
(6,1) 0.130276	(6,2) -0.910914	(6,3) -0.0348001
(7,1) 1.01877	(7,2) -0.454185	(7,3) -0.400659
(8,1) 0.79058	(8,2) -0.455408	(8,3) -0.0794131
(9,1) 0.20918	(9,2) -0.742243	(9,3) 0.380168
(10,1) 1.16832	(10,2) -0.228391	(10,3) -0.0133577
(11,1) -2.74357E-3	(11,2) -0.913611	(11,3) 0.463169
(12,1) -1.24445	(12,2) -1.65595	(12,3) 0.967364
(13,1) 1.14675	(13,2) -0.402705	(13,3) 0.423092
(14,1) 1.54076	(14,2) -0.345832	(14,3) 0.250903
(15,1) -1.44811	(15,2) -1.71577	(15,3) 1.493
(16,1) -0.32842	(16,2) 0.964389	(16,3) -1.74778
(17,1) -1.46138	(17,2) 0.676581	(17,3) -0.822435
(18,1) -0.744004	(18,2) 0.733204	(18,3) -1.13915
(19,1) -1.39604	(19,2) 0.389246	(19,3) -0.651897
(20,1) 1.00041	(20,2) -0.468165	(20,3) 1.80015
(21,1) -0.545392	(21,2) 0.841807	(21,3) -0.343062
(22,1) -2.69186	(22,2) -0.18712	(22,3) 0.765234
(23,1) 0.0117278	(23,2) 1.58048	(23,3) 0.11648
(24,1) 0.0525331	(24,2) 1.35174	(24,3) 0.0826268
(25,1) -0.694683	(25,2) 1.00925	(25,3) 0.393151
(26,1) -0.15939	(26,2) 1.18037	(26,3) 0.165628
(27,1) 0.903186	(27,2) 1.8043	(27,3) 0.391462
(28,1) 0.602262	(28,2) 1.56183	(28,3) 2.70293

Jn 10.95

```
Variable: WORKAREA.VARS2   (length = 28 3)
-------------------------------------------------------------------------
( 1,1)  0.158934    ( 1,2) -0.142507    ( 1,3) -1.71884
( 2,1) -0.214544    ( 2,2) -1.85751     ( 2,3)  0.73908
( 3,1)  0.476905    ( 3,2)  0.2494      ( 3,3) -0.836967
( 4,1)  0.758571    ( 4,2)  0.328998    ( 4,3) -0.908911
( 5,1)  0.328553    ( 5,2) -1.07513     ( 5,3)  0.732046
( 6,1)  0.938307    ( 6,2) -0.915216    ( 6,3)  1.47355
( 7,1)  1.19974     ( 7,2) -0.212435    ( 7,3)  1.53846
( 8,1) -0.291317    ( 8,2) -0.923449    ( 8,3)  0.0589722
( 9,1) -0.112614    ( 9,2) -0.533696    ( 9,3) -1.71532
(10,1)  1.10585     (10,2) -0.238967    (10,3)  1.56244
(11,1)  0.533445    (11,2) -0.0614735   (11,3) -0.0200055
(12,1) -2.4378      (12,2)  0.985571    (12,3)  1.13398
(13,1)  0.51633     (13,2) -1.02206     (13,3)  0.684083
(14,1)  0.559903    (14,2)  0.348146    (14,3) -1.74224
(15,1) -0.149963    (15,2) -0.871102    (15,3) -0.874379
(16,1)  0.12754     (16,2)  0.767172    (16,3)  0.281201
(17,1)  0.226364    (17,2) -0.525594    (17,3) -0.0205816
(18,1) -0.100704    (18,2)  1.96047     (18,3)  0.602872
(19,1) -2.43441     (19,2) -0.812485    (19,3) -1.04327
(20,1)  1.08666     (20,2)  0.409313    (20,3) -0.0954664
(21,1) -0.16008     (21,2) -0.05951     (21,3) -0.805952
(22,1) -2.00154     (22,2) -0.77785     (22,3)  0.62749
(23,1) -1.67657     (23,2)  0.886239    (23,3)  0.8737
(24,1)  0.0125214   (24,2) -0.0390579   (24,3) -0.751274
(25,1)  0.0811176   (25,2)  0.766454    (25,3) -0.604188
(26,1) -0.214827    (26,2)  2.55713     (26,3)  0.763708
(27,1)  0.999768    (27,2)  1.65496     (27,3) -0.707147
(28,1)  0.683873    (28,2) -0.345815    (28,3)  0.772975
-------------------------------------------------------------------------
```

Jn 10.96

```
: W1 GETS 28 1 TAKE VARS1
: W3 GETS 28 -1 TAKE VARS1
: WW2 GETS 28 2 TAKE VARS1
: W2 GETS 28 -1 TAKE WW2
: Z1 GETS ,W1
: Z2 GETS ,W2
: Z3 GETS ,W3
: Z1
0.771094 0.301141 -0.234152 0.591967 0.711676 0.130276 1.01877 0.79058
      0.20918 1.16832 -2.74357E-3 -1.24445 1.14675 1.54076 -1.44811
      -0.32842 -1.46138 -0.744004 -1.39604 1.00041 -0.545392 -2.69186
      0.0117278 0.0525331 -0.694683 -0.15939 0.903186 0.602262
```

Jn 10.97

Sample Correlations

	Z1	Z2	Z3	Z4	Z5	Z6
Z1	1.0000	.0000	.0000	.6908	.0000	.0000
	(28)	(28)	(28)	(28)	(28)	(28)
	.0000	1.0000	1.0000	.0000	1.0000	1.0000
Z2	.0000	1.0000	.0000	.0000	.4646	.0000
	(28)	(28)	(28)	(28)	(28)	(28)
	1.0000	.0000	1.0000	1.0000	.0127	1.0000
Z3	.0000	.0000	1.0000	.0000	.0000	.1206
	(28)	(28)	(28)	(28)	(28)	(28)
	1.0000	1.0000	.0000	1.0000	1.0000	.5411
Z4	.6908	.0000	.0000	1.0000	.0000	.0000
	(28)	(28)	(28)	(28)	(28)	(28)
	.0000	1.0000	1.0000	.0000	1.0000	1.0000
Z5	.0000	.4646	.0000	.0000	1.0000	.0000
	(28)	(28)	(28)	(28)	(28)	(28)
	1.0000	.0127	1.0000	1.0000	.0000	1.0000
Z6	.0000	.0000	.1206	.0000	.0000	1.0000
	(28)	(28)	(28)	(28)	(28)	(28)
	1.0000	1.0000	.5411	1.0000	1.0000	.0000

Coefficient (sample size) significance level

Jn 10.98

10.6.3. Kanooniliste komponentide tõlgendamine. Loomulik on kanoonilisi komponente ka sisuliselt tõlgendada. Tuleb märkida, et siin on oluline erinevus võrreldes faktorite tõlgendamisega - nimelt tabelis (jn 10.92) antud kordajad ei ole korrelatsioonikordajad. Tunnuste X_i, Y_j ja komponentide U_h, V_g korrelatsioonikordajate arvutamine on aga täiesti teostatav, kasutades leitud maatriksit A ja tunnusrühma **X** korrelatsioonimaatriksit R_x:

$$R(\mathbf{X}, U) = R_x A.$$

Samuti on arvutatavad ka teiste kanooniliste komponentide korrelatsioonid nende aluseks olevate tunnustega:

$$R(\mathbf{Y}, V) = R_y B.$$

Arvutame vastavad korrelatsioonikordajad esimese kanooniliste komponentide paari jaoks (s.o maatriksite $R(\mathbf{X}, U)$ ja $R(\mathbf{Y}, V)$ esimesed veerud):

	U_1		V_1
kasv	-0,186	**tulu**	-0,052
kaal	-0,677	**teadtoid**	-0,861
king	-0,541	**arvutoskus**	0,294
sugu	0,461		

Nende korrelatsioonikordajate abil saame kanoonilisi komponente tõlgendada samuti kui faktoreid.

Teemegi seda:

U_1 +	U_1 −
väike kaal	suur kaal
väike king	suur king
naine	mees
. . . . .	. . . . .
kasv keskmine	kasv keskmine

V_1 +	V_1 -
vähe teadustöid	palju teadustöid
. . . . .	. . . . .
arvutioskus pigem üle keskmise	arvutioskus pigem alla keskmise

Tekib mulje, et U_1 iseloomustab ühest küljest noori naisi, teisest küljest suhteliselt vanemaid mehi (arvestusega, et keskmiselt nooremad on pikemad). Faktor V_1 iseloomustab peamiselt teadustööde rohkust, samal ajal seostub see arvutioskuse madalama tasemega. Korrelatsioon nende faktorite vahel kõneleb sellest, et aspirantide seas (noored) saledad naised paistavad silma suhteliselt madala teadusproduktsiooni ja veidi parema arvutioskusega (faktorite plusspooled), seevastu kogukamad mehed on produktiivsemad, kuid see ei seostu kõrgema arvutioskuse tasemega.

Juhime tähelepanu veel sellele, et esimene kanooniline korrelatsioon on kindlasti suurem kui ükski korrelatsioonikordaja kummagi rühma tunnuste vahel. Veendumaks selles kirjutame välja vastava korrelatsioonimaatriksi fragmendi ja kriipsutame alla statistiliselt olulised korrelatsioonid (olulisuse nivool 0,05). Ühtlasi meenutame, et esimese kanoonilise korrelatsiooni väärtus on 0,69.

	kasv	**kaal**	**king**	**sugu**
tulu	0,330	0,199	0,314	-0,328
teadtoid	0,276	0,492	0,472	-0,446
arvutoskus	-0,042	-0,116	-0,130	0,177

Üldiselt on käesolevas näites kahe tunnuste rühma vahelised seosed üsna nõrgad, 12 seosekordajast on statistiliselt olulisi (nivool 0,05) kõigest 3. Kõrgeim korrelatsioon 0,492 annab 24,2 %-lise kirjeldatuse, kanoooniline korrelatsioon aga peaaegu kaks korda kõrgema, 47,6 %-lise kirjeldatuse.

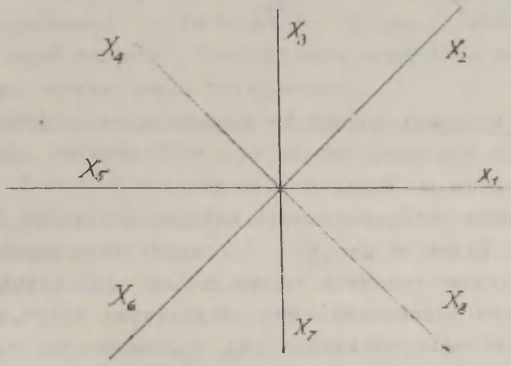
Samal viisil oleks põhimõtteliselt võimalik interpreteerida ka järgmisi kanoonilisi komponente, kuid arvestades nende mitteolulisust, ei ole sellel erilist mõtet.

11. OBJEKTIHULGA ANALÜÜS

§ 11.1. Paljutunnuseliste üksikobjektide piltlik esitus

Kahe ja kolme tunnuse abil kirjeldatavaid objekte saab silmale hästi haaratavalt esitada mitmesugustel graafikutel ja diagrammidel, kus telgedeks on vastavate tunnuste väärtuste skaalad. Rohkem kui kolme tunnuse puhul muutub niisuguse esituse võimalikkus juba üsna küsitavaks, kuid mitmesuguseid graafilisi pilte, mis võimaldaksid silmaga haarates teha otsustusi objektide sarnasuse ja erinevuse kohta, on välja töötatud mitmeid. Mõnega nendest tutvume ka järgnevas punktis.

11.1.1. Objektide esitamine hulknurk-kujunditena. Üks võimalus paljutunnuseliste objektide visuaalseks kujutamiseks on järgmine. Vastaku igale tunnusele telg (kiir), kusjuures teljed on paigutatud keskpunkti ümber korrapärase kimbu (tähe) kujuliselt, vt jn 11.1.

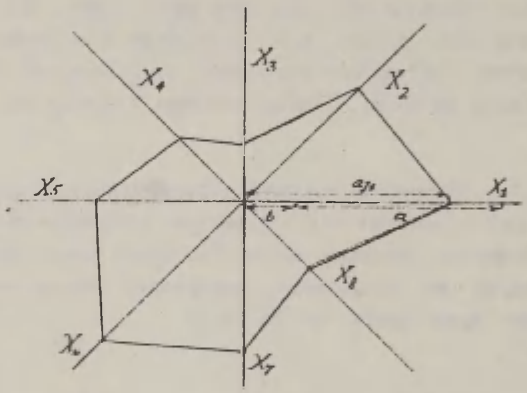


Jn 11.1

Olgu meil antud üks konkreetne objekt (järjekorranumbriga j), millele vastavad uuritavate tunnuste väärtused x_{j1}, x_{j2}, ... x_{jk}. Seda objekti iseloomustavad lõigud a_{ji} telgedel x_i (i = 1, ..., k), lõikude pikkused arvutatakse vastavalt valemile

$$a_{ji} = \frac{x_{ji} - \min x_i}{\max x_i - \min x_i} \cdot a + b,$$

kus a ja b on sobivalt valitud konstandid. Ühendanud lõikude a_{ji} otspunktid omavahel, saame k-nurga (kus k on kasutatud tunnuste arv), mis iseloomustab j-ndat objekti (vt jn 11.2). Arusaadavalt vastavad kahele ligikaudu ühesuguste väärtustega tunnusele ka ligikaudu ühetaolised kujundid.



Jn 11.2

11.1.2. Hulknurk-kujundite moodustamise protseduur *Star Symbol Plot*. Hulknurk-kujundite moodustamise protseduur paikneb allmenüüs *Q. Multivariate methods* 9. kohal.

Protseduuri tellimuspaneeli kujutab joonis 11.3. Paneeli täitmiseks tuleb kõigepealt (lahtrisse *Data vectors*) märkida kasutatavate tunnuste nimed selles järjestuses, nagu neid soovitakse kujunditel näha. Siinjuures tuleb soovitada seda, et sisuliselt lähedased (ka tugevasti korreleeritud) tunnused tuleks paigutada loetelusse kõrvuti, see teeb kujundid ülevaatlikumaks.

Star Symbol Plot

Data vectors: kasv&king&kaal&vanus&teadtoid&tul&pere

Labels: synnipaik

Shortest ray: 0.1

Jn 11.3

Joonisel 11.3 on kasutatud tunnuste nimede eraldamiseks sümbolit &, nii on kirjapilt lühem. Kasutada võib ainult arvtunnuseid, nõutav on nende võrdne pikkus.

Tunnuste arv peab olema vähemalt kolm, maksimumpiiri määrab kujundite ülevaatlikkus: enamasti pole üle 7-8 tunnuse mõtet korraga vaadelda.

Objektidele on võimalik omistada ka tekstiline märgend, mis peab olema salvestatud ülejäänud tunnustega sama pikkusega sümboltunnusena. Selle asemel võib aga allkirjana kasutada mingit muud tunnust, mis küll objekte ei klassifitseeri, vaid iseloomustab neid täiendavalt.

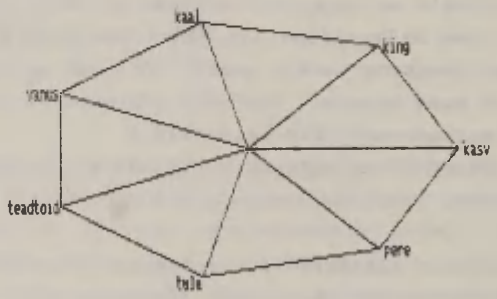
Tellimusse tuleb märgendit sisaldava tunnuse nimi trükkida lahtrisse *Labels*. Joonisel 11.3 on selleks tunnus **synnipaik**.

Tellitavaks parameetriks on tunnuse miinimumile ja maksimumile vastavate lõikude suhe $\frac{a}{a+b}$, s.o arv 0...0.5, mis tuleb trükkida kohale *Shortest ray* (lühim kiir). Väga hästi sobib kasutamiseks vaikimisi väärtus 0.1.

Käivitamisele järgneb kujundite moodustamine vastavalt andmestikus olevate objektide arvule alates esimesest (vasakus allnurgas). Juba jooniselt 11.4 selgub, et rohkem kui 50 objekti puhul muutub joonis üsna ebaülevaatlikuks.

Jn 11.4

Pärast ekraanipildi (jn 11.4) kustutamist võtmega **ESC** ilmub ekraanile aken tekstiga *Plot key*: selle operatsiooni valimisel ilmub ekraanile kujundite võti, vt jn 11.5, mis identifitseerib kõik kasutatavad teljed. On kasulik meeles pidada, et esimene telg paikneb "kella 3" kohal, st on horisontaalne ja paremale suunatud.

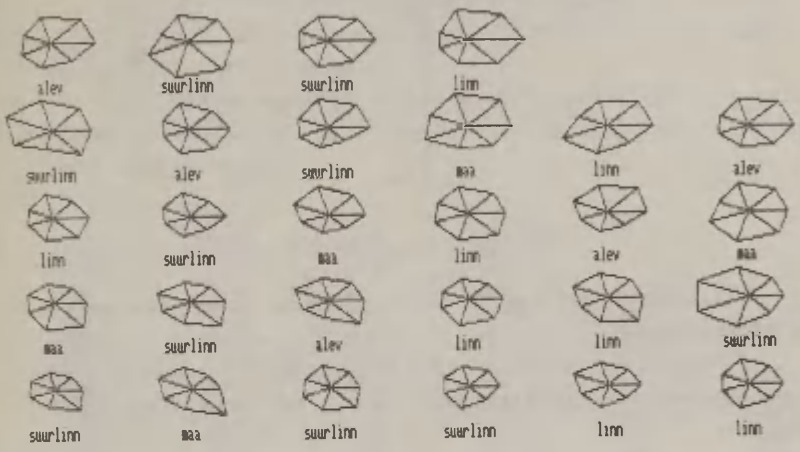


Jn 11.5

Joonist 11.4 silmitsedes näeme, et sarnaseid kujundeid leida on siit üsna raske, võib-olla on suhteliselt ühesugused väikesemõõdulised kujundid kõige alumises reas (peale teise, mille puhul pere on eriti suur) ja kolmanda rea al-

guses. Küll aga torkavad silma üksikud ülejäänutest tugevasti erinevad (suured) kujundid.

Seda, et miinimumile ja maksimumile vastavate lõikude suhe määrab oluliselt kujundite ilmekuse, demonstreerib joonis 11.6, kus on kujutatud täpselt sama andmestikku, ainult parameetri *Shortest ray* väärtuseks on valitud 0,5.



Jn 11.6

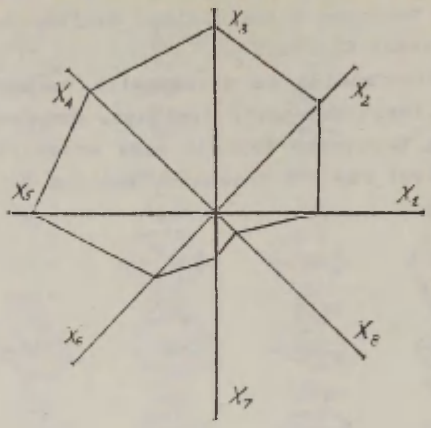
11.1.3. Päikesekiir-kujundite moodustamise protseduur *Sun Ray Plot*. Eelnevaga analoogilise võimaluse kujutada üksikobjekte graafiliste kujunditena pakub protseduur *Sun Ray Plot* (10. protseduur allmenüüs *Q. Multivariate Methods*).

Erinevused eelmisest meetodist on järgmised.

1. Üksiktunnuste skaalad on määratud mitte vastava tunnuse minimaal- ja maksimaalväärtuste kaudu, vaid σ-ühikutes, kusjuures skaala pikkus on tellija poolt määratav (näiteks -3σ...3σ). Seega kõigi tunnuste skaalad on ühesugused.

2. Skaala miinimumpunktile vastava lõigu pikkus on 0.

3. Kõik teljed trükitakse välja terves ulatuses, kuigi murdjoonega ühendatakse iga konkreetse objekti puhul seda iseloomustavad punktid teljel. Seega erinevalt eelmisest graafikust moodustuvad hulknurkse kujundi ümber veel "päikesekiired", mis aitavad hinnata iga konkreetse väärtuse paiknemist σ-skaalal, vt jn 11.7.



Jn 11.7

Tellimuspaneeli (vt jn 11.8) täitmine on sarnane eelmise protseduuriga:

- lahtrisse *Data vectors* trükitakse tunnuste nimed (arvtunnused, võrdse pikkusega, nende arv peab olema vähemalt 3);

Sun Rsv Plot

Data vectors: kasv
ving
kaal
vanus
teadtoid
tulu
pere

Labels: nimed

Sigma multiple: 2

Jn 11.8

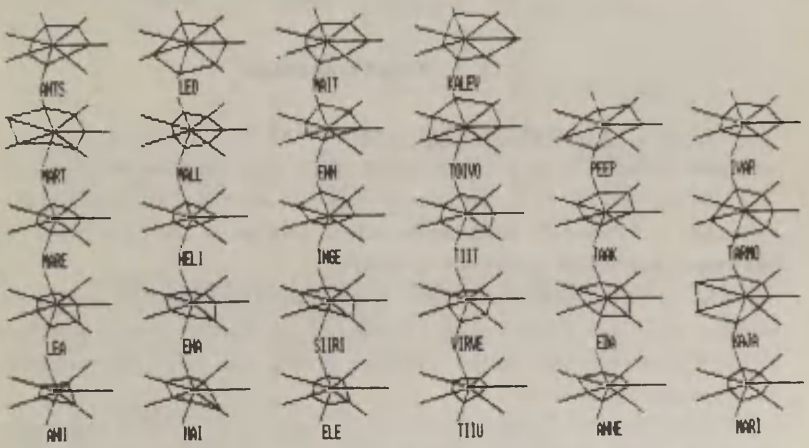
- lahtrisse *Labels* trükitakse sama pika sümboltunnuse nimi, mis sisaldab kas objektide nimesid või märgendeid, mida soovitakse näha kujundite allkirjadena (käesolevas näites on selleks tunnus **nimed**, mis sisaldab objektide eesnimesid);

- lahtrisse *Sigma multiple* trükitakse arv h, mis määrab skaalale pikkused standardhälbe ühikutes

$$-h\sigma, \ldots, h\sigma.$$

Enamikul juhtudest on h sobivaim väärtus 2...3, kusjuures vaikimisi väärtuseks on 3.

Käivitamisel saadud ekraanipildil vastab jällegi igale objektile kujund, vt jn 11.9. Käesoleval juhul on objektide allkirjadeks nimed.



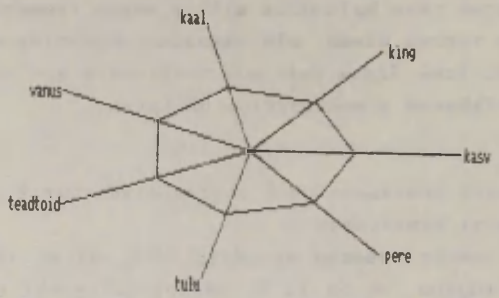
Jn 11.9

Silma järgi otsustades näiteks tunduvad olevat üsna sarnased **MARE**, **VIRVE**, **MARI**, **ELE** ja **LEA**, aga samuti **ANTS** ja **MAIT** jne.

Kujundite võti on saadav jällegi pärast kujundeid, see on kujutatud joonisel 11.10.

Kui käesolevas paragrahvis vaadeldud meetodite puhul on tellijal võimalus visuaalselt objektide sarnasuse või erinevuse üle otsustada, siis järgmises paragrahvis asume vaatlema

meetodeid objektide automaatseks klassifitseerimiseks teatavate formaliseeritud kriteeriumide alusel.



Jn 11.10

§ 11.2. Klasteranalüüs

11.2.1. Klasteranalüüsi idee. Vaatleme objekte k-mõõtmelises ruumis, $k \geq 2$, kusjuures kõik tunnused on arvulised.

Juhul kui $k = 2$, on punktihulk tasandil (korrelatsiooniväljal) lihtsalt kujutatav ning kergesti on leitavad ka omavahel sarnaste punktide kogumid.

Ka kolmemõõtmelisel juhul on võimalik punkte reaalses ruumis kujutada ning leida lähedased ja kauged punktid visuaalselt. Rohkem kui kolme tunnuse puhul aga selline visuaalne protseduur puudub.

Küll aga on võimalik leida objektidevahelised kaugused ka siis, kui kasutatavaid tunnuseid on märksa rohkem kui kolm. Kõige sagedamini kasutatakse selleks nn eukleidilist kaugust j-nda objekti o_j ja g-nda objekti o_g vahel, mis arvutatakse järgmiselt:

$$d_E(o_j, o_g) = \left[\sum_{i=1}^{k} (x_{ij} - x_{ig})^2\right]^{1/2} \qquad (11.1)$$

Nagu näeme, on valemiga (11.1) esitatud eukleidiline kaugus üldistuseks igapäevaelust tuntud kaugusele, mille puhul koordinaatide arv k ei ületa arvu 3.

Tuleb aga öelda, et kaugus (11.1) ei ole hästi kasutatav siis, kui tunnused on väga erinevad. Näiteks vaadeldes tunnuseid **vanus** ja **laste arv**, näeme, et esimene on teisest keskmiselt üle kümne korra suurem ning seetõttu kaob teise mõju üsna ära. Selle vältimiseks standardiseeritakse tunnused (jagatakse standardhälbega) ning saadakse kaugus

$$d_{ES}(O_j,\ O_g) = \left[\sum_{i=1}^{k}\left(\frac{x_{ij} - x_{ig}}{s_i}\right)^2\right]^{1/2} \qquad (11.2)$$

Standardiseeritud eukleidilise kauguse puhul on kõigi tunnuste toime ühesugune, sõltumata nende skaalast.

Kui andmestikus on n objekti, siis määrab nende omavahelise paiknemise n×n kauguste maatriks, mis sisaldab iga objekti kauguse iga teise objektini, vt jn 11.11.

	1	2	3	4
1	0	d(1, 2)	d(1, 3)	d(1, 4)
2	d(1, 2)	0	d(2, 3)	d(2, 4)
3	d(1, 3)	d(2, 3)	0	d(3, 4)
4	d(1, 4)	d(2, 4)	d(3, 4)	0

Jn 11.11

Kauguste maatriks on sümmeetriline, tema peadiagonaalil on nullid, sest iga objekti kaugus temast enesest võrdub nulliga. Fragmenti kauguste maatriksist võib näha joonisel 11.25.

Klasteranalüüsi eesmärgiks on objektide rühmitamine niiviisi, et ühte rühma e klastrisse sattuvad objektid oleksid omavahel sarnased (lähedased), eri rühmade objektid aga üldiselt erinevad (kauged).

Klasteranalüüsi aluseks on enamasti kauguste maatriks või sellega teatud mõttes analoogiline sarnasuskordajate maatriks.

Objektide rühmitamine toimub tavaliselt järgmiselt. Oletame, et esimesel hetkel on meil n klastrit (rühma), millest igaühes on parajasti üks objekt. Siis on klastritevahelisteks kaugusteks parajasti objektidevahelised kaugused. Teeme järgmise sammu:

Otsime välja kaks mingis mõttes kõige lähemat olemasolevat klastrit ning ühendame need üheks klastriks. Klastrite arv väheneb sellega ühe võrra. Seejärel arvutame välja selle uue klastri kaugused kõigist teistest klastritest. Loomulikult ülejäänud klastrite omavahelised kaugused jäävad endiseks.

Teeme järgmise sammu jne, kuni on leitud vajalik arv klastreid.

Erinevad klasteranalüüsi meetodid erinevad üksteisest selle poolest, kuidas arvutatakse kahe klastri omavaheline kaugus. Nimetame siin mõningaid võimalusi, kusjuures klastreid tähistame suure O-ga, objekte - väikese o-ga, lisades vajaduse korral indeksid.

1) $d(O_i, O_g) = \max_{o_j \in O_i,\ o_f \in O_g} d(o_j, o_f)$, s.o suurim kaugus;

2) $d(O_i, O_g) = \min_{o_j \in O_i,\ o_f \in O_g} d(o_j, o_f)$ - vähim kaugus;

3) $d(O_i, O_g) = \overline{d(o_j, o_f)}$, $o_j \in O_i$, $o_f \in O_g$ - keskmine kaugus;

4) $d(O_i, O_g) = \mathrm{med}(o_j, o_f)$, $o_j \in O_i$, $o_f \in O_g$ - kauguste mediaan;

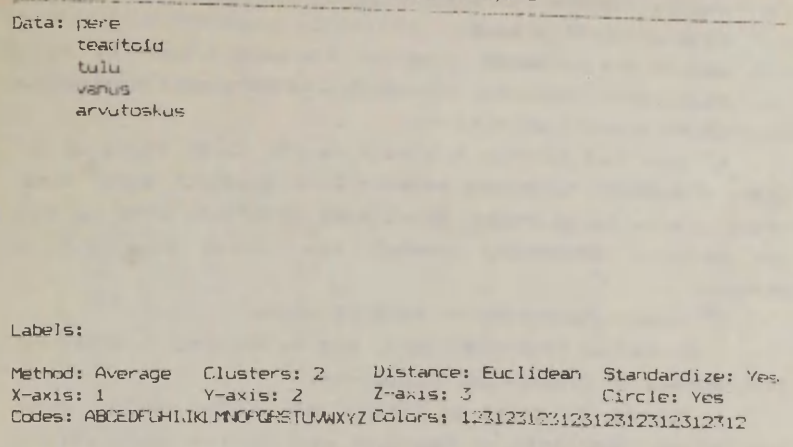
5) $d(O_i, O_g) = d(\bar{o}_i, \bar{o}_g)$ - klastrite keskpunktide vaheline kaugus. Siin $\bar{o}$ tähistab punkti, mille koordinaatideks on klastrisse O kuuluvate punktide koordinaatide keskmised.

Klasteranalüüsi tulemus sõltub oluliselt

- valitud tunnuste loetelust ja arvust,
- valitud objektidevahelisest kaugusest või sarnasusekordajast,
- valitud klasteranalüüsi meetodist,
- soovitavast klastrite arvust.

Seda arvestades ei saa klasteranalüüsi pidada rangeks hindamis- või hüpoteeside kontrollimise meetodiks, pigem on õige pidada seda statistilisi hüpoteese genereerivaks meetodiks.

11.2.2. Klasteranalüüsi (*Cluster Analysis*) tellimus. Klasteranalüüs on 6. protseduur allmenüüs *Q. Multivariate Methods*. Klasteranalüüsi tellimust kujutab joonis 11.12.



Jn 11.12

1° Andmete sisestamiseks (aknasse *Data*) on kaks võimalust:

- trükkida tunnuste nimed;
- trükkida kauguste maatriksi nimi.

Tunnuste nimede trükkimisel tuleb jälgida, et kõik tunnused oleksid arvulised, samuti, et nad oleksid võrdse pikkusega.

Maatriksi võib sisestada siis, kui ülesannet lahendatakse korduvalt ning varasematel etappidel on juba sobiv kauguste (või sarnasuskordajate) maatriks välja arvutatud.

Kuid ka sel juhul on maatriksi sisestamine halvem variant, sest sel juhul ei ole võimalik saada klasteranalüüsi graafilisi illustratsioone.

2° Aknasse *Labels* (märgendid) võib trükkida objektide identifikaatoreid (näiteks nimesid) sisaldava tunnuse nime. Kui see jääb tühjaks, identifitseeritakse objektid järjekoranumbrite järgi.

3° Meetodi märkimiseks on 6 võimalust:

Average - keskmine kaugus (vt p 11.2.1, kaugus 3)), mis on vaikimisi väärtuseks;

Centroid - rühmakeskmine kaugus 5);

Furthest - suurim kaugus 1);

Median - mediaankaugus 4);

Nearest - vähim kaugus 2);

Seeded - ette antud tüüpobjektide meetod.

Nagu näha, erinevad meetodid rühmadevaheliste kauguste arvutamise eeskirjade poolest.

4° Soovitav rühmade e klastrite arv tuleb trükkida aknasse *Clusters*. Vaikimisi väärtus 2 on enamasti selle parameetri jaoks liiga väike. Soovitavat klastrite arvu ei vaja ega arvesta protseduur *Seeded*, mis lähtub tüüpobjektide arvust.

5° Kaugus *Distance*. On kaks võimalust:

- *Euclidean* (eukleidiline), mis tuleb märkida siis, kui lähteandmed on sisestatud tunnustena;

- *Matrix* (kauguste maatriks), mis tuleb märkida siis, kui lähteandmestikuks on kauguste maatriks; viimane väärtus on ühtlasi vaikimisi väärtuseks ning see tuleb tunnustena sisestamisel kindlasti ära muuta.

6° Standardiseerimine *Standardize* - see aken tuleb täita ainult siis, kui arvutatakse eukleidilised kaugused, *Yes* tähistab valemit (11.2) (see on ühtlasi vaikimisi väärtuseks), *No* valemit (11.1).

7° *X-axis*, *Y-axis* ja *Z-axis* tähistavad tunnuseid, millega määratud projektsiooni kasutatakse tulemuse graafiliseks illustreerimiseks. On kaks võimalust:

- märkides ainult x- ja y-telje, kasutades selleks tunnuse järjekorranumbrit tellimuses, ja jättes z-telje kohale 0 (vaikimisi), saame tasandilise graafiku - korrelatsioonivälja;

- märkides ka z-teljele vastava tunnuse järjekorranumbri tellimuses, saame ruumilise punktiparve.

8° *Circle* - kas tasandilisel graafikul klastrid ümbritsetakse joonega (*Yes* (jah), vaikimisi) või mitte (*No*).

9° *Codes* - rühmade koodid. Esimesse rühma kuuluvad objektid tähistatakse graafikul tähega *A*, teise rühma kuuluvad objektid tähega *B* jne. Koode võib soovi korral muuta.

10° *Colors* - on võimalik valida rühmi iseloomustavaid värve graafikul.

11.2.3. Klasteranalüüsi tulemus. Klasteranalüüsi tulemusena fikseeritakse objektide kuulumine klastritesse, tehakse kindlaks igasse klastrisse kuuluvate objektide arv (*Frequency*) ja vastav objektide protsent üldarvust, vt jn 11.13.

Results of Clustering by Average Method

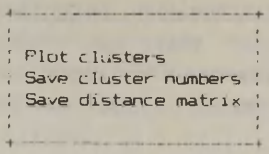
Observation	Cluster	Cluster	Frequency	Percentage
1. Obs. 1	1	1	19	67.8571
2. Obs. 2	1	2	6	21.4286
3. Obs. 3	1	3	1	3.5714
4. Obs. 4	2	4	1	3.5714
5. Obs. 5	2	5	1	3.5714
6. Obs. 6	1			
7. Obs. 7	1			
8. Obs. 8	1			
9. Obs. 9	3			
10. Obs. 10	1			
11. Obs. 11	1			
12. Obs. 12	4			
13. Obs. 13	1			
14. Obs. 14	2			
15. Obs. 15	2			
16. Obs. 16	1			
17. Obs. 17	2			
18. Obs. 18	1			
19. Obs. 19	5			
20. Obs. 20	1			
21. Obs. 21	2			
22. Obs. 22	1			
23. Obs. 23	1			
24. Obs. 24	1			
25. Obs. 25	1			
26. Obs. 26	1			
27. Obs. 27	1			
28. Obs. 28	1			

Jn 11.13

Tulemuste tabeli päisesse on märgitud ka kasutatud meetod.

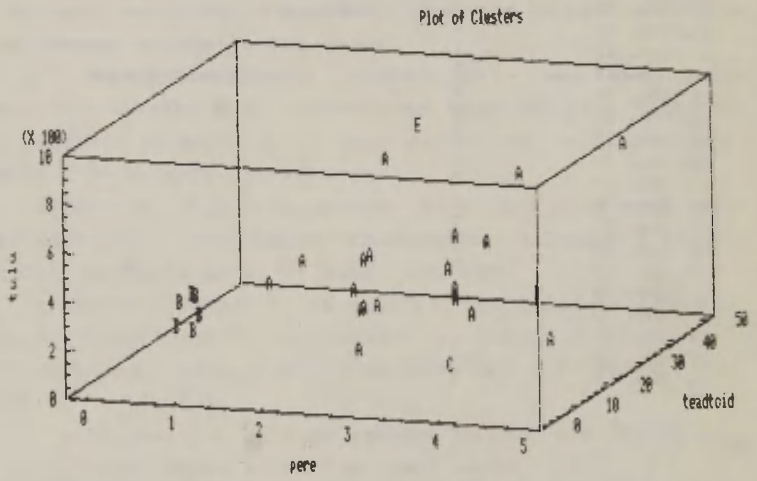
Järgnevalt ilmub ekraanile aken (vt jn 11.14), kus on kolm valikuvõimalust:

- teha graafik *Plot clusters*,
- salvestada klastrite numbrid uue tunnusena *Save cluster numbers*,
- salvestada kauguste maatriks *Save distance matrix*.



Jn 11.14

Joonisel 11.15 on punktihulk kujutatud 3-mõõtmelises ruumis, kus telgedeks on vastavalt pere, teadtoid ja tulu (viimasel skaalal on ühikuks sajarublane). Esimesse rühma kuuluvad objektid on tähistatud tähega A, neid on 19 tükki, teise rühma kuulub 6 tihedasti koos paiknevat B-tähte, kolmanda, neljanda ja viienda rühma ainsateks esindajateks on C, D ja E.



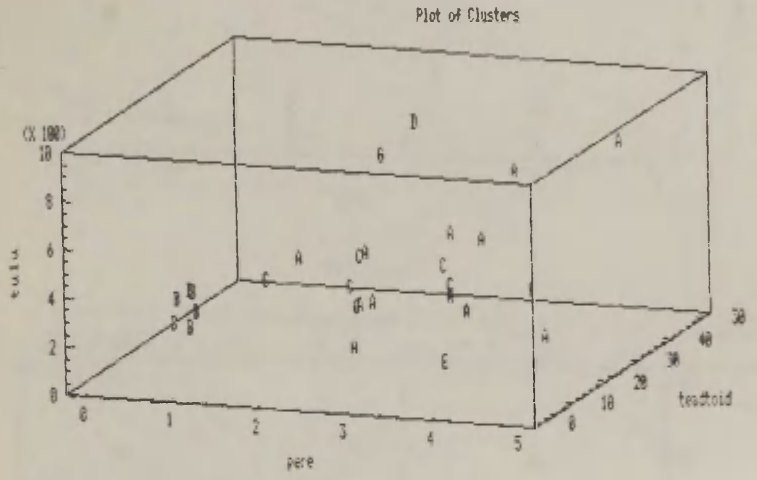
Jn 11.15

Joonisel 11.16 on mõnevõrra muudetud tellimuse väljund. Siin soovitakse 7 klastrit; tulemusena on saadud 3 mitmeelemendilist ja 4 üheelemendilist rühma, vt jn 11.17.

Results of Clustering by Average Method

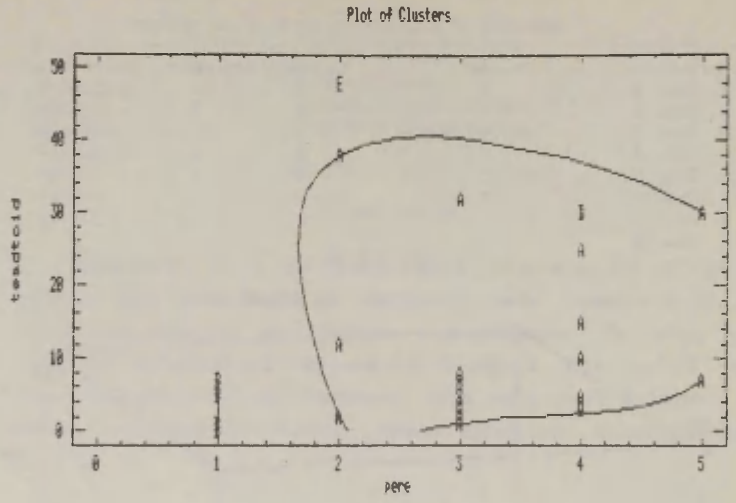
Observation	Cluster	Cluster	Frequency	Percentage
1. Obs. 1	1	1	12	42.8571
2. Obs. 2	1	2	6	21.4286
3. Obs. 3	1	3	6	21.4286
4. Obs. 4	2	4	1	3.5714
5. Obs. 5	2	5	1	3.5714
6. Obs. 6	3	6	1	3.5714
7. Obs. 7	3	7	1	3.5714
8. Obs. 8	1			

Jn 11.16

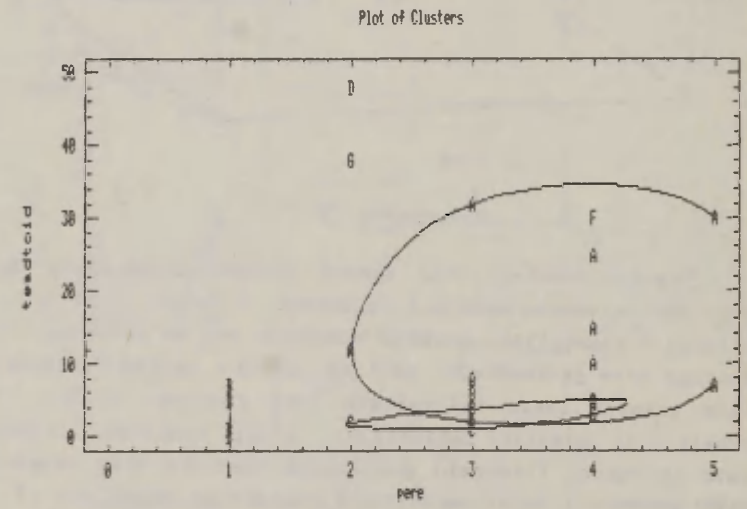


Jn 11.17

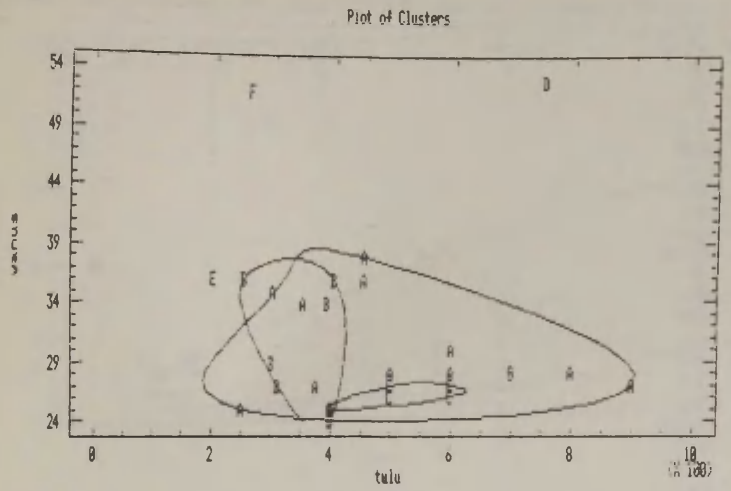
Tsentroidmeetodi abil saadud lahendused langevad täpselt ühte keskmise meetodil saadutega. Joonisel 11.18 on kujutatud 5-klastriline graafik tasandil, mis on määratud tunnustega **pere** ja **teadtoid**, joonisel 11.19 - vastav 7-klastriline graafik samas teljestikus ning joonisel 11.20 - 7-klastriline graafik teljestikus, mille määravad tunnused **tulu** ja **vanus**. Viimaseid graafikuid vaadates saab selgeks, miks punktid D, E, F ja G, mis ruumilisel graafikul 11.17 ülejäänutest kuigivõrd ei eristugi, moodustavad omaette klastrid.



Jn 11.18

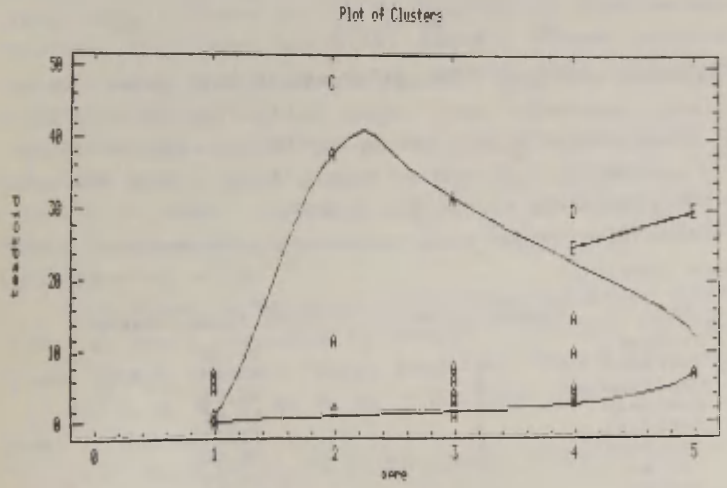


Jn 11.19



Jn 11.20

Joonisel 11.21 aga on esitatud graafik ning joonisel 11.22 tulemuste tabel suurima kauguse meetodil saadud klastrite kohta.



Jn 11.21

Results of Clustering by Complete Linkage or Furthest Neighbor Method

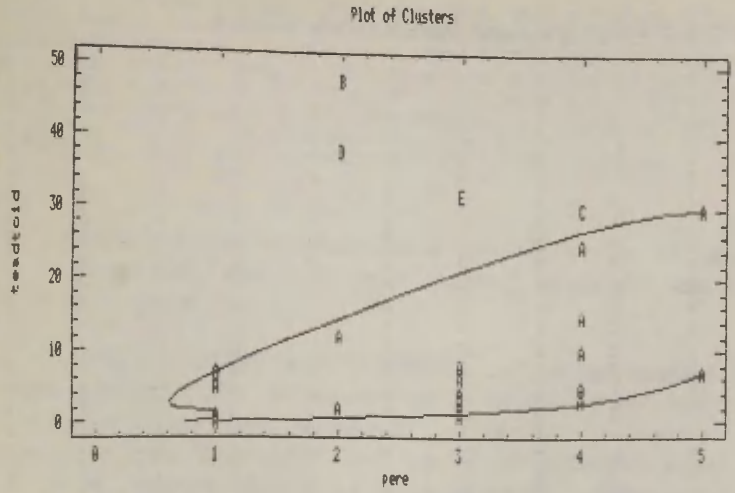
Observation	Cluster	Cluster	Frequency	Percentage
1. ANU	1	1	23	82.1429
2. MAI	1	2	1	3.5714
3. ELE	1	3	1	3.5714
4. TIIU	1	4	2	7.1429
5. ANNE	1	5	1	3.5714
6. MARI	1			
7. LEA	1			
8. EHA	1			
9. SIIRI	2			
10. VIRVE	1			
11. EDA	1			
12. KAJA	3			
13. MARE	1			
14. HELI	1			
15. INGE	1			
16. TIIT	1			
17. JAAK	1			
18. TARMO	4			
19. MART	5			
20. MALL	1			
21. ENN	1			
22. TOIVO	1			
23. PEEP	1			
24. IVAR	1			
25. ANTS	1			
26. LEO	4			
27. MAIT	1			
28. KALEV	1			

Jn 11.22

Jätkates analüüsi teistegi meetodite abil tuleb tõdeda, et enamiku meetodite puhul satub valdav osa objekte ühte rühma, drastilisim selles mõttes on vähima kauguse (lähima naabri) meetod, millele vastab joonis 11.23 - kõik klastrid peale ühe sisaldavad ainult ühe objekti.

Kokkuvõttes saame kümne erineva klasteranalüüsi kohta järgmise tabeli:

Meetod	Klastrite arv	Suurima rühma osakaal
Average	5	67,86
Average	7	42,86
Centroid	5	67,86
Centroid	7	42,86
Furthest	5	82,14
Furthest	7	75,00
Median	5	82,14
Median	7	71,43
Nearest	5	85,71
Nearest	7	71,43



Jn 11.23

Kuivõrd interpreteerimise seisukohalt on huvipakkuvam selline tulemus, kus objektid on enamvähem ühtlaselt klastrite vahel ära jagatud, mitte selline, kus üks klaster sisaldab peaaegu kõiki elemente ning ülejäänud on 1-elemendilised, siis tunduvad keskmiste kauguste ja tsentroidmeetod (vähemalt käesoleva andmestiku korral) olevat eelistatud. Lähima naabri meetodi kalduvus objekte ühte rühma "kokku korjata" on üldtuntud.

Meetodite võrdlemiseks on tabelis joonisel 11.24 kõigi objektide jaoks välja trükitud klastrite numbrid, mis on salvestatud pärast vastava lahenduse leidmist protseduuriga *Save cluster number* (salvesta rühma number), vt teine rida aknas joonisel 11.14.

On huvitav, et objektide vahekorrad kujunevad erinevate meetodite puhul küllaltki erinevaiks. Vaadeldes näiteks kaheksat esimest objekti, saame järgmised rühmastruktuurid:

(1, 2, 3, 4, 5, 6, 7, 8) - üks rühm (*Furthest* 5, *Furthest* 7, *Nearest* 5);

(1), (2, 3, 4, 5, 6, 7, 8) - kaks rühma (*Median* 5, *Median* 7);

row	CLUSTa	CLUSTa	CLUSTc	CLUSTc	CLUSTf	CLUSTm	CLUSTm	CLUSTn	CLUSTn	CLUSTf
1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	2	2	1	2	1
3	1	1	1	1	1	2	2	1	1	1
4	2	2	2	2	1	2	2	1	1	1
5	2	2	2	2	1	2	2	1	1	1
6	1	3	1	3	1	2	2	1	1	1
7	1	3	1	3	1	2	2	1	1	1
8	1	1	1	1	1	2	2	1	2	1
9	3	4	3	4	2	1	1	1	3	2
10	1	3	1	3	1	2	2	1	1	1
11	1	1	1	1	1	2	2	1	2	1
12	4	5	4	5	3	3	3	2	4	3
13	1	3	1	3	1	2	2	1	1	1
14	2	2	2	2	1	2	2	1	1	1
15	2	2	2	2	1	2	2	1	1	1
16	1	1	1	1	1	2	2	1	1	1
17	2	2	2	2	1	2	2	1	1	1
18	1	1	1	1	4	2	4	1	1	4
19	5	6	5	6	5	4	5	3	5	5
20	1	3	1	3	1	2	2	1	1	1
21	2	2	2	2	1	2	2	1	1	1
22	1	1	1	1	1	2	6	4	6	6
23	1	7	1	7	1	5	7	5	7	7
24	1	1	1	1	1	2	2	1	1	1
25	1	1	1	1	1	2	2	1	1	1
26	1	1	1	1	4	2	4	1	1	4
27	1	1	1	1	1	2	2	1	1	1
28	1	3	1	3	1	2	2	1	1	1

Jn 11.24

(4, 5), (1, 2, 3, 6, 7, 8) - kaks rühma (*Average* 5, *Centroid* 5);

(2, 8), (1, 3, 4, 5, 6, 7) - kaks rühma (*Nearest* 7);

(4, 5), (6, 7), (1, 2, 3, 8) - kolm rühma (*Average* 7, *Centroid* 7).

Objektipaare, mis kõigi meetodite puhul ühte rühma kuuluvad, tuleb välja üsna vähe - (4, 5), (2, 8), (6, 7) -, ning ei leidu ühtegi 3-elemendilist ühisklastrit!

Jooniselt 11.24 näeme veel, et kõigi meetodite puhul erinevad ülejäänutest ning moodustavad 1-elemendilised rühmad 12. ja 19. objekt.

Valides aknast joonisel 11.14 protseduuri *Save distances* - säilita kauguste maatriks -, on meil võimalik salvestada 28×28 kauguste maatriks. Kuivõrd selle trükkimiseks kuluks üle 3 leheküljе, esitame sellest ainult 8×8 fragmendi, mis kujutab täielikku kauguste maatriksit 8 esimese objekti vahel, vt jn 11.25.

0	3.3081	1.20546	1.98504	3.44613	3.73872	3.99511	2.37899
3.3081	0	2.74674	3.94413	3.21095	3.03864	2.34353	1.2178
1.20546	2.74674	0	1.59987	2.75036	2.71317	2.89125	1.76779
1.98504	3.94413	1.59987	0	2.36503	2.72005	3.66655	2.95319
3.44613	3.21095	2.75036	2.36503	0	1.88517	3.01774	2.57072
3.73872	3.03864	2.71317	2.72005	1.88517	0	1.68273	2.75874
3.99511	2.34353	2.89125	3.66655	3.01774	1.68273	0	2.45701
2.37899	1.2178	1.76779	2.95319	2.57072	2.75874	2.45701	0

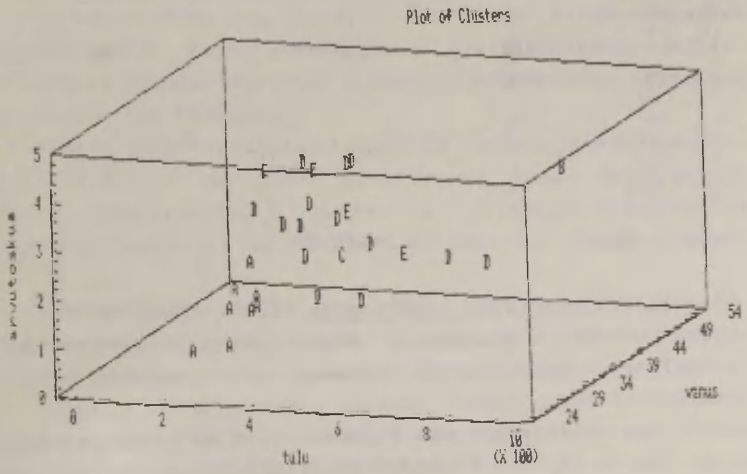
Jn 11.25

Paneme tähele, et kirjeldatud kauguse mõttes on lähimad 1. ja 3. objekt, samuti 2. ja 8. objekt, kaugeimad aga 1. ja 7. ning 2. ja 4. objekt.

11.2.4. Klastrite moodustamine etteantud keskpunktide ümber (*Seeded*). Kui valida meetodiks *Seeded*, tuleb anda ette klastrite tüüpilised elemendid (liidrid), selleks ilmub ekraanile aken tekstiga *Enter vector of starting seeds*, vt jn 11.26, millele vastuseks tuleb märkida "tüüpelementide" järjekorranumbrid.

Enter vector of starting seeds: 1 12 19 22 27

Jn 11.26



Jn 11.27

Selle meetodi puhul moodustub just nii palju klastreid, kui palju on tüüpelemente, ning iga objekt loetakse kuuluvaks sellesse klastrisse, mille tüüpelement on talle lähim.

Meetodi rakendamise tulemusi kirjeldavad joonised 11.27 ja 11.28.

Results of Clustering by Seeded Method

Observation	Cluster	Cluster	Frequency	Percentage	Seed
1. ANU	1	1	8	28.5714	1
2. MAI	4	2	1	3.5714	12
3. ELE	1	3	1	3.5714	19
4. TIIU	1	4	4	14.2857	22
5. ANNE	4	5	14	50.0000	27
6. MARI	5				
7. LEA	5				
8. EHA	5				
9. SIIRI	1				
10. VIRVE	5				
11. EDA	5				
12. KAJA	2				
13. MARE	5				
14. HELI	1				
15. INGE	1				
16. TIIT	5				
17. JAAK	5				
18. TARMO	5				
19. MART	3				
20. MALL	5				
21. ENN	1				
22. TOIVO	4				
23. PEEP	4				
24. IVAR	1				
25. ANTS	5				
26. LEO	5				
27. MAIT	5				
28. KALEV	5				

Jn 11.28

11.2.5. Tunnuste rühmitamine. Kõiki rühmitamismeetodeid saab kasutada ka tunnuste rühmitamiseks, lähtudes näiteks korrelatsioonimaatriksist (viimast on soovitatud ka juhendis [5]).

Seda tehes saame aga tulemuse, mis on kaugel ootuspärasest (vt jn 11.29). Erinevatesse klastritesse satuvad omavahel tihedasti korreleeritud tunnused kaal, kasv, ja **king**, seevastu aga arvutoskus ja abielus, mis on praktiliselt sõltumatud, kuuluvad samasse klastrisse. Põhjus on ärusaadav:

Results of Clustering by Average Method

Observation	Cluster	Cluster	Frequency	Percentage
1. vanus	1	1	6	54.5455
2. sugu	1	2	2	18.1818
3. kasv	1	3	2	18.1818
4. kaal	2	4	1	9.0909
5. king	3			
6. pere	3			
7. tulu	2			
8. teadtoid	4			
9. abielus	1			
10. arvutoskus	1			
11. eritulu	1			

Jn 7.29

korrelatsioonimaatriksis tähistavad suured kordajad lähedasi tunnuseid, kauguste maatriksis aga kaugeid objekte. Seetõttu ei saa tunnuste rühmitamiseks kasutada vahetult korrelatsioonimaatriksit, küll aga võib kasutada näiteks maatriksit, mille elementideks on suurused $1-r_{ij}^2$ või $\arccos r_{ij}$.

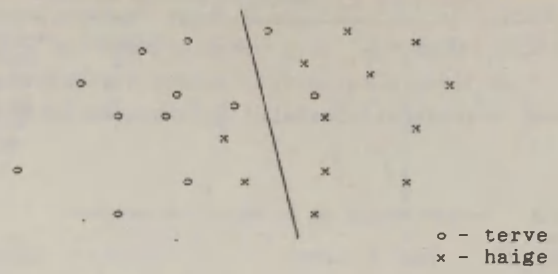
§ 11.3. Diskriminantanalüüs

11.3.1. Diskriminantanalüüsi meetodi idee ja eesmärk. Diskriminantanalüüs on sageli rakendatav mitmemõõtmelise analüüsi meetod, mille eesmärgiks on klassifitseerida mitmemõõtmelises ruumis mõõdetud objektid teatavatesse etteantud klassidesse (rühmadesse).

Tüüpiliseks diskriminantanalüüsi rakendusvaldkonnaks on meditsiiniline ja tehniline diagnostika. Toome näite meditsiinist. Olgu uuritud n_1 tervet ja n_2 teatava diagnoosiga patsienti, kusjuures neil kõigil on mõõdetud samad tunnused $X_1, \ldots, X_k$.

Diskriminantanalüüsi eesmärgiks on: etteantud teadaoleva klassikuuluvusega objektidele (käesolevas näites - diagnoositud patsientidele) tuginedes

- leida k-mõõtmelises ruumis eeskiri (diskrimineeriv funktsioon, ka pind), mis võimalikult hästi eristaks klasse (terveid ja haigeid), vt jn 11.30;
- selle eeskirja järgi otsustada tundmatu objekti x_0 puhul, kumba rühma ta kuulub (panna diagnoos).



Jn 11.30. Diskrimineeriv joon

Kui diskrimineeriv pind (joon) on lineaarne, siis räägime lineaarsest diskriminantanalüüsist. Lineaarne diskriminantanalüüs hindamis- ja hüpoteeside kontrollimise meetodina on rakendatav siis, kui on täidetud järgmised eeldused:

1) mõlemas rühmas on tunnused mitmemõõtmelise normaaljaotusega,

2) kõigis rühmades on kovariatsioonimaatriksid võrdsed.

Sõnastatud põhiülesandega seostub rida täiendavaid ülesandeid:

- lahendada diskrimineerimisülesanne rohkem kui 2 rühma puhul;
- hinnata eksliku klassifitseerimise tõenäosust;
- kontrollida tehtud otsustuste olulisust;
- leida minimaalne hulk diskrimineerivaid tunnuseid jne.

Kui eristatavate rühmade arv on suurem kui kaks (terved ja patsiendid mitme erineva diagnoosiga), siis on diskriminantanalüüsi metoodikat arendatud mitmes erinevas suunas:

1. Klassikaline diskriminantanalüüs, mille puhul põhieesmärgiks on leida nn klassifitseerimiskriteeriumid (ka informandid), mille alusel (valimisse kuuluv või "uus") objekt paigutatakse ühte olemasolevatest klassidest.

Klassifitseerimiskriteeriumide aluseks võib olla

- objekti kaugus (eukleidiline või Mahalanobise kaugus) iga rühma keskpunktist,
- tihedusfunktsioonide järgi arvutatud tõepärasuhe jne.

Klassifitseerimise juures saab arvesse võtta rühmade arvukust, eksliku klassifitseerimise ohtlikkust (riskifunktsiooni) jne.

2. Kanooniline diskriminantanalüüs, mille eesmärgiks on diskrimineerimisel kasutatavate lähtetunnuste hulka vähendada (komponentanalüüsiga sarnasel põhimõttel). Selleks asendatakse lähtetunnused nende selliste lineaarkombinatsioonidega, mis kõige paremini rühmi eristavad. Selliselt saadud diskrimineerivad funktsioonid määravad maksimaalse hajuvuse sihid (peakomponendid) rühmade keskpunktide poolt kirjeldatud ruumis, nende maksimaalne võimalik arv on min (k, s-1), kus k on tunnuste arv, s etteantud rühmade arv.

Diskriminantfunktsioonide olulisuse kontrollimisel kasutatakse nende ja rühmitustunnuste (täpsemalt - rühmitustunnustele vastavate dihhotoomiliste tunnuste) kanoonilist analüüsi ning oluliseks loetakse ainult need diskriminantfunktsioonid, millele vastavad olulised kanoonilised korrelatsioonikordajad.

SG diskriminantanalüüsi protseduur hõlmab mõlemat varianti.

11.3.2. Diskriminantanalüüsi protseduur *Discriminant Analysis*. Diskriminantanalüüs on süsteemis **Statgraphics** realiseeritud protseduurina *Discriminant Analysis*, mis on seitsmes protseduur allmenüüs *Q. Multivariate Methods*.

```
                    Discriminant Analysis
-------------------------------------------------------------

Classification factor: sugu

Data vectors: vanus
              abielus
              tulu
              teadtoid
```

Jn 11.31

Diskriminantanalüüsi tellimus on esitatud joonisel 11.31. Tellimuspaneelile tuleb kõigepealt trükkida rühmitav tunnus *Classification factor*. See peab olema diskreetne arvtunnus, sümboltunnust saab kasutada ainult pärast arvuliseks kodeerimist (väärtuste järjestus pole siinjuures oluline). Tähtis on see, et rühmitaval tunnusel ei oleks liiga palju väärtusi, see tähendaks materjali jaotamist väga paljudeks osadeks, mida enamasti ei õnnestu vajaliku usaldatavusega

tena. Rühmitava tunnuse väärtus määrab, millisesse klassi objekt kuulub.

Järgmiseks tuleb sisestada tunnused, mille väärtuste järgi toimub diskrimineerimine. Tuleb kohe märkida, et ka siin ei maksa liiale minna: liiga suure argumenttunnuste arvu korral väheneb tihti otsustuste olulisus, sest hinnata on tarvis paljusid parameetreid.

Esimene diskriminantanalüüsi tulemus ilmub ekraanile kaheosalise tabelina, mis kirjeldab teostatavat diskriminantanalüüsi tervikuna (kanoonilise diskriminantanalüüsi mõttes), vt jn 11.32.

Discriminant Analysis for sugu

Discriminant Function	Eigenvalue	Relative Percentage	Canonical Correlation
1	.4275727	100.00	.54728

Functions Derived	Wilks Lambda	Chi-Square	DF	Sig.Level
0	.7004897	8.5434138	4	.07358

Jn 11.32

Pealkirjas on ära märgitud klassifitseeriva (funktsioon-) tunnuse nimi.

Tabeli ülemises osas on märgitud põhimõtteliselt eraldatavate diskriminantfunktsioonide arv. Praegu on see s-1, kus s on rühmitava tunnuse erinevate väärtuste arv. Märgitakse ka seda, kui suure osa summaarsest hajuvusest kirjeldab iga diskrimineeriv funktsioon (ühe diskrimineeriva funktsiooni korral on see alati 100 %), samuti märgitakse ära kanoonilised korrelatsioonid lähtetunnuste ja rühmitavate dihhotoomiliste tunnuste vahel (viimaste arv on s-1), seega ei saa kanooniliste korrelatsioonide arv olla suurem kui $\rho = \min(k, s-1)$, kus k on esialgsete tunnuste arv. Kahe rühma korral võrdub ainus kanooniline korrelatsioon mitmese korrelatsioonikordajaga 2-väärtuselise rühmitustunnuse prognoosimisel regressioonanalüüsi abil.

Tabeli alumises osas on esitatud andmed rühmitamise olulisuse kohta. Seal vastab üks rida igale diskrimineeriva-

le funktsioonile (statistikute väärtused on arvutatud enne vastava diskrimineeriva funktsiooni arvutamist); esimene veerg *Functions Derived* annab varem eraldatud diskriminantfunktsioonide arvu, mis h-ndal sammul on h-1, järgneb Wilksi Λ-statistik,

$$\Lambda = \prod_{i=h}^{p} (1 - \rho_i^2), \qquad (11.3)$$

seejärel Wilksi Λ põhjal arvutatud χ^2-statistiku väärtus,

$$W = - m \ln \Lambda,$$

$$m = n - \frac{1}{2}(f_1 - f_2 + 1),$$

Statistik W on ligikaudu χ^2-jaotusega vabadusastmete arvuga $f_1 f_2$, kus $f_1 = k + 1 - h$, $f_2 = s - h$, k on argumenttunnuste arv, s on rühmade arv ja h - sammu järjekorranumber. Viimases veerus on antud χ^2-statistiku olulisuse tõenäosus, kusjuures nullhüpoteesiks on

$$H_0: \rho_i = 0, \quad i = h, \ldots, p.$$

Järgmistel joonistel - 11.33 ja 11.34 - esitatakse diskriminantfunktsioonid, mis saadakse järgmisel sammul klahvi ESC kasutamise tulemusena.

Discriminant Analysis for sugu

Standardized Discriminant Function Coefficients

	1
vanus	0.97269
abielus	0.15575
tulu	0.29845
teadtoid	-1.55020

Jn 11.33

Discriminant Analysis for sugu

Unstandardized Discriminant Function Coefficients

	1
vanus	0.12982
abielus	0.32143
tulu	0.00175
teadtoid	-0.13193
CONSTANT	-3.58876

Jn 11.34

Kahe rühma korral määrab ainus standardiseerimata diskriminantfunktsioon rühmi kõige paremini eristava sihi, mis arvutatakse lähtetunnuste lineaarkombinatsioonina

$$\sum_{i=1}^{k} x_i b_i + b_0.$$

Klahvi ESC vajutamise järel saame ekraanile kõigepealt standardiseeritud diskriminantfunktsiooni kordajad, vt jn 11.33. Standardiseerimata diskriminantfunktsioonist saame standardiseeritud funktsiooni

$$\sum_{i=1}^{k} x_i \beta_i$$

kordajad β_i lihtsalt teisendusega

$$\beta_i = b_i / s_i, \quad i = 1, \ldots, k,$$

kus s_i on i-nda tunnuse standardhälve, s.o ruutjuur rühmasisesest ühisest dispersioonist, mis on maatriksi S_s diagonaalielemendiks,

$$S_s = \frac{1}{n_1+n_2-2} [S_1(n_1 - 1) + S_2(n_2 - 1)], \qquad (11.4)$$

kus S_1 ja S_2 on vastavalt 1. ja 2. rühma objektide kovariatsioonimaatriksid. Seejärel saame järgmise sammuna standardiseerimata diskriminantfunktsiooni, vt jn 11.34.

Veelkordse ESC-klahvile vajutamise tulemusena saame rühmade nn tsentroidid (*Centroids*) c_1 ja c_2, vt jn 11.35. Need on arvutatud vastavalt eeskirjale

$$c_j = \sum_{i=1}^{k} \bar{x}_i^{(j)} b_i + b_0, \quad j = 1, 2,$$

seega tsentroid on diskriminantfunktsiooni väärtus, kui tunnuste väärtustena kasutatakse rühmakeskmisi.

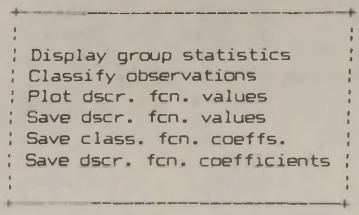
Discriminant Analysis for sugu

Group Centroids

	1
1	-0.72758
2	0.54569

Jn 11.35

Sellega lõpeb nn kohustuslike tulemuste osa. Kasutades akent, vt jn 11.36, on võimalik saada veel rida lisatulemusi



Jn 11.36

Kõigepealt saame üksikute rühmade arvkarakteristikud (*Display group statistics*) - kõigi tunnuste keskmised ja standardhälbed, vt jn 11.37.

Discriminant Analysis for sugu

Group	1	2	TOTAL
COUNTS	12	16	28
MEANS			
vanus	31.3333	31.0625	31.1786
abielus	0.66667	0.68750	0.67857
tulu	531.250	416.250	465.536
teadtoid	17.8333	6.43750	11.3214
STD. DEVIATIONS			
vanus	7.30297	7.62862	
abielus	0.49237	0.47871	
tulu	196.817	147.507	
teadtoid	12.2685	11.3547	

Jn 11.37

Seejärel on võimalik saada (ESC-klahvi abil) rühmasisene ühine kovariatsioonimaatriks S_e, vt (11.4). Rühmasisene korrelatsioonimaatriks R_e arvutatakse rühmasisese kovariatsioonimaatriksi põhjal standardsel viisil, vt jn 11.38.

Kui diskriminantfunktsioone on vähemalt 2, on nende korrelatsioonivälju võimalik ekraanil näha.

Samuti on võimalik rida tulemusi salvestada. Lisaks juba leitud diskriminantfunktsioonide kordajatele (joonised

Discriminant Analysis for sugu

Within-Group Covariance Matrix

	vanus	abielus	tulu	teadtoid
vanus	56.1386	0.40946	-167.356	55.6242
abielus	0.40946	0.23478	21.3942	1.13542
tulu	-167.356	21.3942	28941.6	956.875
teadtoid	55.6242	1.13542	956.875	138.062

Within-Group Correlation Matrix

	vanus	abielus	tulu	teadtoid
vanus	1.00000	0.11278	-0.13130	0.63182
abielus	0.11278	1.00000	0.25954	0.19943
tulu	-0.13130	0.25954	1.00000	0.47869
teadtoid	0.63182	0.19943	0.47869	1.00000

Jn 11.38

11.33 ja 11.34) saab salvestada ka klassifitseerivate funktsioonide (nn informantide) kordajad a_{ij}, mis vastavad rühmale indeksiga j, j = 1, 2, ..., p (*Save class. fcn. coeffs.*):

$$I_{jg} = \sum_{i=1}^{k} a_{ij}x_{ig} + a_{oj}, \qquad (11.5)$$

vt jn 11.39.

Variable: WORKAREA.CLCOEFF2 (length = 5 2)

(1,1)	1.72494	(1,2)	1.89023
(2,1)	-1.02415	(2,2)	-0.614831
(3,1)	0.0616403	(3,2)	0.063874
(4,1)	-0.984589	(4,2)	-1.15258
(5,1)	-34.2766	(5,2)	-38.7302

Jn 11.39

Arvutades mingi g-nda objekti jaoks välja kõigi informantide väärtused, on lihtne teha otsustust tema kuuluvuse kohta: objekt loetakse kuuluvaks sellesse rühma, millele vastava informandi väärtus on suurim. See tulemus kuulub klassikalisse diskriminantanalüüsi.

Valides aknast joonisel 11.36 rea *Save dscr. fcn. values* - säilitada diskriminantfunktsioonide väärtused -, saame salvestada vaadeldava andmestiku jaoks arvutatud diskriminantfunktsiooni kui uue tunnuse väärtused, mis on leitud, kasutades valemit (11.5), vt jn 11.40.

```
Variable: WORKAREA.DSCRMAT2   (length = 28 1)
-------------------------------------------------------------------------
( 1,1)  0.15289
( 2,1)  0.879145
( 3,1)  0.152178
( 4,1)  0.418155
( 5,1)  1.44826
( 6,1)  0.399867
( 7,1)  0.89462
( 8,1)  0.441245
( 9,1)  1.03983
(10,1)  0.632867
(11,1)  1.53598
(12,1) -1.40387
(13,1)  0.28411
(14,1)  0.460226
(15,1)  0.863682
(16,1) -0.299111
(17,1)  0.717718
(18,1) -1.52721
(19,1) -0.0185389
(20,1)  0.531799
(21,1) -0.221206
(22,1) -1.76656
(23,1) -3.7392
(24,1) -0.159772
(25,1) -0.659819
(26,1) -2.14126
(27,1)  0.628642
(28,1)  0.45532
-------------------------------------------------------------------------
```

Jn. 11.40

Diskriminantfunktsiooni väärtuste järgi rühmitamine toimub rühma tsentroidide põhjal - otsustus rühma kuuluvuse kohta tehakse vastavalt diskriminantfunktsiooni väärtuse märgile: kui esimese rühma tsentroid on negatiivne, siis loetakse esimesse rühma kuuluvaks need objektid, millele vastab negatiivne diskriminantfunktsiooni väärtus ja vastupidi. Nii on võimalik klassifitseerida ka seni klassifitseerimata objekte.

Lõpuks aknast joonisel 11.36 rea *Classify observations* valikuga saadud klassifitseerimistulemuste tabel võimaldab otsustada klassifitseerimisprotsessi edukuse üle - siin on antud tabel, vt jn 11.41, mille read vastavad tegelikele rühmadele (vastavalt rühmitava tunnuse väärtustele) ja veerud rühmitamise protsessis tekkinud rühmadele. Diskrimineerimiseks on kasutatud klassifitseerivaid funktsioone, vt

jn 11.39. Näeme, et meeste (rühm 1) ära tundmisel töötas käesolev protseduur vaevalt rahuldavalt: 12 meesvastajast tunti ära vaid 8. Seevastu naiste määramisel eksis protseduur vaid ühel korral 16-st. Meenutame seejuures, et materjal oli üldse vaevalt eristatav - esimeses tulemuste tabelis (jn 11.32) oli olulisuse tõenäosus 0,07, seega kogu protseduuri usaldusväärsus tervikuna on kahtlustäratav: valitud tunnused ei sobi hästi soo eristamiseks.

Classification Results for suyu

Predicted Group (count,percentage)					
Actual Group	1		2		TOTAL
1	8	66.67	4	33.33	12 100.00
2	1	6.25	15	93.75	16 100.00

Jn 11.41

11.3.3. Diskriminantanalüüs mitme (rohkem kui 2) rühma korral. Vaatleme käesolevat juhtu näite varal, kus samade argumentide alusel prognoositakse tunnust pere (millel on 5 võimalikku väärtust), vt jn 11.42.

Discriminant Analysis

```
Classification factor: pere

Data vectors: vanus
              abielus
              tulu
              teadtoid
```

Jn 11.42

Kui kahe rühma juhul rühmade keskpunktid (tsentroidid) paiknevad ühel sirgel, siis s rühma keskpunktid määravad teatava kujundi (s-1)-mõõtmelises ruumis. Selles ruumis on võimalik (komponentanalüüsi põhimõttel) leida suurima hajuvuse suund, mis vastab esimesele diskrimineerivale funktsioonile, sellega ristuv teine diskrimineeriv funktsioon, mis kirjeldab maksimumi jääkhajuvusest jne. Sellist lahutust kirjeldavadki diskriminantfunktsioonidele vastavad omavek - torid, ning diskrimineerivate funktsioonide suhtelised osakaalud (mis väljendatakse protsentides) on arvutatud suhtena

$$\frac{\lambda_i}{\sum_{i=1}^{p-1} \lambda_i}, \quad i = 1, \ldots, s-1.$$

Kanoonilised korrelatsioonid lähtetunnuste ja rühmi kirjeldavate dihhotoomiliste tunnuste vahel iseloomustavad diskrimineerimisülesande lahendatavust üldse ja võimaldavad kontrollida hüpoteese oluliste diskriminantfunktsioonide arvu kohta.

Olulisuse kontroll toimub Λ-statistiku (vt valem 11.3) järgi ning loomulikult on mõistlik arvestada ainult olulisi diskriminantfunktsioone.

Käesoleval juhul on esimene diskriminantfunktsioon oluline, st võetakse vastu hüpotees H_1: $\rho_1 \neq 0$ ($p = 0{,}042 < 0{,}05$), esimene kanooniline korrelatsioon erineb nullist. Kuna järgmised diskriminantfunktsioonid on mitteolulised, siis pole erilist lootust eristada korrektselt kõiki viit rühma, kuid teatava lahutuse saab siiski adekvaatselt teha, vt jn 11.43.

Discriminant Analysis for pere

Discriminant Function	Eigenvalue	Relative Percentage	Canonical Correlation
1	1.0047394	63.79	.70794
2	.4071962	25.69	.53793
3	.1725069	10.88	.38757
4	.0005196	.03	.02279

Functions Derived	Wilks Lambda	Chi-Square	DF	Sig.Level
0	.3021666	26.927480	16	.04230
1	.6057652	11.278413	9	.25711
2	.8524305	3.592431	4	.46396
3	.9994807	.011688	1	.91391

Jn 11.43

Joonisel 11.44 on kujutatud standardiseerimata diskriminantfunktsioonide kordajate tabel; kuivõrd rühmi on 5, on diskriminantfunktsioone 5 - 1 = 4. Joonisel 11.45 on rühmade keskpunktid.

Discriminant Analysis for pere

Unstandardized Discriminant Function Coefficients

	1	2	3	4
vanus	-0.05117	-0.05096	0.20406	0.04352
abielus	2.39710	1.17432	0.30199	-0.44662
tulu	-0.00330	0.00193	0.00453	0.00665
teadtoid	-0.02865	0.05294	-0.09077	-0.09453
CONSTANT	1.82948	-0.70550	-7.64676	-3.08120

Jn 11.44

Discrim. ant Analysis for pere

Group Centroid

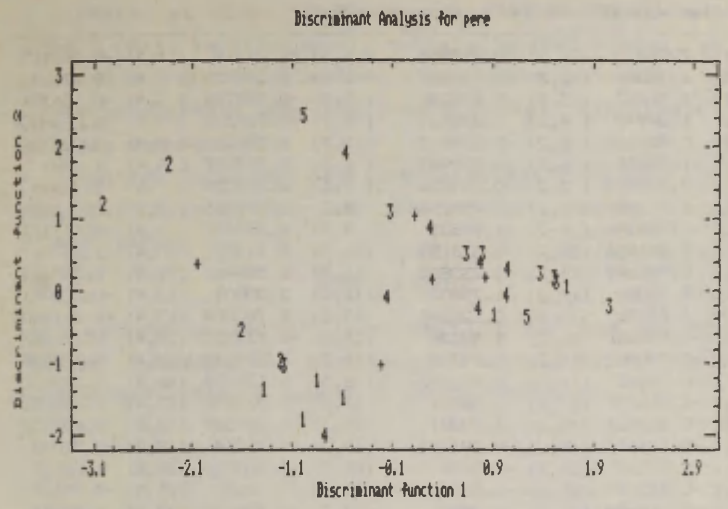
	1	2	3	4
1	-0.19990	-1.02288	-0.03675	0.01432
2	-2.03577	0.37334	-0.18117	-0.01188
3	0.83014	0.17982	-0.42892	-0.01073
4	0.30511	0.14417	0.54266	-0.01026
5	0.13022	1.02599	0.01760	0.06477

Jn 11.45

Rühmade tsentroidide tabelit vaadates näeme, et esimene diskriminantfunktsioon eristab tugevasti 2. rühma (2-liikmelisi peresid) ülejäänutest: sellele vastab väärtus -2, ülejäänud rühmade kordajad on nullilähedased või positiivsed. See ongi kõige tugevam erinevus antud materjalis. Teine diskriminantfunktsioon teeb teatava vahe sisse 1. ja 5. rühma vahel, ülejäänud diskriminantfunktsioonide toime on üsnagi segane.

Kuivõrd rühmade arv on suurem kui 2, on käesoleval juhul võimalik diskriminantfunktsioone ka graafikul kujutada, joonisel 11.46 on 1. ja 2. diskriminantfunktsiooni väärtused Ristikestega on kujutatud rühmade keskpunktid. Joonisel 11.46 on rühmade 2 ja 1 eristumine päris hästi näha.

Klassifitseerimiskordajad on esitatud joonisel 11.47, diskriminantfunktsioonide väärtused (mille alusel on tehtud otsustused) - joonisel 11.48.



Jn 11.46

Variable: WORKAREA.CLCOEFF (length = 5 5)

(1,1)	1.80766	(1,2)	1.79984	(1,3)	1.61255	(1,4)	1.83952
(2,1)	-4.32173	(2,2)	-7.1148	(2,3)	-0.547492	(2,4)	-1.55472
(3,1)	0.061127	(3,2)	0.0690513	(3,3)	0.058106	(3,4)	0.0641707
(4,1)	-0.942096	(4,2)	-0.800002	(4,3)	-0.869764	(4,4)	-0.945051
(5,1)	-35.8568	(5,2)	-40.6299	(5,3)	-31.6537	(5,4)	-39.7715

(1,5)	1.69965
(2,5)	-1.13046
(3,5)	0.0645718
(4,5)	-0.852788
(5,5)	-37.2626

Jn 11.47

Variable: WORKAREA.DSORMAT (length = 28 4)

(1,1)	2.065	(1,2)	-0.216965	(1,3)	-1.29316	(1,4)	-0.965415
(2,1)	1.24506	(2,2)	-0.36539	(2,3)	0.519915	(2,4)	-0.670191
(3,1)	1.51267	(3,2)	0.178286	(3,3)	-0.795784	(3,4)	-0.156394
(4,1)	1.64978	(4,2)	0.0704236	(4,3)	-0.727524	(4,4)	0.0836866
(5,1)	0.921162	(5,2)	-0.329315	(5,3)	1.35815	(5,4)	0.227788
(6,1)	-1.20834	(6,2)	-0.959962	(6,3)	-0.259537	(6,4)	1.1882
(7,1)	0.778924	(7,2)	0.409256	(7,3)	0.608359	(7,4)	1.35596
(8,1)	1.04528	(8,2)	-0.0591546	(8,3)	0.269835	(8,4)	-0.664614
(9,1)	-0.758594	(9,2)	-1.99535	(9,3)	0.332471	(9,4)	-0.467313
(10,1)	0.801446	(10,2)	0.513155	(10,3)	0.31352	(10,4)	1.21791
(11,1)	0.756149	(11,2)	-0.232858	(11,3)	1.58446	(11,4)	0.560484
(12,1)	-2.33561	(12,2)	1.75602	(12,3)	2.50809	(12,4)	-0.768443
(13,1)	1.54131	(13,2)	0.125346	(13,3)	-0.705009	(13,4)	-0.0618606
(14,1)	-0.575162	(14,2)	-1.48334	(14,3)	-0.733901	(14,4)	0.156542
(15,1)	-0.9809	(15,2)	-1.79301	(15,3)	0.377231	(15,4)	-0.323684
(16,1)	0.281657	(16,2)	0.891663	(16,3)	0.131253	(16,4)	0.352121
(17,1)	-1.36925	(17,2)	-1.36807	(17,3)	0.511996	(17,4)	0.426292
(18,1)	-0.562524	(18,2)	1.90881	(18,3)	-0.279385	(18,4)	0.650532
(19,1)	-0.15167	(19,2)	-0.091223	(19,3)	1.72014	(19,4)	-2.37078
(20,1)	-1.17969	(20,2)	-1.0129	(20,3)	-0.168762	(20,4)	1.28273
(21,1)	-0.845024	(21,2)	-1.23396	(21,3)	-1.00646	(21,4)	-0.484691
(22,1)	-0.119648	(22,2)	1.09461	(22,3)	-0.458336	(22,4)	-1.90488
(23,1)	-3.00201	(23,2)	1.22979	(23,3)	-2.21407	(23,4)	-0.797178
(24,1)	1.37825	(24,2)	0.239901	(24,3)	-0.86391	(24,4)	-0.613836
(25,1)	-1.59714	(25,2)	-0.532478	(25,3)	-0.759158	(25,4)	0.329906
(26,1)	-0.984615	(26,2)	2.41738	(26,3)	-0.484705	(26,4)	0.799737
(27,1)	0.641815	(27,2)	0.517118	(27,3)	0.540099	(27,4)	1.11588
(28,1)	1.05166	(28,2)	0.322223	(28,3)	-0.0258105	(28,4)	0.501504

Jn 11.48

Toome näite ka klassifitseerimisülesande lahendamise kohta klassifitseerimiskordajate abil.

Oletame, et meil on tegemist 40-aastase abielus vastajaga, kelle tulu on 800 rbl ja teadustööde arv samuti 40. Küsime, kui suur võiks olla tema pere? Tähistame selle objekti $x_0 = (x_{10}, x_{20}, x_{30}, x_{40})$.

Klassifitseerimisfunktsiooni väärtusteks saame, kasutades valemit (11.5) ja tabelist joonisel 11.47 võetud kordajaid:

$$I_1(x_0) = 43,3456$$

$$I_2(x_0) = 47,4899$$

$$I_3(x_0) = 43,9870$$

$$I_4(x_0) = 45,7899$$

$$I_5(x_0) = 47,1389$$

Näeme, et kõige tõepärasem on selle objekti kuulumine rühma 2 (2-liikmeline pere), sellele järgneb oletus, et tema pere võiks olla 5-liikmeline. Nagu näha, ei paikne rühmad argumenttunnustega määratud ruumis sugugi lineaarselt pereliikmete arvu suhtes.

Kontrollimaks oma järeldust arvutame veel ka 1. diskriminantfunktsiooni väärtuse (kasutades andmeid jooniselt 11.44). Leiame

$$d_1 = -1{,}6062,$$

mis, arvestades tabelit jooniselt 11.45, kus on esitatud rühmade keskpunktid diskriminantfunktsioonidega määratud teljestikus, kõneleb samuti 2. rühma valimise poolt.

JÄRELSÕNA

Raamatu teises osas seadsid autorid oma eesmärgiks käsitleda põhilisi matemaatilise statistika meetodeid, millele kuulub keskne koht süsteemis SG.

Ruumi- ja ajapuudusel jäid käsitlemata protseduurid, mis on seotud mõningate spetsiaalsete valdkondade ja probleemiseadetega, mis on küll statistika ja tõenäosusteooria alusel arenenud, kuid mida käesoleval ajal loetakse iseseisvateks valdkondadeks. Anname neist lühikese ülevaate.

Aegridade analüüsile on pühendatud protseduurid allmenüüdest *Time Series Analysis, Smoothing* ja *Forecasting*. Nimetatud protseduuride loetelu on üpriski pikk, haarates 25 nimetust. Umbes pooled neist (allmenüüst *Time Series Analysis*) on küllaltki väikesemahulised, sisaldades vaid suhteliselt lihtsate tehete graafilisi illustratsioone. Seevastu aga osa protseduure (eriti allmenüüs *Forecasting*) realiseerivad üpris keerukaid matemaatilisi mudeleid. Iseloomulik on siinjuures üldine eeldus mõõtmisperioodide võrdsetest vahedest (konstantne samm). Ka graafiline protseduur *Component Line Chart* allmenüüst *Plotting Functions* kuulub sisuliselt aegridade analüüsi ja kirjeldamise juurde.

Kvaliteedi kontrolli meetoditele on pühendatud allmenüü *Quality Control*, kus on realiseeritud rida suhteliselt lihtsaid protseduure kontrollimaks kvaliteedi kontrolli teooriale vastavaid statistilisi hüpoteese, mis on üldiselt sõnastatavad järgmiselt:

H_0: jaotuse parameeter vastab etteantud standardile,

H_1: jaotuse parameeter erineb antud standardist,

kusjuures otsustamine toimub enamasti järjendanalüüsi põhimõttel, iga üksikvaatluse tulemust on võimalik valimisse lisada ja iga sammu järel hüpoteese uuesti kontrollida.

Katseplaneerimisele (allmenüü *Experimental Design*) on pühendatud neli protseduuri, neist üks (*Response Surface Plotting*) sisaldab üksnes graafilist väljundit. Siin on võimalik koostada mõningaid katseplaane (nt täis- ja murdfaktorplaane) ning pärast katse teostamist ja katsetulemuste sisestamist hinnata (dispersioonanalüüsi alusel) faktorite mõju olulisust.

Allmenüü *Sampling* 'valimi määramine' võimaldab hinnata eelinformatsiooni või pilotaažuuringu põhjal valimi mahtu, mis on teatavates tingimustes vajalik sisuka hüpoteesi tõestamiseks antud olulisuse nivool.

Viimane allmenüü *Mathematical Functions* 'matemaatilised funktsioonid' sisaldab loetelu üldkasutatavatest matemaatikaprotseduuridest - numbriline diferentseerimine ja integreerimine, võrrandisüsteemide ja omaväärtusülesande lahendamine, kiire Fourier' teisendus, juurte ja faktoriaalide arvutamine, algarvude leidmine, lineaarplaneerimise ülesande lahendamine simpleksmeetodil. Need kõik pakuvad iseseisvat huvi ning võimaldavad lahendada väikesi statistikavälised ülesandeid. Olulisem on aga nähtavasti võimalus andmeanalüüsi tulemusi rikastada ja paindlikumaks muuta mitmesuguste mittestandardsete võtete (teisenduste, protseduuride) kasutamise teel.

Autorid ei seo end lubadusega kirjeldada edaspidi ka neid SG osi põhjalikumalt.

Tartu, oktoober 1991

Kirjandus

1. Dixon, W. J., Massey, F. J. Introduction to Statistical Analysis. New York, 1957.

2. Durbin, J., Watson, G. S. Testing for Serial Correlation in Least Squares Regression. - Biometrika, 1951, 37, 409-428.

3. Parring, A.-M. Sissejuhatus matemaatilisse statistikasse. Tartu, 1989.

4. Parring, A.-M. Statistilise andmetöötluse algõpetus. Tartu, 1991.

5. STATGRAPHICS. Statistical graphics system by Statistical Graphics Corporation. TUTORIAL. A PLUS * WARE ® PRODUCT. STSC, Inc., 1987.

6. Tiit, E.-M. Andmeanalüüs I-II. Tartu, 1982, 1983.

7. Tiit, E.-M. Andmeanalüüs personaalarvutil programmipaki STATGRAPHICS abil. I osa. Põhimõisted. SG andmekäsitlus. Tartu, 1991.

8. Tiit, E. Matemaatiline statistika I. Tartu, 1972.

9. Tiit, E., Parring, A.-M., Matemaatiline statistika II-IV. Tartu, 1973-1977.

10. Tiit, E.-M., Parring, A.-M., Möls, T. Tõenäosusteooria ja matemaatiline statistika. Tallinn, 1977.

Rbl.17.-