H. Tammet

STATISTIKA-MEETODID ARVUTI JAOKS II

KASUTAJALE

Digikoopia eessõna

Käesolev fail on osaline fotokoopia 1976-ndal aastal ilmunud raamatust, milles kirjeldatakse Tallinna Pedagoogilises Instituudis (TpedI) koostatud statistikaprogrammide süsteemi. Programmisüsteem on orienteeritud rakendustele pedagoogilises ja sotsioloogilises uurimistöös ning selle koostamisel peeti silmas TPedI, ENSV Pedagoogika Teadusliku Uurimise Instituudi (PTUI) ja Vabariikliku Õpetajate Täiendusinstituudi vajadusi ning võimalusi. Süsteemi kasutajate ettevalmistamiseks pidasin kümnest loengust koosneva erikursuse TPedI õppejõududele (78 kuulajat) ning eraldi loengukursuse ka PTUI teaduritele. Käsiraamat ilmus ametkondliku rotaprinditrükisena ja suur osa selle tiraažist ladustati TPedI raamatukogu hoidlasse. Raamatuhoidlas puhkes tulekahju, raamatud said kustutusvee läbi kannatada ja hävitati. Sellepärast on raamatust säilinud vähe eksemplare.

Programmisüsteemi tehniliseks baasiks oli insenertehniliste arvutusete jaoks projekteeritud väikese võimsusega arvuti Nairii-2, mille arvutuskiirus oli ca 1000 ujukomatehet sekundis, operatiivmälu ca 10 kB, välismäluks paberist perfolint (väljastamiskiirus 80 baiti sekundis ja lugemiskiirus 150 baiti sekundis) ning printeriks elektriline kirjutusmasin kiirusega 10 tähemärki sekundis. Masinkoodis kirjutatud programmisüsteem suutis selle arvuti abil teha faktoranalüüsi kuni 30 tunnuse jaoks, mis ei jäänud oluliselt alla seitsmekümnendate aastate suurte arvutite statistikasüsteemide piirangutele.

TPedI sai ühe arvuti Nairii-2 aastal 1971 ja teise samasuguse aastal 1975. Statistikaprogrammid leidsid aktiivset kasutamist ja koormasid TPedI arvuteid ca 1000 tundi aastas. Kolme aasta 1975–1977 jooksul töödeldud statistiliste andmestike arv oli:

	Raamatus kirjeldatud statistikasüsteemi abil TpedI-s	Teistsuguse tarkvara abil Tartu Ülikoolis
Väikesed andmestikud	54	99
Keskmised ja suured andmestikud	122	113

40 aastat vana programmisüsteem omab ainult ajaloolist väärtust. Käsiraamat sisaldab aga ka olulisemate statistikameetodite sissejuhatavat kirjeldust. Digikoopiasse ongi võetud ainult meetodeid kirjeldav ja konkreetsest arvutist ning programmidest sõltumatu osa, mis ei aegu.

Hannes Tammet
30. aprill 2015

Eesti NSV Haridusministeerium
Eesti NSV Pedagoogika Teadusliku Uurimise Instituut

H. Tammet

S T A T I S T I K A M E E T O D I D

arvuti NAIRII-2 kasutajale

"Valgus" Tallinn 1976

S I S U K O R D

L I S A D

SISSEJUHATUS

Matemaatiline statistika on võitnud kindla koha mitme teadusala uurimismeetodite arsenalis. Nende teaduste hulka kuulub ka pedagoogika. Matemaatilise statistika formaalne aparatuur leiab pedagoogilise kogemuse üldistamisel enesele loomuliku rakenduse. Statistikameetodite tarvitamist pidurdas varem arvutustöö tülikus, arvutustehnika ja programmvarustuse areng aga kõrvaldab selle kitsaskoha. E. Vilde nim. Tallinna Pedagoogilise Instituudi käsutuses on juba mõnd aega elektronarvuti NAIRII-2. 1975.a. valmis pedagoogikauurimistöö ülesannete lahendamiseks sobiv statistikaprogrammide süsteem. Käesolev raamat kavatseb tutvustada nimetatud programmisüsteemi rakendusvõimalusi laiemale pedagoogikauurijate ja -üliõpilaste ringkonnale.

Matemaatilise statistika kasutamisest ei maksa loota revolutsiooni uurimistöö produktiivsuses. Tõuseb kõigepealt tulemuste kvaliteet. Matematiseerumist peetakse teaduse küpsuse tundemärgiks. Niiviisi kindlustab statistikameetodite kasutamine pedagoogika kui teaduse reputatsiooni.

Pedagoogikaprobleemi uurimise matemaatilise statistika kaasabil võib jagada nelja etappi ehk staadiumi: 1) formaliseerimine, 2) andmekogumine, 3) andmetöötlus, 4) tõlgendamine. Esimesel etapil koostatakse matemaatiline mudel, mis võimaldab lõpmata keerulist reaalset objekti kirjeldada lõpliku arvuhulga abil, peegeldades seejuures vaatlusaluse probleemi seisukohalt olulisi omadusi. Formaliseerimise tulemusena saadakse matemaatiline mudel, mis konkreetsete arvudega täidetakse andmekogumise etapil. Andmetöötluse käigus opereeritakse ainult formaalse mudeliga ning tehakse järeldusi mudeli piirides. Tõlgendamine seab arvutustulemused vastavusse reaalse uurimisobjektiga. Alles siin jõuame sisuliste järeldusteni.

Matemaatilised meetodid töötavad vaid uurimistöö formaliseeritud etappidel ja seepärast ei anna arvutusmeetodite tundmine ja korrektne tarvitamine iseenesest uurimistööle sisulist väärtust. Uurimistöö väärtus oleneb kõigepealt formaliseerimis-

etapi ja tõlgendusetapi lahendustest. Halb on, kui matemaatikat pruugitakse saamatult, veel halvem, kui matemaatilisele töötlusele ei järgne asjatundlikku sisulist analüüsi. Formaliseerimine ja arvutustulemuste tõlgendamine on igas uurimistöös spetsiifiline ja nõuab kõigepealt probleemi sisu tundmist. Seetõttu pole võimalik formaliseerimis- ja tõlgendamisküsimusi käesolevas raamatus piisavalt käsitleda.

Matemaatiline statistika on tõsine distsipliin ning ka algteadmiste omandamiseks kulub humanitaarharidusega uurijal palju vaeva. Siit tekib kiusatus korraldada uurimistöö matemaatikust konsultandi abiga,ise statistikameetodeid tundmata. Niiviisi aga töö ei laabu. Formaliseerimis- ja tõlgendamisetappi ei õnnestu veeretada konsultandi õlule. Kui statistikamõisted pole iseenesel selged ega andmetöötluse võimalused tuttavad, ei tule matemaatilise mudeli koostamisest ega ka andmetöötluse tulemuste tõlgendamisest midagi head. Öeldu ei tähenda, et uurija peaks loobuma konsultandi abist. Et küsida, peab aga teadma, mida ja kuidas küsida. Käesolev raamat sisaldabki parajasti niisugust statistikateadmiste miinimumi, mis on tarvilik konsultandi poole pöördumisel.

Esimeses osas tutvustatakse matemaatilise statistika mõisteid ja meetodeid niipalju, kui on vahetult tarvilik NAIRII-2 statistikaprogrammide süsteemi kasutajale. Erinevalt elektronarvutite-eelsetest käsiraamatutest ei kirjeldata siin üldse arvutusretsepte. Käsitlusviis on kohandatud matemaatilise ettevalmistuseta lugejale.

Teine osa teeb lugejale kättesaadavaks E. Vilde nimelises Tallinna Pedagoogilises Instituudis koostatud statistikaprogrammide tarvitamisjuhendid. Ka mõni statistikameetod leiab siin üksikasjalikumat käsitlemist kui esimeses osas.

Lisades on esitatud teises osas kirjeldatud programmide abil saadud arvutustulemuste näidised.

Kirjeldatav programmisüsteem on koostatud E.Vilde nimelises Tallinna Pedagoogilises Instituudis ENSV Pedagoogika Teadusliku Uurimise Instituudi ja Vabariikliku Õpetajate Täiendusinstituudi toetusel. Tööst võtsid osa matemaatikud Linda Pallas, Ann Indla ja Satu-Orvokki Vaikla. Käsikirja viimistlemisel osutas autorile tõsist abi Ann Indla.

Esimene osa

M E E T O D I D

1.1. TÕENÄOSUSTEOORIA ALGMÕISTED

Sündmus on tõenäosusteooria käsituses katse tulemus. Katse võib olla mündi heitmine kulli ja kirja mängus, sündmus - mündi peatumine, kiri ülal. Konkreetsed katsed ja sündmused on korduvmatud. Katse üldisemas mõttes võib olla korratav. Näiteks mündi heitmise katse on hõlpsalt korratav kasvõi sada korda järjest. Veel üldisema käsitluse juures kujutletakse katset, mida on võimalik korrata muutumatutes tingimustes kuitahes palju kordi. Niisugused kujutletavad katsed ja sündmused ongi tõenäosusteooria uurimisobjekt.

Juhuslik sündmus on sündmus, mille toimumises ei saa ette kindel olla, näiteks: "Andres saab matemaatika kontrolltöö hindeks "nelja"".

Suhteline sagedus on juhusliku sündmuse toimumiskordade t ja korduskatsete arvu k suhe t/k. Suhtelist sagedust võib avaldada nii absoluutarvuna kui protsentides. Mündi kümnekordsel heitmisel võime saada kirja suhteliseks sageduseks 0,4 ehk 40 %.

Tõenäosus on suurus, millele suhteline sagedus läheneb katse piiramatul kordamisel. Mündi heitmisel on kirja tõenäosus 50%. Sündmuse "Andres saab matemaatika kontrolltöö hindeks "nelja"" tõenäosus võib olla näiteks 63 %. Kummagi nimetatud sündmuse tõenäosust ei saa täpselt mõõta. Esimene katse on paremini korratav ja siin näib tõenäosus olevat katseliselt määratav. Sügavamal järelemõtlemisel on aga kerge veenduda, et erinevus kahe näite vahel pole põhimõttelist laadi.

Arutlustest tõenäosuse olemuse üle siinkohal loobutakse. Asjast huvitatud lugejale soovitame pöörduda ükskõik millise tõenäosusteooria õpiku poole.

Juhuslik suurus on suurus, mille väärtus oleneb juhuslikest

sündmustest. Juhuslikud on näiteks Andrese järgmise matemaatika kontrolltöö hinne ja kolmeteistkümnenda tänaval vastu tuleva inimese pikkus.

Diskreetse juhusliku suuruse, näiteks kontrolltöö hinde võimalike väärtuste hulga kahe naaberväärtuse vahel pole vahepealseid väärtusi. Diskreetse juhusliku suuruse võimalike väärtuste hulk on tavaliselt lõplik.

Pideva juhusliku suuruse, näiteks inimese pikkuse kahe erineva võimaliku väärtuse vahele mahub alati kolmas võimalik väärtus. Pideva juhusliku suuruse võimalike väärtuste hulk on lõpmatu.

Ümardatud juhuslik suurus saadakse pidevast juhuslikust suurusest viimase väärtuste ümardamisel. Inimese pikkus ümardatakse tavaliselt täissentimeetritesse. Ümardatud juhuslik suurus on diskreetse juhusliku suuruse erijuhtum. Niisugune päritolu lubab teoorias vaadelda paralleelselt ümardatud suurusega tema ümardamata pidevat originaali.

Statistilise analüüsi praktikas tuleb meil tegemist ainult diskreetsete suurustega, millest osa on aga päritolult ümardatud suurused. Teoorias käsitleme paralleelselt diskreetsete suurustega ka pidevaid juhuslikke suurusi ja lubame endale mõnel juhul vabadust käsitleda diskreetseid suurusi ümardatud suurustena ka siis, kui uuritav nähtus annab selleks vähe alust.

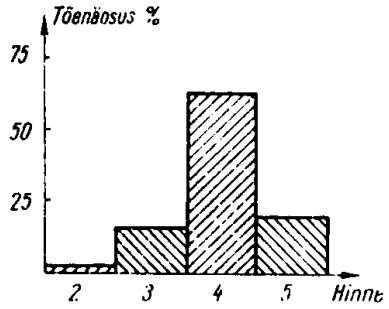
Diskreetne tõenäosusjaotus. Sündmuse "diskreetne juhuslik suurus saab katses konkreetse väärtuse" tõenäosus oleneb vaatlusalusest väärtusest. Tabel, kus iga võimaliku väärtuse järel on näidatud vastav tõenäosus, kirjeldab tõenäosusjaotust (vt. tabel 1). Näitlikumalt saab tõenäosusjaotust kirjeldada tulpdiagrammi abil (vt. joonis 1).

Pidev tõenäosusjaotus. Pideva juhusliku suuruse ümardamise korral saab tõenäosusjaotust kujutada tulpdiagrammi abil. Mida vähem ümardada, seda tihedamalt on diagrammil tulpi. Piirjuhul muutub diagrammi ülaserv sujuvaks kõveraks. See kõver kujutabki pidevat tõenäosusjaotust (vt. joon. 2). Kõvera punkti kõrgust argumenditeljest nimetatakse tõenäosustiheduseks. Funktsiooni $f(\xi)$, mis kirjeldab tõenäosustiheduse sõltuvust argumendist ξ,

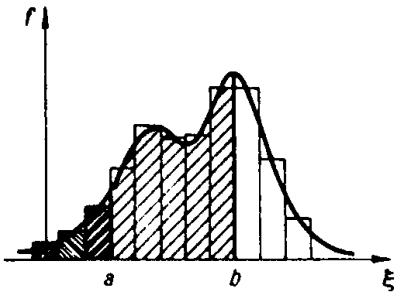
Tabel 1

Kontrolltöö hinde tõenäosusjaotus (näide)

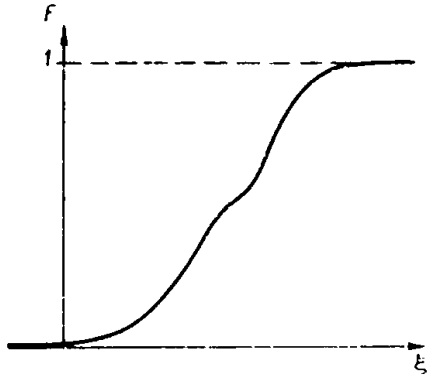
Hinne	Tõenäosus
2	2 %
3	15 %
4	63 %
5	20 %



Joon. 1. Kontrolltöö hinde tõenäosusjaotuse tulpdiagramm tabeli 1 järgi.



Joon. 2. Pideva ja ümardatud juhusliku suuruse tõenäosusjaotus.



Joon. 3. Pideva juhusliku suuruse (vt. joon.2) kumulatiivne jaotusfunktsioon.

nimetatakse juhusliku suuruse tihedusfunktsiooniks. Tõenäosusjaotust kujutav kõver on tihedusfunktsiooni graafik.

Vahemikku kuulumise tõenäosuse arvutamiseks on diskreetse suuruse korral tarvis vaid liita vahemikku kuuluvate väärtuste tõenäosused. Graafiliselt kujutab tõenäosuste summat vahemikku kuuluvate tulpade üldpindala jaotusdiagrammil. See tähelepanek võimaldab lihtsalt määrata vahemikku kuulumise tõenäosust ka pideva juhusliku suuruse jaoks. Tihedusfunktsiooni graafikule kantakse vahemikku piiravad vertikaaljooned (vt. joon. 2 jooned a ja b), nende joontega kõveraalusest kujundist eraldatud pindala näitabki otsitavat tõenäosust.

Jaotusfunktsioon. Olgu juhusliku suuruse tähis x. Sündmuse " $x < \xi$ " tõenäosust esitavat funktsiooni $F(\xi)$ nimetatakse juhusliku suuruse kumulatiivseks jaotusfunktsiooniks ehk lihtsalt

jaotusfunktsiooniks. Kumulatiivse jaotusfunktsiooni näide on esitatud joonisel 3.

Mitmemõõtmeline juhuslik suurus. Kui ühest juhuslikust sündmusest oleneb korraga mitme suuruse väärtus, siis käsitletakse neid suurusi tavaliselt komplektis. Üksiksuurusi nimetatakse komplekti komponentideks, komplekti, mille komponendid on juhuslikud suurused, mitmemõõtmeliseks juhuslikuks suuruseks, matemaatilises teoorias ka juhuslikuks vektoriks. Kui komplekt koosneb ühestainsast komponendist, räägitakse ühemõõtmelisest juhuslikust suurusest, see on tavalise juhusliku suuruse sünonüüm.

Mitmemõõtmelise juhusliku suuruse iga komponenti võib kirjeldada iseseisvalt kui tavalist juhuslikku suurust. Komplektne käsitlus osutub tarvilikuks vaid komponentidevaheliste seoste uurimisel.

1.2. MOMENDID

Keskväärtus on juhusliku suuruse väärtuste tõenäosuskeskpunkt. Juhusliku suuruse oodatavaks väärtuseks ennustatakse tavaliselt keskväärtust, seepärast nimetatakse keskväärtust ka matemaatiliseks ooteväärtuseks.

Kui juhuslik suurus võib omada vaid kaht väärtust ξ_1 ja ξ_2, kumbagi võrdse tõenäosusega $p_1 = p_2 = 0{,}5$, on selle suuruse keskväärtus

$$\mu = p_1\xi_1 + p_2\xi_2 = \frac{\xi_1 + \xi_2}{2} \quad .$$

Üldisemal juhul võib diskreetne juhuslik suurus omandada n erinevat väärtust $\xi_1, \xi_2, \ldots, \xi_n$ tõenäosustega $p_1, p_2, \ldots, p_n$. Keskväärtus on siis

$$\mu = p_1\xi_1 + p_2\xi_2 + \ldots + p_n\xi_n \quad .$$

Edaspidiseks võtame kasutusele summa lühendatud tähistusviisi. Märk $\sum_{i=1}^{n}$ nõuab talle järgnevast avaldisest summa moodustamist, andes indeksile i väärtuseks 1, 2, ..., n. Keskväärtuse avaldis kirjutatakse järgmiselt:

$$\mu = \sum_{i=1}^{n} p_i\xi_i \qquad (1)$$

Keskväärtus ei ole juhuslik suurus. Kuna tõenäosusi ei saa katseliselt täpselt mõõta, pole ka keskväärtust võimalik katse teel täpselt määrata.

Pideva juhusliku suuruse keskväärtust on võimalik ligikaudselt arvutada ümardamisvõtte abil, täpselt aga ainult integreerimise teel

$$\mu = \int_{-\infty}^{+\infty} f(\xi)\,\xi\,d\xi \quad . \tag{2}$$

Juhuslike suuruste summa keskväärtus on alati võrdne liidetavate keskväärtuste summaga.

Ülesannetes, kus tegemist mitme juhusliku suuruse keskväärtustega, tähistatakse x keskväärtust μ_x ehk $\mu(x)$.

Dispersioon iseloomustab juhusliku suuruse hajuvust keskväärtuse ümber.

Nimetame juhusliku suuruse erinevuse keskväärtusest juhusliku suuruse hälbeks. Hälve on juhuslik suurus. Osa hälbeid on positiivsed, osa negatiivsed ja hälbe keskväärtus on null. Öeldakse, et hälve on tsentreeritud. Hälbe ruut pole kunagi negatiivne. Dispersioon σ^2 on hälbe ruudu keskväärtus. Diskreetse suuruse puhul

$$\sigma^2 = \sum_{i=1}^{n} p_i\left(\xi_i - \mu\right)^2 \quad , \tag{3}$$

pideva juhusliku suuruse puhul

$$\sigma^2 = \int_{-\infty}^{+\infty} f(\xi)\left(\xi - \mu\right)^2 d\xi \quad . \tag{4}$$

Kaht juhuslikku suurust nimetatakse teineteisest sõltumatuteks siis, kui ühe suuruse konkreetse väärtuse määramine ei anna mingit informatsiooni teise suuruse väärtuse ennustamiseks. Kahe sõltumatu juhusliku suuruse summa dispersioon võrdub liidetavate dispersioonide summaga. Nii kasulikku omadust pole ühelgi teisel juhusliku suuruse hajuvust iseloomustaval näitajal.

Dispersiooni mõõtühikuks on juhusliku suuruse mõõtühiku ruut, seetõttu pole dispersioon otseselt hälvetega võrreldav.

Standardhälve σ võrdub ruutjuurega dispersioonist

$$\sigma = \sqrt{\sigma^2} \quad . \tag{5}$$

Ruutjuur võrdsustab mõõtühiku juhusliku suuruse mõõtühikuga ja standardhälve on otseselt hälvetega võrreldav. Tavaliselt on 65...70 % hälvetest absoluutväärtuse poolest standardhälbest väiksemad ja 30...35 % suuremad. Standardhälvet kolm ja enam korda ületavad hälbed on väga haruldased. Sobiv mõõtühik ja lihtne seos dispersiooniga teevad standardhälbest kõige tarvitatavama hajuvuse mõõdu.

Kahe sõltumatu juhusliku suuruse x ja y summa z standardhälve arvutatakse liidetavate standardhälvete σ_x ja σ_y järgi Pythagorase valemi abil

$$\sigma_z = \sqrt{\sigma_x^2 + \sigma_y^2} \qquad (6)$$

Variatsioonikordajaks nimetatakse juhusliku suuruse standardhälbe ja keskväärtuse suhet

$$v = \frac{\sigma}{\mu} \quad , \qquad (7)$$

mis sageli avaldatakse protsentides. Variatsioonikordaja ei sõltu juhusliku suuruse mõõtühikute valikust ja teda võib teisiti nimetada suhteliseks standardhälbeks.

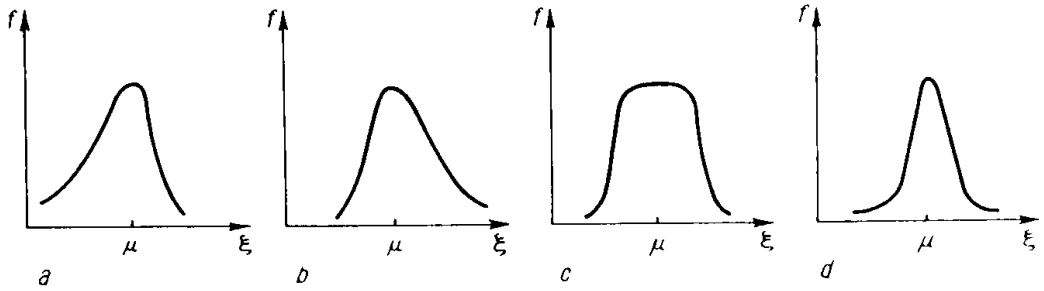
Variatsioonikordajat kasutatakse ainult niisuguste juhuslike suuruste kirjeldamisel, mille väärtus on alati positiivne.

Algmomendiks nimetatakse juhusliku suuruse mingi astme keskväärtust. Valitud astet nimetatakse momendi järguks. Keskväärtus ise on esimest järku algmoment.

Tsentraalmomendiks nimetatakse juhusliku suuruse hälbe mingi astme keskväärtust. Esimest järku tsentraalmoment on alati null. Dispersioon on teist järku tsentraalmoment. Kolmandat järku tsentraalmomenti nimetatakse asümmeetriaks ja neljandat järku tsentraalmomenti ekstsessiks. Asümmeetria mõõtühikuks on juhusliku suuruse mõõtühiku kuup, ekstsessi mõõtühikuks neljas aste.

Asümmeetriakordaja on asümmeetria ja standardhälbe kuubi suhe ega sõltu mõõtühikute valikust. Jaotust nimetatakse sümmeetriliseks siis, kui jaotuskõver on keskväärtuse suhtes peegelsümmeetriline (vt. joon. 4). Sümmeetrilise jaotuse asümmeetriakordaja on null, vastupidine väide pole aga alati õige. Vasa-

kult väljavenitatud jaotuse asümmeetria on negatiivne, paremalt väljavenitatud jaotuse oma positiivne (vt. joon. 4). Niiviisi näitab asümmeetriakordaja märk, kummal pool on suured hälbed tõenäolisemad.



Joon. 4. Jaotuskõverate näiteid:
a - negatiivse asümmeetriaga jaotus, b - positiivse asümmeetriaga jaotus, c - väikese ekstsessiga sümmeetriline jaotus, d - suure ekstsessiga sümmeetriline jaotus.

Ekstsessikordaja on ekstsessi ja standardhälbe neljanda astme suhe ega sõltu samuti mõõtühikute valikust. Ekstsessikordaja pole kunagi väiksem kui üks. Järsult lõigatud tiibadega jaotuse ekstsessikordaja on väike, väljavenitatud tiibadega jaotuse ekstsessikordaja suur (vt. joon. 4). Seega iseloomustab ekstsessikordaja suurte hälvete tõenäosust.

Paljud autorid kasutavad ekstsessi iseloomustamiseks näitarvu, mis on ülaldefineeritud ekstsessikordajast kolme võrra väiksem.

1.3. TÜÜPJAOTUSED

Diskreetse ühtlase jaotuse korral on juhusliku suuruse kõik võimalikud väärtused võrdtõenäolised. Niisuguse omadusega on näiteks täringu heitmisel saadav silmade arv. Ühtlane jaotus esineb loteriitüüpi katsetes.

Alternatiivne ühtlane jaotus on diskreetse ühtlase jaotuse sageli ettetulev erijuht. Võimalikke väärtusi on siin kaks, tähistame need a ja b. Keskväärtus on $(a+b)/2$, standardhälve $|b-a|/2$, asümmeetria null ja ekstsessikordaja 1. Püstkriipsulud

standardhälbe valemis tähendavad, et avaldisest b-a tuleb võtta absoluutväärtus, s.o. negatiivse väärtuse puhul miinusmärk ära jätta.

Pideva ühtlase jaotusega juhusliku suuruse tõenäosustihedus on mingis vahemikus (a, b) kõikjal ühesugune, väljaspool seda vahemikku aga null. Kirjutis (a, b) tähistab kõigi nende väärtuste hulka, mis on suuremad kui a ning väiksemad kui b. Väärtusi a ja b nimetatakse vahemiku rajadeks. Ühtlase jaotuse kõver on ristkülikukujuline. Keskväärtus võrdub (a+b)/2, standardhälve 0,58(b-a)/2, asümmeetriakordaja on null ja ekstsessikordaja 1,8.

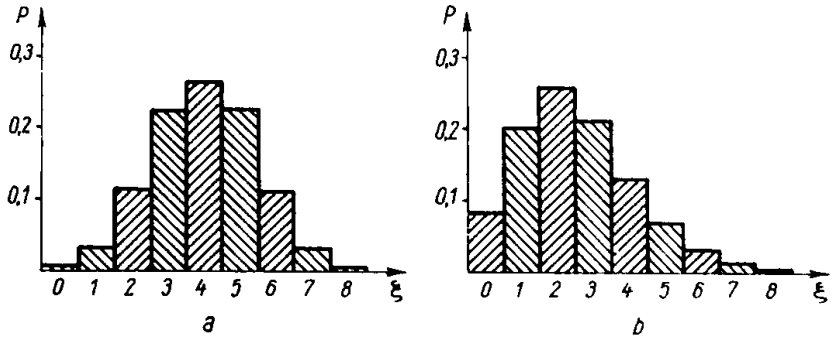
Binomiaaljaotusele allub sõltumatute juhusliku tulemusega katsete õnnestumiste arv kindlast katsete hulgast koosnevas katseseerias. Tähistame ühe katse õnnestumise tõenäosuse p ja katsete arvu n. Siis on õnnestunud katsete arvu keskväärtus $\mu = pn$ ja standardhälve $\sigma = \sqrt{(1-p)\mu}$. ξ õnnestumisjuhu tõenäosus on

$$p(\xi) = \frac{n!}{\xi!(n-\xi)!}\, p^{\xi}(1-p)^{n-\xi} \quad . \qquad (8)$$

Faktoriaalidest moodustatud tegur on binomiaalkordaja, siit jaotuse nimi.

Joonisel 5 on näidatud binomiaaljaotuse tõenäosused juhul p = 0,5, n = 8.

Kui keskväärtus ületab kümmet, käsitletakse binomiaalselt jaotatud juhuslikku suurust tavaliselt kui normaalselt jaotatud juhusliku suuruse ümardatud väärtust.



Joon. 5. Diskreetseid tõenäosusjaotusi:
a - binomiaaljaotus p = 0,5, n = 8; b - Poissoni jaotus $\mu = 2,5$.

Poissoni jaotusele (loe: puassooni) allub sõltumatute sündmuste arv juhul, kui võimalikke sündmusi on väga palju, iga üksiksündmuse tõenäosus aga väga väike. Praktikas tuleb sageli ette juhuslikke suurusi, mis on jaotatud ligilähedaselt Poissoni jaotusele. Näiteks vigade arv etteütluses, üle 190 cm pikkuste poiste arv koolis jne. Kui Poissoni jaotusele alluva suuruse keskväärtus on μ, siis standardhälve on

$$\sigma = \sqrt{\mu} \quad . \tag{9}$$

Tõenäosuste arvutamise valem

$$p(\xi) = \frac{\mu^{\xi}}{\xi!} e^{-\mu} \tag{10}$$

Jaotusdiagrammi näide on joonisel 5.

Kui keskväärtus ületab kümmet, käsitletakse Poissoni jaotusega juhuslikku suurust sageli kui normaalselt jaotatud suuruse ümardatud väärtust.

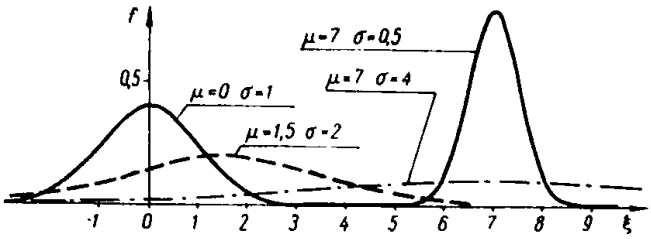
Normaaljaotus ehk Gaussi jaotus on tähtsaim pideva juhusliku suuruse tüüpjaotus (vt. p. 1.4). Jaotuskõver sarnaneb binomiaaljaotuse diagrammiga. Katsete arvu suurenemisel lähenebki binomiaaljaotus normaaljaotusele.

Normaaljaotuse keskväärtus μ ja standardhälve σ on mõlemad vabad parameetrid. Tihedusfunktsioon avaldub valemiga

$$f(\xi) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\xi-\mu)^2}{2\sigma^2}} \tag{11}$$

Normaaljaotuse asümmeetria on null ja ekstsessikordaja täpselt 3. Absoluutväärtuse poolest standardhälbest suuremate hälvete tõenäosus on 32 %, üle kahe korra suuremate hälvete tõenäosus 4,5 % üle kolme korra suuremate hälvete tõenäosus 0,3 %. Põhimõtteliselt on võimalikud kuitahes suured hälbed, nende tõenäosus on aga kaduvväike.

Joonisel 6 on esitatud näitena mõned normaaltihedusfunktsiooni graafikud. Maksimum asub kohal μ ja kõvera käänupunktid kohtadel $\mu-\sigma$ ja $\mu+\sigma$.



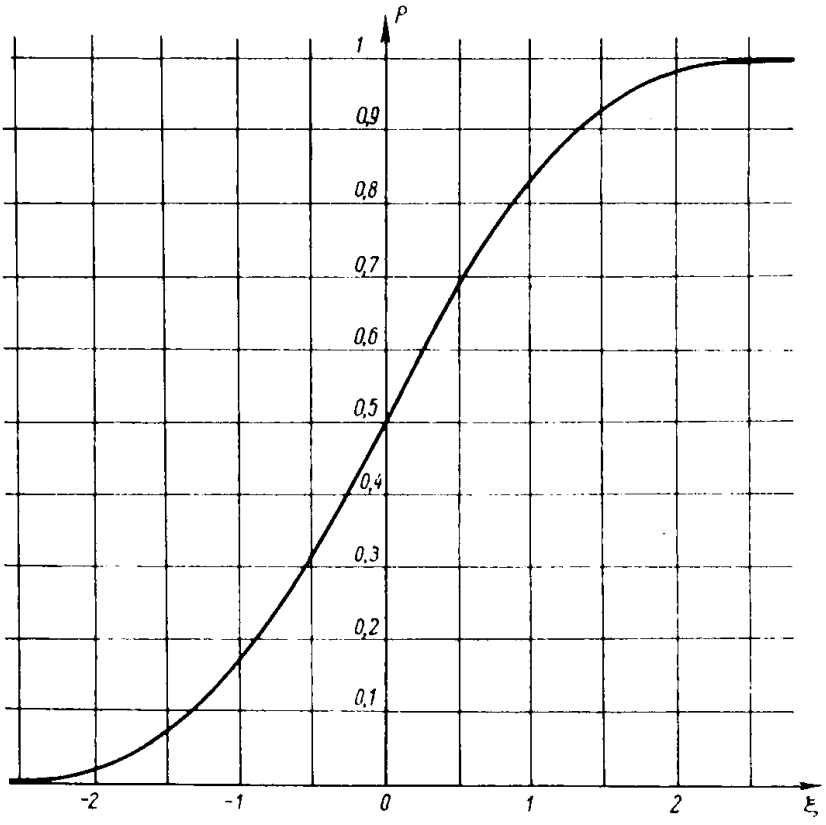
Joon. 6. Normaaljaotuse tihedusfunktsioonide graafikud (näited).

Standardiseeritud normaaljaotuseks nimetatakse normaaljaotust parameetritega $\mu = 0$, $\sigma = 1$. Kui normaaljaotusega juhuslikul suurusel x on teistsugused parameetrid, siis teisendatud suuruse

$$z = \frac{x - \mu}{\sigma} \qquad (12)$$

jaotus on standardiseeritud normaaljaotus. Et mõõtskaala teisendus (12) on küllalt lihtne, kasutatakse seda kõigis normaaljaotusega seotud praktilistes arvutustes. Standardiseeritud normaaljaotuse tihedusfunktsiooni ja jaotusfunktsiooni tabeleid võib leida igast tõenäosusteooria või statistika tabelitekogust ning ka paljudest õpikutest. Jaotusfunktsiooni nimetatakse lühemalt normaaljaotusfunktsiooniks, tõenäosusintegraaliks ehk Laplace'i funktsiooniks ja tähistatakse tavaliselt $\Phi(\xi)$. $\Phi(\xi)$ on tõenäosus selleks, et standardiseeritud normaaljaotusega juhusliku suuruse väärtus on väiksem argumendist ξ . Normaaljaotusfunktsiooni graafik on kujutatud joonisel 7.

Paljudes ülesannetes on tarvis teada argumendi väärtust, mille juures normaaljaotusfunktsioon saavutab antud taseme p. Võrrandi $p = \Phi(\xi)$ lahendit $\xi = \Psi(p)$ nimetatakse normaaljaotusfunktsiooni pöördfunktsiooniks ehk normaalkvantiilfunktsiooniks. Ka see funktsioon on tabuleeritud. NAIRII-2 statistikaprogrammide süsteem sisaldab kõigi nimetatud funktsioonide arvutamise alamprogramme ning vabastab arvutaja tabelite kasutamise vajadusest.



Joon. 7. Graafik normaaljaotusfunktsiooni $p = \Phi(\xi)$ ja normaalkvantiilfunktsiooni $\xi = \Psi(p)$ väärtuste leidmiseks.

1.4. TÕENÄOSUSTEOORIA SEADUSI

Tšebõšovi võrratus. Tõenäosus, et juhusliku suuruse hälve keskväärtusest on absoluutväärtuse poolest k või enam korda suurem standardhälbest, oleneb jaotusseadusest. Tšebõšov tõestas, et ühegi lõpliku keskväärtuse ja standardhälbega jaotuse korral pole see tõenäosus suurem kui $1/k^2$. Tšebõšovi võrratusel on praktilist huvi pakkuvate jaotusseaduste korral väga suur tagavara. Näiteks garanteerib see, et standardhälvet enam kui kolmekordselt ületavate hälvete tõenäosus pole üle 11,1 %, normaaljaotuse korral on aga nimetatud tõenäosus vaid 0,3 %.

Aritmeetilise keskmise omadused järelduvad summa keskväärtuse ja dispersiooni arvutamise reeglitest (vt. p. 1.2). Juhusliku suuruse x n erineva sõltumatu väärtuse ehk n ühesuguse jao-

tusega juhusliku suuruse x_1, x_2, ... x_n aritmeetiline keskmine avaldub kujul

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} . \qquad (13)$$

Aritmeetiline keskmine on juhuslik suurus keskväärtusega $\mu_{\bar{x}}$ ja standardhälbega $\sigma_{\bar{x}}$. Keskväärtuse arvutamiseks liidame avaldise (13) lugeja liidetavate keskväärtused. Niiviisi saame keskväärtuse lugejaks n ning aritmeetilise keskmise keskväärtus osutub võrdseks keskmistatava suuruse keskväärtusega:

$$\mu_{\bar{x}} = \mu$$

Dispersioonide liitmisel saame lugeja dispersiooniks $n\sigma^2$ ja standardhälbeks $\sqrt{n}\,\sigma$. Aritmeetilise keskmise standardhälve on lugeja standardhälbest n korda väiksem

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad (14)$$

Keskmistamine vähendab hajuvust. Nelja väärtuse keskmistamisel väheneb standardhälve kaks korda, saja väärtuse keskmistamisel kümme korda.

Suurte arvude seadus väidab: keskmistatavate väärtuste arvu tõkestamatul suurendamisel tõenäosus, et aritmeetilise keskmise erinevus keskmistatava suuruse keskväärtusest ületaks kuitahes väikest konstantset positiivset suurust, läheneb nullile. Seadus kehtib lõpliku keskväärtuse ja standardhälbega jaotuste korral ning järeldub Tšebõšovi võrratusest ja aritmeetilise keskmise omadustest. Arvu n suurendamisel läheneb aritmeetilise keskmise standardhälve nullile (valem 14) ja kuitahes väikese konstantse positiivse hälbe suhe standardhälbesse kasvab tõkestamatult. Tšebõšovi võrratuse järgi kahaneb siis tõkestamatult tõenäosus, et aritmeetilise keskmise hälve aritmeetilise keskmise keskväärtusest ületaks vaadeldavat konstantset suurust. Et aritmeetilise keskmise keskväärtus võrdub keskmistatava suuruse keskväärtusega, on suurte arvude seadus ülaltoodud arutlusega tõestatud.

Tsentraalne piirteoreem. Olgu x_1, x_2, ..., x_n lõpliku keskväärtuse ja standardhälbega iseseisvad sõltumatud juhuslikud suurused, mis võivad alluda igaüks erinevale jaotusseadusele, ja

$x = x_1 + x_2 + \ldots + x_n$ nende summa. Tsentraalne piirteoreem väidab, et liidetavate arvu tõkestamatul suurendamisel ja nende samaaegsel ühtlasel vähendamisel läheneb summa jaotus alati normaaljaotusele, millised ka poleks liidetavate jaotused. Ühtlane vähendamine tähendab seda, et iga liidetava dispersiooni suhe dispersioonide summasse peab lähenema nullile.

Praktikas on juba nelja-viie võrdse standardhälbega ühtlaselt jaotatud liidetava summa jaotus eristamatu normaaljaotusest.

Tsentraalse piirteoreemi rakendamisel on komistuskiviks mõned "salakavalad" jaotused, millel pole üldse lõplikku standardhälvet. Niisugune on näiteks jaotuskõvera kuju poolest normaaljaotusega üsna sarnane Cauchy jaotus (tihedusfunktsioon $f(\xi) = \frac{1}{\pi(1+\xi^2)}$). Cauchy jaotusega liidetavate summa on alati Cauchy jaotusega ning siin tsentraalne piirteoreem ei kehti.

Normaaljaotushüpotees väidab, et uuritava juhusliku suuruse jaotus on normaaljaotus. Enamik statistilisi andmetöötlusmeetodeid lähtub normaaljaotushüpoteesist. Tegelik tõenäosusjaotus erineb alati vähem või rohkem teoreetilisest normaaljaotusest. Paljud normaaljaotushüpoteesile rajatud statistikameetodid on aga üsna tundetud jaotusseaduse hälbimise suhtes ja nende rakendamine pea alati õigustatud. See väide ei kehti sugugi kõigi meetodite kohta, mistõttu uurija peab oskama hinnata ühelt poolt statistikameetodite tundlikkust jaotusseaduse variatsioonide suhtes, teiselt poolt normaaljaotushüpoteesi ja tegeliku jaotuse kooskõla.

Normaaljaotushüpoteesi toetuseks võib öelda järgmist.

1) Kui uuritava suuruse juhuslikel hälvetel on palju ühtlaselt väikesi põhjusi, võib hälvet vaadelda kui paljude väikeste osahälvete summat. Normaaljaotushüpotees tugineb siis tsentraalsele piirteoreemile.

2) Andmetöötlusel opereeritakse peamiselt aritmeetiliste keskmistega. Aritmeetiline keskmine arvutatakse aga paljude ühtlaselt väikeste liidetavate summa kaudu ning on ligilähedaselt normaalselt jaotatud, olenemata uuritava suuruse jaotusseadusest.

3) Normaaljaotus on binomiaaljaotuse ja Poissoni jaotuse piirjaotus üleminekul pidevale mõõtskaalale.

4) Ühe ja sama standardhälbega jaotusseaduste hulgas lisab

normaaljaotusseadus keskväärtuse ja standardhälbega esitatud informatsioonile kõige vähem kõrvalinformatsiooni.

5) Normaaljaotus on ainus pidev tõenäosusjaotus, mille korral aritmeetilise keskmise arvutamine on parim moodus keskväärtuse hindamiseks. Seega on normaaljaotushüpotees teatud mõttes samaväärne soovitusega arvutada aritmeetiline keskmine.

1.5. MÕÕTSKAALAD

Mõõtmine seab reaalsele objektile vastavusse tunnuse formaalse väärtuse. Üldise käsitluse puhul nimetatakse mõõtmiseks igasugust väärtuse määramise protsessi. Väärtuseks võib olla mitte ainult arv, vaid ka nimi. Olenevalt mõõdetavast tunnusest sooritatakse mõõtmine vahetu vaatluse teel või aparatuuri abil.

Ka õpilaste teadmiste igapäevane kontroll koolis on mõõtmine, tulemuseks jõudlushinded.

Mõõtskaala on tunnuse võimalike väärtuste hulk, kus iga reaalne väärtus on varustatud formaalse nimega. Väärtuste formaalseteks nimedeks on enamasti mõõtarvud. Mõõtskaaladel on kolm põhitüüpi: nimeskaala, järjeskaala ja meetriline skaala. Konkreetse tunnuse mõõtskaala tüüp oleneb tunnuse väärtuste sisulisest interpreteerimisest, seepärast saab mõõtskaala tüübi üle otsust teha ainult uuritava probleemi sisu tundes.

Nimeskaala ehk nominaalskaala väärtustest pole ükski teistest suurem või väiksem, parem või halvem. Öeldakse, et väärtused pole võrreldavad. Väärtuste hulk on lõplik. Nimeskaalad on näiteks õpilaste soo skaala (väärtused: poiss, tüdruk), koolide skaala (väärtusteks on koolide numbrid), huvialade skaala jne.

Nimeskaala kasutamise korral nimetatakse mõõtmist sageli määramiseks ehk klassifitseerimiseks.

Nimeskaala on mõõtmistulemuste informatiivsuse poolest kõige madalama tasemega skaala. Kõiki teisi skaalasid on võimalik mõnest omadusest loobumise ja lihtsustamise teel degradeerida nimeskaalaks.

Järjeskaala ehk ordinaalskaala väärtustel on loomulik jär-

jekord ning neid saab paigutada ühesel viisil "pingeritta", mis algab kõige väiksemast ja lõpeb kõige suurema väärtusega. Väärtuste nimedeks on tavalised arvud, harvem alfabeetilises järjestuses tähed. Arvule lisatakse mõõtühiku kohale sageli sõna "palli". Skaala sammud pole võrreldavad: ei saa öelda, kas hüpe kolmelt pallilt neljale on suurem, võrdne või väiksem kui hüpe kahelt pallilt kolmele. Järjeskaala mõõtarvud on tingarvud ja aritmeetilised tehted nendega, näiteks liitmine ja keskmise arvutamine, ei oma mõtet.

Järjeskaala on näiteks koolihinnete skaala, millest järgnevas tuleb pikemalt juttu.

Meetrilise ehk kvantitatiivse skaala korral on täpselt teada, mitu korda üks skaalavahemik on suurem või väiksem kui teine. Kui nimetatud teadmisest loobuda, degradeerub meetriline skaala järjeskaalaks. Meetrilisele skaalale vastandamisel nimetatakse nimeskaalat ja järjeskaalat kvalitatiivseteks skaaladeks.

Ühtlase meetrilise skaala korral vastavad tunnuse väärtuse sisuliselt võrdsetele hüpetele mõõtarvude võrdsed hüpped. Ühtlaste meetriliste skaalade hulka kuuluvad pea kõik füüsikaliste suuruste mõõtskaalad, näiteks pikkuse ja massi skaalad. Eespool juhusliku suuruse mõistet käsitledes pidasime silmas peaasjalikult ühtlases meetrilises skaalas mõõdetavaid suurusi. Ühtlase meetrilise skaala mõõtarvude liitmisel ja lahutamisel saab tulemustele omistada sisulist mõtet.

Suhteskaala on ühtlase meetrilise skaala täiuslikum erijuht. Suhteskaalat saab rakendada siis, kui on võimalik määrata tunnuse reaalsete väärtuste suhet. Mõõtarvude suhe võrdub väärtuste suhtega. Pikkuse, massi ja enamiku teiste füüsikaliste suuruste skaalad on suhteskaalad.

Vaheskaala on ühtlase meetrilise skaala vähem täiuslik variant. Erinevalt suhteskaalast on vaheskaala nullpunkt kokkuleppeline ning mõõtarvude suhtele ei saa omistada sisulist mõtet. Vaheskaalas mõõdetakse näiteks kellaaega. Kuigi üks tund on teisega võrdne, ei saa öelda, et kell kaks oleks poole suurem kui kell üks.

Laiendatud skaala sisaldab põhiskaalat ja veel üht põhiskaa-

last väljaspool seisvat lisaelementi, mida nimetatakse küsimärgiks ehk lüngaks ja mis tähistab määramata jäänud väärtust.

Alternatiivseks eh dihhotoomseks nimetatakse ainult kahest väärtusest koosnevat skaalat. Alternatiivne skaala koostatakse tavaliselt kui nimeskaala (väärtuste paarid "ei - jah", "poiss - - tüdruk" jne.), aga teda võib alati kuulutada ühtlaseks meetriliseks skaalaks. Et skaalavahemikke on vaid üks, ei saa siin tekkida mingeid probleeme. Väärtuste mõõtarvude valik on suvaline, tavaliselt valitakse arvud 0 ja 1 või 1 ja 2.

Mõõtskaala metriseerimine tähendab meetrilise skaala konstrueerimist kvalitatiivses skaalas mõõdetud tunnuse jaoks.Näiteks tuule tugevust mõõdeti varem Beaufort'i järjeskaalas, nüüd aga ka kiiruse suhteskaalas. Meteoroloogia käsiraamatutest võib leida üleminekutabeli pallidelt meetritele sekundis. Metriseerimise võimalus võib põhjustada vaidlusi mõõtskaala tüübi määramisel. Iseloomulik näide on koolihinnete skaala probleem.

Mõne ülesande puhul osutub mõõtskaala metriseerimisel otstarbekaks regressioonihinnete meetod (vt. p. 1.19).

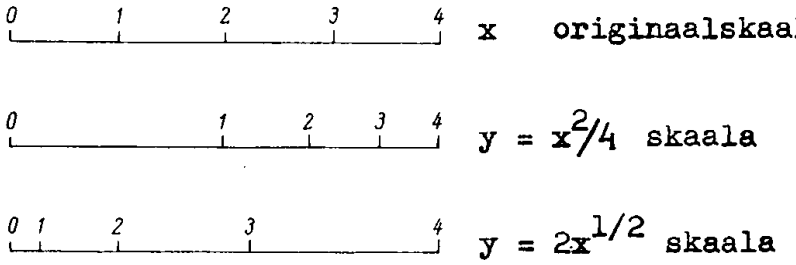
Koolihinnete skaala koosneb neljast väärtusest: 2, 3, 4 ja 5. Kui see oleks vaid järjeskaala, poleks võimalik hinnete aritmeetilisele keskmisele mingit sisulist tähendust omistada.Koolihinnete skaalat ühtlaseks meetriliseks skaalaks kuulutades peaks aga väitma, et kolme erinevus kahest on täpselt niisama suur kui nelja erinevus kolmest ja viie erinevus neljast. See väide on teatud mõttes samaväärne hinnete aritmeetilisele keskmisele tähenduse andmisega. Kolme ja viie keskmine on täpselt võrdne neljaga vaid siis, kui vahed kolme ja nelja ning nelja ja viie vahel on võrdsed. Vahede võrdsuse eeldust ei loeta tavaliselt vastuvõetavaks. Jääb veel üle võimalus tõlgendada skaalat kui ebaühtlast meetrilist skaalat. Kui aga ka vahede võrreldavust eitada, peab koolihinnete skaala ikkagi degradeerima järjeskaalaks, mida õige sageli tehaksegi.

1.6. SKAALATEISENDUSED

Astmeteisendus. Olgu mittenegatiivne suurus x avaldatud meetrilises mõõtskaalas. Arvutame uue suuruse

$$y = x^r , \qquad (15)$$

kus r on suvaliselt valitud arv. Suurused x ja y kirjeldavad sisult üht ja sama tunnust, ainult teine teises mõõtskaalas. Üleminek objekti kirjeldamiselt suuruse x abil tema kirjeldamisele suuruse y abil on tunnuse mõõtskaala astmeteisendus (vt. joon.8).



Joon. 8. Skaala astmeteisendus (koos mastaabiteisendusega). Ühele ja samale objektile vastavad väärtused on joonisel kohakuti.

Monotoonseks teisenduseks nimetatakse suvalist skaalateisendust, mis ei aja segi tunnuse väärtuste järgi korraldatud objektide järjekorda ehk pingerida. Mittenegatiivsete väärtuste piirides on astmeteisendus monotoonne.

Skaala ühtlustamine. Ebaühtlane meetriline skaala degradeerub standardsete andmetöötlusmeetodite puhul ühele tasemele järjeskaalaga. Erinevalt järjeskaalast on aga ebaühtlast meetrilist skaalat võimalik eeltöötluse staadiumis katseandmetele tuginemata parandada, minnes sobiva monotoonse teisenduse abil üle ühtlasele skaalale. Lihtsamatel juhtudel aitab astmeteisendus, mis olenevalt astmenäitajast skaalat ühest otsast tihendab ning teisest venitab (vt. joon. 8).

Kui ühtlustav teisendus on varakult teada, võib soovitada juba andmete kogumisel kasutada uut skaalat.

Klassiteisendust ehk kategoriseerimist kasutatakse pidevalt skaalalt diskreetsele skaalale üleminekuks või diskreetse skaala väärtuste arvu koondamiseks. Vaatlusaluse tunnuse x väärtuste jaoks valitakse klassirajad a_1, a_2, ..., a_{k+1}, kus k on moodustatavate klasside üldarv. Esimesse klassi arvatakse objektid tun-

nuse väärtustega esimesest rajast teiseni, teise klassi väärtustega teisest rajast kolmandani jne. Kui väärtus võrdub alumise rajaga, arvatakse see samasse klassi, kui ülemise rajaga, siis järgmisse klassi. i-ndasse klassi kuulumise tingimus on $a_i \leqslant x < a_{i+1}$. Teisenduse tulemusena asenduvad tunnuse väärtuste originaalmõõtarvud klassinimede või numbritega.

Ühtlase klassiteisenduse korral on klasside laiused võrdsed. Klassi laiust nimetatakse siin sammuks. Klassi nimeks pannakse tunnuse väärtuse originaalmõõtarv klassi keskel c (klassi keskpunkt). Klassifitseerimisskeemi kirjeldatakse kolme arvu abil, need on

1) esimese klassi keskpunkt c_1 ,
2) samm h ,
3) klasside üldarv k .

i-ndasse klassi kuulumise tingimus on

$$c_i - \frac{h}{2} \leqslant x \leqslant c_i + \frac{h}{2} , \qquad (16)$$

kusjuures $c_i = c_1 + (i - 1)\,h$. Esimese klassi alumisest rajast väiksemad ning viimase klassi ülemisest rajast suuremad väärtused langevad niiviisi hoopis vaatluse alt välja (asenduvad küsimärkidega). Öeldakse, et kirjeldatud juhul on äärmised klassid kinnised. Erikokkuleppel võib aga kõigist klassidest väljalangevad väärtused tingimisi arvata äärmistesse klassidesse. Siis öeldakse, et äärmised klassid on lahtised.

Järjed on tunnuse väärtuste järjenumbrid kasvavas pingereas. Kõige väiksema väärtuse järg on üks, järgmisel kaks jne. Suurim järg võrdub väärtuste arvuga n. Järgesid saab määrata järjeskaala või meetrilise skaala puhul.

Kohtade jagamise puhul on kõigi kohta jagavate väärtuste järjeks jagatavate kohtade aritmeetiline keskmine. Kohtade jagamisel ei tule järg alati täisarv, näiteks pingereas alt teist ja kolmandat kohta jagava väärtuse järg on 2,5.

Järjehinded arvutatakse järgede põhjal niiviisi, et tulemus oleks alati täisarv. NAIRII-2 programmides kasutatakse paaritute arvude süsteemi: kui kohtade jagamist pole, pannakse järjehinneteks arvud 1, 3, 5, ..., 2n-1. Kui j on järg, siis järjehinne

$h = 2j - 1.$

Järjeteisendus asendab tunnuse originaalväärtuse järjehindega. Erinevalt astmeteisendusest tehakse järjeteisendus konkreetse väärtuste hulga (valimi, vt. p. 1.7) järgi ning on seetõttu kordusmõõtmiste seerias juhusliku loomuga. Selle kompensatsiooniks kasutab aga järjeteisendus ära katseandmetes peituva informatsiooni originaalmõõtskaala loomuse ja tõenäosusjaotuse kohta. Järjeteisenduse tulemus on invariantne mõõtskaala eelnevate monotoonsete funktsionaalteisenduste suhtes. See tähendab, et eelneva astmeteisenduse või mastaabiteisendusega ei saa järjeteisenduse tulemust muuta. Niiviisi osutub järjeteisenduse tulemus täiesti vabaks mõõtskaala valiku mõjust.

Järjeteisendust kasutatakse nii järjeskaala kui meetrilise skaala korral. Ta on kasulik statistiliste hüpoteeside kontrollimisel ning statistiliste seoste uurimisel juhul, kui tunnuste tõenäosusjaotus on ebatüüpiline või tundmatu.

Lihtsal järjeteisendusel on üks oluline tehnilist laadi puudus. Teisendatud tunnuse tõenäosusjaotus tuleb keskeltläbi ühtlane, standardsed andmetöötlusmeetodid ja arvutusprogrammid on aga enamasti orienteeritud normaaljaotusega lähteandmetele.

Normaalhinded ehk normaalskoorid (inglise keeles "normal scores") arvutatakse järgedest ehk järjehinnetest niisuguse skaalateisenduse abil, mille korral tõenäosusjaotuseks tuleks ühtlase jaotuse asemel normaaljaotus. Blomi lähenduses on normaalhinne

$$g = \Psi\left(\frac{j - 0{,}375}{n + 0{,}25}\right) = \Psi\left(\frac{h + 0{,}25}{2(n + 0{,}25)}\right), \qquad (17)$$

kus Ψ on normaaljaotusfunktsiooni pöördfunktsiooni tähis, j järg ja h järjehinne. Normaalhinnete väärtused on enamasti vahemikus -2 ... 2. NAIRII-2 programmides esitatakse normaalhinded sajaga korrutatult ja täisarvuks ümardatult.

Pedagoogikauurimistöös on originaalväärtusi normaalhinnetega asendav modifitseeritud järjeteisendus üsna tihti skaala valimise probleemi parim lahendus.

1.7. STATISTIKA ALGMÕISTED

Statistika kui teadus tekkis hilisel keskajal ja tähendas algselt riigiõpetust. Riigiuurijate ülesandeks jäi ka riigi olukorra kirjeldamiseks tarvilike andmete kogumine ja analüüsimine. Aegamisi kujuneski sõna "statistika" tähenduseks andmete kogumine ja analüüsimine.

Kaasajal on statistika õige lai mõiste. Üks statistika ülesandeid on endiselt andmete kogumine majanduselu ja rahvastiku kohta, nende andmete süstematiseerimine ja kokkuvõtete tegemine. Seda osa statistikateadusest, mis uurib eelnimetatud tegevuse meetodeid, võiks nimetada klassikaliseks statistikaks. Klassikalist statistikat käsitleb üksikasjalikult õpik [1].

Klassikaline statistika on vanem teadus kui tõenäosusteooria. Statistika näol leidis tõenäosusteooria eest avara tegevusvälja oma meetodite praktiliseks rakendamiseks. Puhas tõenäosusteooria ise aga ei õpeta andmeid töötlema ega praktilisi järeldusi tegema. Juhuslike suuruste katselise uurimise õpetust nimetatakse matemaatiliseks statistikaks. Matemaatiline statistika on tõenäosusteooriast väljakasvanud distsipliin.

Matemaatilise statistika rakendusväljaks on kõik vaatlust ja eksperimenti kasutavad teadused. Psühholoog ja füüsik, lingvist ja bioloog, pedagoog ja insener kasutavad ühtviisi aktiivselt statistilisi meetodeid. Viimasel ajal on matemaatiline statistika üha enam omandanud vaatluse ja eksperimendi üldteooria maine (vt. näiteks [15]).

Objekt ehk indiviid on uurimisalune üksus. Pedagoogikas on objektiks enamasti õpilane, vahel aga ka õppeaine, test, õpiku rubriik, kool jne. Järgnevas kasutatakse kõikjal üldist terminit; kui see lugejat häirib, võib ta sõna objekt olenevalt probleemist asendada konkreetsema terminiga, näiteks "õpilane".

Tunnus on suurus, mille abil kirjeldatakse objekte. Kui objektiks on õpilane, võib tunnuseks olla jõudlushinne, psühholoogilise testi tulemus, kehakaal jne. Tunnus võib olla otseselt mõõdetav ehk primaarne või primaarsete tunnuste järgi arvutamise teel leitav ehk sekundaarne. Kui terminit "tunnus" kasutatakse lisaselgitusteta, siis mõeldakse tavaliselt primaarset tunnust.

Sekundaarseid tunnuseid nimetatakse ka kombineeritud tunnusteks.

Konkreetset objekti iseloomustab tunnuse väärtus. Määramata jäänud väärtust nimetatakse küsimärgiks ehk lüngaks.

Tunnusekomplekt on kõigi nende tunnuste hulk, mida me kasutame konkreetses uurimistöös objektide kirjeldamiseks. Konkreetse objekti tunnuste väärtuste hulka nimetatakse tunnusekomplekti väärtuseks. Erijuhul võib tunnusekomplekt koosneda ühestainsast tunnusest.

Sageli on uurimistöö ühes osas vaatluse all vaid osa tunnuseid. Sel juhul võib üldise tunnusekomplekti vastavat osa vaadelda kui iseseisvat tunnusekomplekti. Tunnusekomplektist osa eraldamist nimetatakse tunnuste valimiseks.

Objekti matemaatiliseks mudeliks on tunnusekomplekti väärtus. Uurimistöö arvutusstaadiumis esindab tunnusekomplekti väärtus objekti täielikult, sisaldades kõike, mis konkreetse objekti kohta teada. Rääkides objektist arvutusstaadiumis, mõtleme selle all tunnusekomplekti väärtust.

Andmetabelis eraldatakse iga objekti jaoks üks rida, iga tunnuse jaoks üks veerg. Objekti nime või numbri järel kirjutatakse tunnuste väärtused (vt. tabel 2).

Tabel 2

Andmetabeli näide

Tunnus / Õpilane	Nr.	Pikkus cm	VTK märk	Matemaatikahinne
		1	2	3
Liiv	1	169	puudub	3
Tamm	2	176	kuld	4
Rebane	3	171	hõbe	5
Pedak	4	172	hõbe	5
Mägi	5	166	puudub	5
Kraav	6	177	hõbe	3

Üldkogum ehk populatsioon on kõigi kirjeldamisele kuuluvate objektide hulk. Üldkogum on määratud uurimisülesande sisuga ja kõik tema objektid ei pruugi olla uurijale kättesaadavad.

Üldkogum võib olla konkreetne ja täpselt piiritletud, umbmääraselt suur või hoopis tõkestamata. Näiteks eesti õppekeelega koolides möödunud õppeaastal kolmanda klassi lõputunnistuse saanud õpilaste hulk on konkreetne. Kui aga õppeaasta pole fikseeritud, jääb hulk umbmääraseks. Kõikvõimalike kirjanditeemade hulk on tõkestamata.

Konkreetset objektide hulka kirjeldab lõplik üldkogum, tõkestamata hulka lõpmatu üldkogum. Uurimistöö arvutusstaadiumis käsitletakse üldkogumit kui reaalsetest objektidest koosneva reaalse hulga abstraktset kujutist ehk mudelit. Loomult range matemaatiline mudel pole kunagi täiuslikult adekvaatne umbmääraselе tegelikkusele (mõelgem näiteks kõigi õhtukooliõpilaste tegelikkusele vastava nimistu koostamisele) ning seda võib teatud piirides üsna suvaliselt modifitseerida. Üldkogumit kui abstraktset mudelit võib näiteks laiendada kujutletavate objektidega. Kujutletavate objektide arv on reeglina tõkestamata ning laiendamise tulemus lõpmatu üldkogum.

Lõpmatu üldkogum lubab end kirjeldada kui mitmemõõtmelist (erijuhul ühemõõtmelist) juhuslikku suurust, mille väärtusteks on tunnusekomplekti väärtused. Vaatlusaluse objekti valimist ehk objekti sattumist vaatluse alla käsitletakse siin nagu juhuslikku sündmust.

Pedagoogikauurimistöös kirjeldatavate objektide hulk on enamasti umbmääraselt suur ning kujutletavatest objektidest koosnev lõpmatu üldkogum sobivaim matemaatiline mudel. Ka esimesel pilgul selgelt piiritletud kogumi uurimisel võib sisulise huvi objektiks osutuda hoopiski kujutletav lõpmatu üldkogum. Olgu näiteks haridussüsteemis üksainus z-kallakuga eriklass ning tarvis on uurida z-kallaku mõju õpilaste arengule. Uurimisülesandel on mõte ilmselt vaid siis, kui nähakse ette võimalust z-kallakuga klassi korduvaks komplekteerimiseks või mitme z-kallakuga klassi asutamiseks. Tarvis on teada z-kallaku mõju mitte vaatlusaluses klassis, kus vaadeldud mõju on juba realiseerunud ja teha pole enam midagi, vaid kujutletavates uutes z-kallakuga klassides.

Sageli on tarvis kirjeldada mõju ja tulemuse vahelist seost.

Huviobjektiks on seos mingil viisil piiritletud omadustega kollektiivis üldse, aga mitte ühes konkreetses kollektiivis.Ka siin on mõistlikuks matemaatiliseks mudeliks lõpmatu üldkogum.

Valim on üldkogumi vaatlusalune osa ehk vaatlusaluste objektide kogum. Vaatlusaluste objektide arv n on valimi maht. Et iga objekt on arvutusstaadiumis esindatud tunnusekomplekti väärtusega, siis valim on esindatud tunnusekomplekti väärtuste hulga ehk andmetabeliga. Esialgu (kuni seoste käsitlemiseni) vaatleme ainult üksiktunnuseid ja ühe tunnuse valimeid. Valimisse sattunud väärtusi x_1, x_2, ..., x_n nimetatakse valimi komponentideks ja neid käsitletakse nagu juhusliku suuruse konkreetseid väärtusi n järjestikuse katse korral.

Statistikaterminoloogia pole eesti keeles veel stabiliseerunud. Paralleelselt terminiga "valim" on tarvitatud ja tarvitatakse samas tähenduses termineid "väljavõte", "väljavõtukogum", "võend".

Statistikud on valimi komponentide järgi arvutatavad suurused, näiteks aritmeetiline keskmine $\bar{x}$, ekstreemsed väärtused min ja max, valimi haare w = max - min jne. Statistikute moodustamise võimalusi on lõpmata palju. Et valimi komponendid on juhuslikud, siis ka statistikud on juhuslikud suurused igaüks oma jaotuse, keskväärtuse ja standardhälbega. Statistikute juhuslikkus avaldub valimi kordamisel ja nende jaotust saab otseselt uurida kordusvalimite meetodil. Matemaatilise statistika teooria vabastab meid sellest tülikast ülesandest ja õpetab statistikute jaotusparameetreid hindama uurimisaluse valimi enese järgi.

Esindav ehk representatiivne on valim siis, kui valimi moodustamise eeskiri tagab võimaluse hinnata valimi järgi üldkogumi parameetreid ilma süstemaatilise veata. Põhiline esindava valimi moodustamise meetod on juhuslik komplekteerimine. Juhuslikkuse tagamiseks võib kasutada mitmesuguseid loosimismeetodeid, kaasa arvatud loosimine juhuslike arvude tabeli abil, nimede alfabeetilise struktuuri alusel jne. Kui üldkogumi igal elemendil on ühesugune võimalus valimisse sattuda, on esindavus tagatud.Erandiks on kujutletavad objektid, sest neil pole võimalust reaalselt valimisse sattuda. Siin ei sobitata esindavuse huvides aga mitte valimit üldkogumiga, vaid üldkogumit valimiga: kujutleta-

vaid objekte kujutletakse niisugustena, et nende tunnused oleksid sama tõenäosusjaotusega, nagu on reaalsetel objektidel. See postulaat peab olema kooskõlas uurimisülesande sisuga.

Juhul kui üldkogum jaguneb tuntud proportsioonides osakogumiteks, on kasulik ka valim koostada osavalimitest. Üldvalimi esindavuse tingimusteks on siis osavalimite esindavused osakogumite jaoks ning valimi jaotuse proportsioonide identsus üldkogumi jaotuse proportsioonidega.

Lõplik üldkogum ei ole modelleeritav juhusliku suurusega ja teda kirjeldatakse teistes terminites kui lõpmatut üldkogumit. Kui aga üldkogum on valimist palju suurem, siis jäävad kõik olulised erinevused tulemuste tõlgendamise staadiumi. Arvutused tehakse niisamuti nagu lõpmatu üldkogumi puhul. Niiviisi osutuvad lõpmatu üldkogumi uurimise eeskirjad ning arvutusprogrammid teatud utilitaarses mõttes universaalseteks.

Juhul kui üldkogumi maht on vaid mõned korrad suurem valimi mahust, põhjustab lõpmatu üldkogumi uurimisele orienteeritud arvutuseeskirjade kasutamine vaid väikest informatsioonikadu (usaldusvahemikud tulevad veidi laiemad kui tarvis). Ka äärmusjuhul, kui mahud on võrdsed (vaatluse all on kogu üldkogum), saab lõpmatu üldkogumi analüüsi programmide abil kõik tarvilikud tulemused, osa tulemusi (näiteks usalduspiirid) osutub siis vaid liigseiks.

Järgnevas lõpliku üldkogumi uurimise probleemidele erilist tähelepanu ei pöörata. Vajaduse korral võib lugeja pöörduda õpiku [1, pt. 9] poole.

1.8. PUNKTHINNANGUD

Punkthinnang on juhusliku suuruse ehk üldkogumi parameetri jaoks valimi järgi arvutatud lähendsuurus. Et kõiki valimi järgi arvutatud suurusi nimetatakse statistikuteks, siis on punktihinnanguks alati mingi statistik. Ühe ja sama parameetri hinnangutena võib kasutada erinevaid statistikuid. Näiteks keskväärtuse hinnanguks võib kasutada valimi üldist aritmeetilist keskmist, ekstreemsete väärtuste keskmist (min+max)/2 või muud statistikut. Matemaatilise statistika teooria peab selgitama, milline statis-

tik on hinnanguks kõige sobivam.

Hindamisviga on hinnangu ja hinnatava parameetri tõelise väärtuse vahe. Praktikas jääb hindamisvea konkreetne väärtus alati tundmatuks, vastasel korral peaks parameeter ette teada olema ja hindamine ise tarbetu. Et statistikud on juhuslikud suurused ja omandavad valimi kordamisel alati uue väärtuse, pole hindamisviga võimalik vältida. Ka hindamisviga ise on juhuslik suurus.

Hindamisvea keskväärtust nimetatakse hinnangu nihkeks, see on hinnangu süstemaatiline viga.

Nihutamata hinnangu nihe on null ja hinnang süstemaatilisest veast vaba. Reeglina leiavad praktilist kasutamist ainult nihutamata või tähtsusetult väikese nihkega hinnangud. Nihutamata hinnangu täpsust kirjeldab hinnangu standardhälve kordusvalimite hulgas.

Efektiivse hinnangu standardhälve on väikseim kõikvõimalike nihutamata hinnangute standardhälvete hulgas. Teiste sõnadega: efektiivne hinnang on kõige täpsem hinnang.

Tõenäosuse p efektiivseks hinnanguks on sündmuse suhteline sagedus (vt. p. 1.1). Hinnangu standardhälve on $\sqrt{p(1-p)/k}$, kus k on katsete üldarv.

Keskväärtuse efektiivse hindamise reegel oleneb uuritava juhusliku suuruse jaotuse iseloomust. Normaaljaotuse korral on efektiivseks hinnanguks valimi aritmeetiline keskmine $\bar{x}$. Aritmeetilise keskmise standardhälve on $\sigma/\sqrt{n}$ (vt. p. 1.4). Ühtlase jaotuse korral on efektiivseks hinnanguks ekstreemsete väärtuste keskmine (min+max)/2. Suure ekstsessikordajaga jaotuste korral tuleb aga, vastupidiselt ühtlasele jaotusele, arvestada eelistatult valimi komponentide "pingerea" keskel seisvaid väärtusi. Praktikas kasutatakse keskväärtuse hinnanguna enamasti jaotusele tähelepanu pööramata aritmeetilist keskmist.

Dispersiooni loomulikuks hinnanguks oleks valimi dispersioon $\sum_{i=1}^{n} (x_i - \mu)^2/n$. Et aga keskväärtus on tundmatu, pole esitatud valemit võimalik kasutada. Keskväärtuse asendamine aritmeetilise keskmisega vähendab hinnangu väärtust, sest aritmeeti-

line keskmine paikneb täpsemini sama valimi tsentris kui keskväärtus. Nihet õnnestub täpselt kompenseerida, vähendades nimetajat ühe võrra. Niiviisi saadakse statistik

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1} , \qquad (18)$$

mis on normaaljaotuse puhul dispersiooni parim nihutamata hinnang.

Standardhälbe loomulikuks hinnanguks on ruutjuur dispersiooni hinnangust

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}} . \qquad (19)$$

Normaaljaotuse korral erineb see hinnang parimast võimalikust hinnangust õige vähe, ja olgugi ta väikese mahuga valimi korral veidi nihutatud, leiab üldist kasutamist.

Aritmeetilise keskmise standardhälbe hinnangu saamiseks jagatakse statistik s ruutjuurega valimi mahust. Tulemus

$$s_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n(n - 1)}} . \qquad (20)$$

Andmete mehhaniseerimata töötlemisel võib väikese mahuga valimi (n = 3...10) aritmeetilise keskmise standardhälbe hindamisel kasulikuks osutuda hästi lihtsalt arvutatav statistik

$$s_{\bar{x}}^{w} = \frac{w}{n} , \qquad (21)$$

kus w on valimi haare. Hinnang (21) on kõlblik ainult normaaljaotuse korral, hinnangut (20) võib kasutada ka normaaljaotusest erineva jaotuse puhul. Normaaljaotuse korral on hinnangu (21) väiksem täpsus kompenseeritav valimi mahu suurendamisega ühe võrra.

1.9. TÄPSUS JA INFORMATSIOON

Mõõtmisviga on mõõtmistulemuse ja mõõdetava suuruse tõelise väärtuse vahe. Käesolevas punktis eeldame, et mõõdetava suuruse tõeliste väärtuste skaala on pidev meetriline skaala, mispuhul mõõtmisvead on vältimatud.

Mida väiksem on mõõtmisviga, seda parem mõõtmise täpsus. Täpsust kirjeldataksegi mõõtmisvea kaudu.

Vaatleme ühe ja sama püsiva tõelise väärtusega suuruse korduvat mõõtmist. Mõõtmistulemus osutub kordusmõõtmiste hulgas juhuslikuks suuruseks. Mõõtmisvea keskväärtust nimetatakse süstemaatiliseks veaks ja hälvet juhuslikuks veaks. Süstemaatilise vea põhjuseks on mõõtmismeetodi puudulikkus. Järgnevas eeldame, et mõõtmismeetod on korrektne ja süstemaatilist viga ei ole.

Kui mõõtmisvea konkreetne väärtus oleks teada, saaks mõõtmistulemusest mõõtmisviga lahutades kätte mõõdetava suuruse absoluutselt täpse väärtuse. Tegelikult on aga teada vaid mõõtmistulemus ja mõõtmisvea konkreetne väärtus jääb tundmatuks. Juhusliku mõõtmisvea reaalselt hindamisele alluv mõõt on mõõtmisvea standardhälve.

Aprioorne informatsioon tähendab neid teadmisi mõõdetava suuruse väärtuse kohta, mis on olemas enne mõõtmist. Vaatleme siinkohal mingi püsiva suuruse, näiteks ühe konkreetse objekti kindla tunnuse või objektide hulga jaoks tunnuse keskväärtuse mõõtmist. Aprioorse informatsiooni allikaks võivad olla analoogiliste objektide varasema uurimise või sama objekti jaoks sooritatud eelmõõtmise tulemused. Me nimetame mõõtmiseks ka väärtuse hindamist "silma järgi", olgugi et niisuguse mõõtmise vea standardhälve on reeglina suur ning täpsus väike.

Aprioorset informatsiooni esitab mõõtmisele kuuluva väärtuse eelhinnang x_0 ja eelhinnangu vea standardhälve σ_0. Informatsiooni on seda rohkem, mida täpsem on eelhinnang ehk mida väiksem standardhälve σ_0.

Kui aprioorset informatsiooni üldse poleks, ei saaks ükski mõõtmistulemuse pähe pakutud arv uurijat üllatada. Kõik oleks siis ühevõrra usutav. Nii see aga tavaliselt pole, mis annab tunnistust aprioorse informatsiooni olemasolust. Tuleb ette sedagi, et mõõtmine pole üldse eelhinnangust täpsem ja aprioorse

informatsiooni väärtus suurem kui mõõtmistulemusel. Näiteks ühe ja sama suuruse üheksakordsel mõõtmisel on kaheksa esimese mõõtmise tulemused üheksanda mõõtmise eel viimase suhtes aprioorne informatsioon, mis määrab mõõdetava suuruse täpsemini, kui eelseisev üheksas mõõtmine seda teha võib.

Informatsioonide ühendamine tähendab mõõdetava suuruse väärtuse jaoks niisuguse hinnangu leidmist, mis kasutab nii aprioorset informatsiooni kui viimast mõõtmistulemust. Lähteandmeteks on eelhinnang x_0 ja selle vea standardhälve σ_0 ning mõõtarv x_m ja mõõtmisvea standardhälve σ_m. Normaaljaotuse korral osutub parimaks ühendatud hinnanguks eelhinnangu ja mõõtarvu kaalutud keskmine

$$x = \frac{\frac{1}{\sigma_0^2} x_0 + \frac{1}{\sigma_m^2} x_m}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma_m^2}} = \frac{\sigma_m^2 x_0 + \sigma_0^2 x_m}{\sigma_m^2 + \sigma_0^2} \qquad (22)$$

Selle hinnangu vea standardhälve on

$$\sigma = \sqrt{\frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma_m^2}}} = \frac{\sigma_m \sigma_0}{\sqrt{\sigma_m^2 + \sigma_0^2}} \; . \qquad (23)$$

Kui informatsioonide ühendamise võimalus jäetakse kasutamata, siis tähendab see informatsiooni raiskamist. Pedagoogikas on indiviidi uurimisel pea alati ette teada mõõdetava tunnuse keskväärtus ja hajuvus kollektiivis, kuhu indiviid kuulub. Kui muid eelteadmisi polegi, annab kollektiivi kuulumise fakt ise mõõtmistulemuse eelhinnanguks kollektiivi keskmise ja standardhälbe.

Kaalutud keskmise üldvalem on

$$\bar{x} = \frac{p_1 x_1 + p_2 x_2 + \ldots + p_n x_n}{p_1 + p_2 + \ldots p_n} \; . \qquad (24)$$

$x_1, x_2, \ldots, x_n$ on siin keskmistatavad suurused ja $p_1, p_2, \ldots, p_n$ - kaalud. Erijuhul $p_1 = p_2 = \ldots = p_n = 1$ osutub kaalutud keskmine lihtsaks aritmeetiliseks keskmiseks. Mõne kaalu suuren-

damine tähendab vaadeldava väärtuse osatähtsuse tõstmist kaalutud keskmise arvutamisel.

Valem (22) on valemi (24) erijuht, kus kaaludeks on täpsust kirjeldavad suurused $1/\sigma_o^2$ ja $1/\sigma_m^2$.

Mõõteinformatsiooni hulga mõõduks on eelhinnangu ja lõpptulemuse vigade standardhälvete suhe

$$\frac{\sigma_o}{\sigma} = \sqrt{1 + \frac{\sigma_o^2}{\sigma_m^2}} \quad . \tag{25}$$

Informatsiooni hulk on suhteline suurus. Kui eelhinnangu täpsus pole teada, ei saa informatsiooni hulka määrata.

Matemaatiline informatsiooniteooria on iseäralik teooria sellepoolest, et informatsiooni enese kui absoluutse suuruse mõistet siin üldse pole. On vaid mõõtmise või teate vastuvõtmise protsessis saadud informatsiooni hulga mõiste. Matemaatilises teoorias on informatsiooni hulga mõõduks tavaliselt ülalnimetatud standardhälvete suhte logaritm.

Juhusliku suuruse mõõtmisel on mõõdetava suuruse tõeliseks väärtuseks juhusliku suuruse konkreetne väärtus. Seni eeldasime, et üht ja sama konkreetset väärtust on võimalik palju kordi mõõta ja niiviisi mõõtmisviga isoleeritult uurida. Praktikas aga saab juhuslik suurus sageli igal kordusmõõtmisel uue konkreetse väärtuse. Juhusliku suuruse hälve ja mõõtmisviga liituvad ning kumbagi pole võimalik summast eraldada. Mõõtmistäpsuse hindamiseks on tarvis korraldada lisamõõtmisi eritingimustes, kus mõõdetava suuruse tõeline väärtus on konstantne, mõõtmisvigu põhjustavad tegurid aga samad mis põhimõõtmiste korral. Kui see pole võimalik, peab loobuma püüdest kirjeldada huvipakkuva tunnuse hajuvust ning seoseid "puhtal" kujul ja käsitama tulemusi kui mõõtmisvigu sisaldava "ebapuhta" tunnuse väärtusi. Juhusliku suuruse keskväärtuse hinnangu keskväärtust juhuslikud mõõtmisvead ei mõjuta, nad vähendavad vaid keskväärtuse määramise täpsust.

Standardhälbe korrigeerimise võtet saab kasutada siis, kui mõõtmisvea standardhälve σ_m on tuntud. Kui uuritava juhusliku suuruse enese standardhälve on σ_o ja mõõtmisviga ei sõltu uuritava suuruse konkreetsest väärtusest, siis mõõtmistulemuste

standardhälve tuleb $\sigma = \sqrt{\sigma_o^2 + \sigma_m^2}$. Juhusliku suuruse standardhälve avaldub siit järgmiselt:

$$\sigma_o = \sqrt{\sigma^2 - \sigma_m^2} \quad . \tag{26}$$

Keskväärtuse hinnangu täpsust peab aga hindama ikka korrigeerimata standardhälbest σ lähtudes, sest mõõtmisvead mõjutavad aritmeetilist keskmist võrdväärselt juhusliku suuruse enese hälvetega.

1.10. STATISTILISED HÜPOTEESID

Statistiline hüpotees on uurimistöös püstitatud sisulise hüpoteesi formaliseerimise tulemus. Formaliseeritud hüpotees sõnastatakse tõenäosusteooria ja matemaatilise statistika keeles ning see väidab midagi tõenäosusjaotuse või jaotuste suhtes. Näiteks sisulise hüpoteesi "1^a klassi parem õppeedukus 1^b klassiga võrreldes pole seletatav ainult juhusest tingitud erinevustega klasside komplekteerimisel" formaliseerimise tulemuseks võib olla statistiline hüpotees "juhusliku suuruse a keskväärtus on suurem kui juhusliku suuruse b keskväärtus".

Nullhüpotees ja alternatiiv on kaks teineteist välistavat statistilist hüpoteesi. Kumba kahest nimetada nullhüpoteesiks, kumba alternatiiviks, on kokkuleppe küsimus.

Hüpoteesi kontrollimiseks mõõdetakse hüpoteesis nimetatud juhuslike suuruste väärtusi. Matemaatilise statistika teoreetiliste meetodite abil koostatakse mõõtmistulemuste kohta käiv tingimus, mis oleks tõese nullhüpoteesi korral suure tõenäosusega täidetud ning tõese alternatiivi korral suure tõenäosusega täitmata. Kui tegelikud mõõtmistulemused rahuldavad seda tingimust, siis eelistatakse nullhüpoteesi, kui ei rahulda, siis alternatiivi.

Statistilisteks testideks või kriteeriumideks nimetatakse hüpoteeside kontrollimise eeskirju.

Hüpoteesi kummutamine on tinglik väljend, mis tähendab, et katseandmete ja hüpoteesi kooskõlalisus kuulutatakse väheusuta-

vaks. Statistikameetodid ei võimalda hüpoteesi ekslikkust kunagi absoluutse kindlusega tõestada. Alati jääb ka võimalus,et ekslikult kummutatakse tõene hüpotees, sest juhuse tõttu on andmed erakordselt hälbinud.

Hüpoteesi vastuvõtmine on tinglik väljend, mis tähendab hüpoteesi ainuvõimaliku alternatiivi kummutamist. Katseandmete ja hüpoteesi kooskõla usutavaks tunnistamine pole piisav alus hüpoteesi vastuvõtmiseks. Vaatleme näiteks hüpoteesi "juhusliku suuruse x keskväärtus on täpselt 671" juhul, kui keskväärtus on tõeliselt 670,999 ja standardhälve 100. Ükski reaalne katse ei suudaks nendes tingimustes hüpoteesi kummutada, mis aga ei anna veel alust hüpoteesi vastu võtta. Kontrolli tulemuseks võib olla veidi erinev väide "juhusliku suuruse x keskväärtuse erinevus arvust 671 on usutavasti väiksem kui 3" ehk "keskväärtuse võimalik erinevus arvust 671 pole uuritava probleemi seisukohalt oluline". Teise näitena vaatleme hüpoteesi "suuruste a ja b keskväärtused on erinevad". Selle hüpoteesi ainuvõimalik alternatiiv on "suuruste a ja b keskväärtused on võrdsed". Kui võrdsuse hüpoteesi õnnestub kummutada, tähendab see keskväärtuste erinevuse hüpoteesi vastuvõtmist.

Esimest liiki veaks nimetatakse tõese hüpoteesi eksikombel kummutamist.

Olulisuse nivoo on esimest liiki vea suurim lubatud tõenäosus. Olgu hüpotees "suuruste a ja b keskväärtused on võrdsed" tõene. Valides olulisuse nivooks 1 % , riskeerime ühel protsendil hüpoteesi korduva kontrollimise juhtudest "avastada" a ja b keskväärtuste olulise erinevuse eeltoodud hüpoteesi eksliku kummutamise teel. Olulisuse nivoo valitakse sõltuvalt esimest liiki vea ohtlikkusest matemaatikaväliste (majanduslike, eetiliste jne.) kaalutluste alusel. Kui eelnimetatud hüpoteesi kontrollimise tulemusest oleneb vaid see, kas erinevuse otsimisest loobutakse või jätkatakse seda täiendavate katsete sooritamise teel, piisab tavaliselt 5-10 % olulisuse nivoost. Teaduslike lõppjärelduste kontrollimisel ning praktika jaoks soovituste koostamisel on aga harva alust valida olulisuse nivood üle ühe protsendi. Olulisuse nivoo tähendab siin väärjärelduse avaldamise ning praktikasse rakendamise lubatud tõenäosust.

Usaldusnivoo on esimest liiki vea vältimise ehk sisulise uurimistöö seisukohalt väärjärelduste vältimise nõutud tõenäosus. Usaldusnivoo võrdub 100 % miinus olulisuse nivoo. Üheprotsendilise olulisuse nivoo korral on usaldusnivoo 99 %. Kumba nivood uurimistöös nimetada, on vaid kokkuleppe küsimus. Matemaatilise statistika teoorias eelistatakse traditsiooniliselt olulisuse nivood, andmetöötluse tulemuste esitamisel aga sageli usaldusnivood.

Usaldatavus, usaldusprotsent ehk suurim usaldusnivoo tähendab suurimat nendest usaldusnivoodest, mille juures konkreetsed andmed lubavad hüpoteesi kummutada. Usaldatavust võib tõlgendada ka kui eksituse puudumise tõenäosust hüpoteesi kummutamisel konkreetsete andmete alusel.

Lihtsa kalkulaatori ja statistikatabelitega varustatud arvutaja valib tavaliselt usaldusnivoo üheselt ette (näiteks 99 %) ning saab arvutamise tulemuseks vastuse ei või jah. Elektronarvutil on aga otstarbekam lasta leida hüpoteesi kummutamise usaldatavus (mis võib tulla näiteks 98,5 %). Otsuse tegemine on hiljem lihtne: kui usaldatavus pole nõutud usaldusnivoost väiksem, loetakse hüpotees kummutatuks, kui ta aga on väiksem, siis mitte.

NAIRII-2 statistikaprogrammid esitavad hüpoteeside kontrollimisel lõpptulemusena hüpoteesi kummutamise usaldatavuse väärtuse.

Teist liiki veaks nimetatakse väära hüpoteesi kummutamata jätmist. Sisuliselt tähendab see mingi efekti avastamata jäämist.

Testi võimsuseks nimetatakse teist liiki vea puudumise tõenäosust. Tähenduse järgi võiks võimsust nimetada ka testi tundlikkuseks. Võimsus oleneb usaldusnivoost: mida suurem usaldusnivoo, seda väiksem võimsus. Seepärast ei võigi valida liig kõrget usaldusnivood. Fikseeritud usaldusnivoo juures oleneb võimsus testi konkreetsest eeskirjast. Matemaatilise statistika teooria abil koostatakse statistilisi teste niiviisi, et nad eeldustekohastes tingimustes ei jääks võimsuse poolest alla ühelegi teisele mõeldavale testile.

1.11. TÕENÄOSUSTE VÕRDLEMINE

Kahe sagedustabeli erinevus. Jagunegu N objektist koosnev valim A mingi tunnuse järgi k-sse rühma, esimeses rühmas n_1 objekti, teises n_2 objekti jne. Arvud n_1, n_2, ..., n_k moodustavad sagedustabeli, mille graafiliseks kujutiseks on histogramm. Summa $n_1 + n_2 + \ldots + n_k = N$. Jagunegu M objektist koosnev teine valim B analoogilisel viisil k-sse rühma, i-ndas rühmas m_i objekti. i-ndasse rühma sattumise tõenäosuse hinnang valimi A järgi on suhteline sagedus n_i/N, valimi B järgi m_i/M. Sagedustabelite suhtelise erinevuse sobivaimaks mõõduks osutub statistik

$$\chi^2 = NM \sum_{i=1}^{k} \frac{(\frac{n_i}{N} - \frac{m_i}{M})^2}{n_i + m_i} . \qquad (27)$$

Kui kõik suhtelised sagedused on täpselt võrdsed, tuleb $\chi^2 = 0$.

Sagedustabeli erinevus teoreetilisest jaotusest. Teoreetilist jaotust võib kirjeldada rühmadesse sattumise tõenäosuste p_i või tõenäosuste M-kordsete $m_i = p_i M$ abil. Viimane kirjeldusviis on vormilt sarnane sagedustabeliga, selline "teoreetiline" sagedustabel pole aga mõjutatud valimi moodustamisel esinevast juhuslikkusest. Ka tabeli maht M on siin vaid tinglik (võib valida $M = 1$, siis on $m_i = p_i$). Kui arvud n_i esitavad empiirilist ja arvud m_i teoreetilist sagedustabelit, arvutatakse tabelite erinevuse mõõt järgmiselt:

$$\chi^2 = NM \sum_{i=1}^{k} \frac{(\frac{n_i}{N} - \frac{m_i}{M})^2}{m_i} . \qquad (28)$$

Kooskõlahüpotees väidab, et valimid A ja B esindavad üht ja sama juhuslikku suurust või ühtviisi jaotatud juhuslikke suurusi ning statistiku χ^2 väärtus võib erineda nullist vaid valimite komplekteerimise juhuslikkuse tõttu.

χ^2-test on üks universaalsemaid matemaatilise statistika meetodeid, mille tuntuim rakendus on kahe sagedustabeli võrdlemine. Kui kooskõlahüpotees on tõene, siis allub statistik χ^2 k-1

vabadusastmega χ^2-jaotusele, mille jaoks on olemas nii tabelid statistikatabelite kogumikus kui alamprogramm NAIRII-2 statistikaprogrammide süsteemis. Nende abil saab konkreetse ülesande korral leida kooskõlahüpoteesi kummutamise usaldatavuse. Nagu arvata võib, tuleb usaldatavus seda suurem, mida suurem on χ^2 väärtus. Vabadusastmete arvu k-1 võib käsitada kui tingarvu, mida on tarvis tabelist vastuse otsimisel.

Rangelt võttes allub valemi (27) või (28) järgi arvutatud statistik χ^2-jaotusele vaid ligikaudselt, lähenduse täpsus on aga praktika jaoks piisavalt hea.

Näide. Olgu ülesandeks võrrelda hinnete jaotust kahes värskelt komplekteeritud klassis (vt. tabel 3)

Tabel 3

Hinnete võrdlus (näide)

Hinne	Õpilasi klassis	
	9^a	9^b
2	3	9
3	10	4
4	12	6
5	8	14

eesmärgiga kontrollida, kas klassid on komplekteeritud võrdsetel alustel ja juhuslikult. Valemi (27) järgi arvutame $\chi^2 = 9,2$. Statistikatabelitest leiame kolme vabadusastme (k=4) korral χ^2 kriitiliseks väärtuseks 95 % usaldusnivool 7,8 ja 99 % usaldusnivool 11,3. Siit järeldub, et 95 % usaldusnivool võib klasside ühtlase komplekteerimise hüpoteesi kummutada, 99 % usaldusnivool aga ei või. Arvutit kasutades (programm TÕENÄOSUSJAOTUSTE VÕRDLUS) on tarvis vaid esitada neli arvupaari n_i, m_i, vastuseks saame kooskõlahüpoteesi kummutamise usaldatavuse väärtuse 97,3%. Väites, et klassid on komplekteeritud erineval alusel, riskeerime 2,7 % tõenäosusega valetada.

Näide. Jagunegu kõik õpilased z-tunnuse järgi kolme tüüpi ja olgu tüüpide sageduse kohta teada üldine norm. Vaatleme õpi-

laste jaotust konkreetses koolis (tabel 4)

Tabel 4

Õpilaste jaotus (näide)

Tüüp	Õpilasi	Norm
A	141	15 %
B	421	50 %
C	286	35 %

ja küsime, kas see jaotus on kooskõlas normiga. Statistik χ^2 arvutatakse siin valemi (28) järgi, võttes m_i-deks arvud 15, 50 ja 35. Arvuti abil saame aga tulemuse hõlpsamalt. Kooskõlahüpoteesi kummutamise usaldatavus tuleb vaadeldud arvude puhul 62 %, mis tähendab, et normist erinevuse väitmiseks pole alust.

Ühtlase jaotuse hüpoteesi kontrollimiseks võrreldakse sagedustabelit χ^2-testi abil "teoreetilise" sagedustabeliga, mille kõigis lahtrites on võrdsed arvud.

Normaaljaotushüpoteesi on samuti võimalik kontrollida χ^2-testi abil, enamasti on aga otstarbekam kontrollida asümmeetriakordaja ja ekstsessikordaja väärtusi. Kui asümmeetriakordaja absoluutväärtus ületab tabelis 5 näidatud väärtuse või ekstsessikordaja ei mahu tabelis 6 näidatud rajade vahele, siis võib normaaljaotushüpoteesi kummutada tabeli päises näidatud usaldusnivool.

Juhusliku sündmuse tõenäosuse hindamine on samaväärne kaht võimalikku väärtust (sündmus toimus - sündmus ei toimunud) omava juhusliku suuruse tõenäosusjaotuse uurimisega. Kui sündmus toimus n juhul N võimalikust, siis vastandsündmus toimus N-n juhul. Sagedustabeli sümboolikas on $k = 2$, $n_1 = n$ ja $n_2 = N - n$.

Kahe juhusliku sündmuse tõenäosuste võrdlemiseks võib arvutada statistiku χ^2 valemi (27) järgi. Eelnimetatud asendust kasutades saame valemit lihtsustada:

$$\chi^2 = \frac{(N + M)\,(nM - mN)^2}{NM(n+m)\,(N+M-n-m)} . \qquad (29)$$

Tabel 5

Asümmeetriakordaja kriitilised väärtused

Valimi maht	Usaldusnivoo 90 %	98 %
25	0,711	1,061
30	0,661	0,982
35	0,621	0,921
40	0,587	0,869
45	0,558	0,825
50	0,533	0,787
60	0,492	0,723
70	0,459	0,673
80	0,432	0,631
100	0,389	0,567
150	0,321	0,464
200	0,280	0,403
250	0,251	0,360
300	0,230	0,329
400	0,200	0,285
500	0,179	0,255
750	0,146	0,208
1000	0,127	0,180
2000	0,090	0,127
3000	0,073	0,104

Tabel 6

Ekstsessikordaja kriitilised rajad

Valimi maht	Usaldusnivoo 90 %	98 %
50	2,13 4,01	1,95 4,92
100	2,35 3,77	2,18 4,40
150	2,45 3,66	2,30 4,14
200	2,51 3,57	2,37 3,98
250	2,55 3,51	2,42 3,87
300	2,59 3,47	2,46 3,79
500	2,67 3,37	2,57 3,60
1000	2,76 3,26	2,68 3,41
2000	2,83 3,18	2,77 3,28
3000	2,86 3,15	2,81 3,22

Siin n ja m on sündmuste esinemiskordade ja N ja M võimaluste arvud.

Vaatlusalusel erijuhul on statistiku tegeliku jaotuse kõrvalekalle teoreetilisest χ^2-jaotusest kõige suurem. Viga aga annab hõlpsalt parandada, sest veidi muudetud statistiku jaotus

$$\chi^2_1 = \frac{(N+M-1)\,(nM-mN)^2}{NM(n+m)\,(N+M-n-m)} \qquad (30)$$

on χ^2-jaotusele õige lähedane. Seepärast kasutataksegi tõenäosuste ehk protsentide praktilisel võrdlemisel statistikut χ^2_1. Otsus tehakse täpselt sama skeemi järgi kui tõenäosusjaotuste võrdlemisel, ainult kooskõlahüpoteesi on siin sobivam nimetada võrdtõenäosushüpoteesiks. Praktiliste ülesannete lahendamisel kasutatakse programmi TÕENÄOSUSTE VÕRDLUS.

Näide. Olgu A-kooli 67-st kümnenda klassi õpilasest 39 (58%) tüdrukud ning B-kooli 73-st kümnenda klassi õpilasest 35 (48%) tüdrukud. Kas võib öelda, et tütarlapsed on koondunud eelistatult ühte vaatlusalustest koolidest? Arvutustulemuseks on $\chi^2_1 = 1{,}5$ ning võrdtõenäosushüpoteesi kummutamise usaldatavus 77 %. Seega on protsentide erinevus seletatav juhusega ning ülaltoodud küsimusele peab vastama eitavalt.

Juhusliku sündmuse tõenäosuse võrdlemisel antud tõenäosusega (normiga) peab lähtuma valemist (28). Kui sündmus toimus n juhul N võimalikust ning normatiivne tõenäosus on p, siis

$$\chi^2 = \frac{(n - pN)^2}{p(1 - p)N} . \qquad (31)$$

Näide. Olgu õppeasutuse 834 õpilasest 456 tüdrukud, tütarlaste protsent mõeldavate õpilaskandidaatide hulgas aga 49. Et $\chi^2 = 10{,}8$ ja kooskõlahüpoteesi kummutamise usaldatavus 99,9 %, võib üsna veendunult väita, et tüdrukute ülekaal õppeasutuses pole ainult juhus.

1.12. KESKMISTE VÕRDLEMINE

Võrdlushüpoteesid. Tähistame sümbolitega ξ ja η kahe juhusliku suuruse jaotuste tsentreid ehk lühemalt öeldes keskmisi. Vaatleme nelja hüpoteesi ja anname neile nimed järgmiselt:

võrdsushüpotees $\xi = \eta$,
erinevushüpotees $\xi \neq \eta$,
ületamishüpotees $\xi > \eta$,
allajäämishüpotees $\xi < \eta$.

Erinevushüpotees on võrdsushüpoteesi vastand ning ületamishüpo-

teesi ja allajäämishüpoteesi ühendus. Allajäämishüpotees ei vaja iseseisvat käsitlemist, sest suuruste nimede vahetamise teel saab teda alati muuta ületamishüpoteesiks.

Otseselt on võimalik kontrollida vaid hüpoteesi, mis määrab üheselt mingi võrreldavate suuruste erinevust kirjeldava statistiku jaotuse. Vaadeldavate hüpoteeside seas on niisugune vaid võrdsushüpotees, seepärast nimetatakse keskmiste võrdlemise ülesannetes nullhüpoteesiks alati võrdsushüpoteesi. Sisulistes uurimisprobleemides aga ei paku võrdsushüpotees enamasti iseseisvat huvi ja teda kasutatakse vaid kui abihüpoteesi, mille kummutamine oleks aluseks ületamishüpoteesi või erinevushüpoteesi vastuvõtmisele. Ületamishüpoteesi nimetatakse ühepoolseks, erinevushüpoteesi kahepoolseks alternatiiviks.

Studenti test ühe valimi jaoks kontrollib hüpoteesi: juhusliku suuruse x keskväärtus võrdub etteantud väärtusega a. Lähteandmeteks on juhuslikku suurust esindav valim $x_1, x_2, \ldots, x_n$, mille järgi arvutatakse aritmeetiline keskmine $\bar{x}$ ja selle standardhälbe hinnang $s_{\bar{x}}$ (valem 20). Võrdsusest suhtelise kõrvalekaldumise hinnanguks on Studenti suhe

$$t = \frac{\bar{x} - a}{s_{\bar{x}}} . \qquad (32)$$

Vabadusastmete arv $f = n - 1$.

Studenti test kahe valimi jaoks kontrollib hüpoteesi: juhuslike suuruste x ja y keskväärtused on võrdsed. Suurust x esindab valim mahuga n_x, aritmeetilise keskmisega $\bar{x}$ ja standardhälbe hinnanguga s_x ning suurust y valim mahuga n_y, aritmeetilise keskmisega $\bar{y}$ ja standardhälbe hinnanguga s_y. Standardhälvete hinnangud arvutatakse valemi (19) järgi. Vabadusastmete arv $f = n_x + n_y - 2$ ja Studenti suhe

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\left(\frac{1}{n_x} + \frac{1}{n_y}\right) \frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{(n_x - 1) + (n_y - 1)}}} . \qquad (33)$$

Otsuse tegemine Studenti testi kohaselt toimub ühe valimi ülesande ja kahe valimi ülesande korral ühtviisi, oleneb aga al-

ternatiivist, mida võrdushüpoteesi kummutamine peaks kinnitama.

Kahepoolse alternatiivi (erinevushüpotees) vastuvõtmine on samaväärne võrdushüpoteesi kummutamisega. Võrdushüpotees kummutatakse tingimusel

$$|t| > t_{q',f} \quad , \tag{34}$$

kus $t_{q',f}$ on f vabadusastmega Studenti suhte kriitiline väärtus kahepoolse testi olulisuse nivool q'.

Ühepoolse alternatiivi (ületamishüpoteesi) vastuvõtmine ei ole samaväärne võrdushüpoteesi kummutamisega, sest tingimus (34) osutub täidetuks ka absoluutväärtuselt suure negatiivse t puhul. Eksitustest hoidumiseks võetakse ületamishüpotees vastu vaid tingimusel

$$t > t_{q,f} \quad , \tag{35}$$

kus q on ühepoolse testi olulisuse nivoo.

Tõese võrdushüpoteesi korral on Studenti suhte jaotus sümmeetriline ning ühe ja sama kriitilise väärtuse puhul sündmuse (34) tõenäosus (kahepoolse testi olulisuse nivoo) parajasti poole suurem kui sündmuse (35) tõenäosus (ühepoolse testi olulisuse nivoo). Siit

$$t_{q',f} = t_{2q,f} \quad . \tag{36}$$

Statistikatabelites on Studenti suhte kriitilised väärtused tavaliselt näidatud ühepoolse testi olulisuse nivoo q järgi. Võrdusest (36) tulenevad ümberarvutusvalemid

$$\begin{aligned} q &= \frac{q'}{2} = 1 - p = \frac{1 - p'}{2} \quad , \\ q' &= 2q = 2(1-p) = 1-p' \quad , \\ p &= 1-q = 1 - \frac{q'}{2} = \frac{1+p'}{2} \quad , \\ p' &= 1-2q = 1-q' = 2p-1 \quad , \end{aligned} \tag{37}$$

kus p on ühepoolse testi ja p' kahepoolse testi usaldusnivoo.

<u>Studenti testi rakendatavuse tingimused</u> on:

1) uuritavate suuruste jaotus peab olema normaalne,

2) kahe valimi võrdlemisel peavad juhuslike suuruste standardhälbed olema võrdsed.

Esimene tingimus ei tee praktikas pea kunagi takistusi. Studenti suhe arvutatakse aritmeetiliste keskmiste kaudu, aritmeetiline keskmine allub aga üsna täpselt normaaljaotusele ka siis, kui keskmistatava suuruse enda jaotus on hoopiski erinev (vt. p. 1.4). Jaotuse iseloom nõuab tähelepanu vaid siis, kui valimi maht ei ületa kuut-seitset objekti.

Teine tingimus võib osutuda häirivamaks. Standardhälvete hinnangute väike erinevus ei tee veel takistusi, see erinevus võib olla tingitud ainult juhusest, ning ka tõeliste standardhälvete väike erinevus ei põhjusta märgatavat eksitust. Tõsisemate erinevuste korral on aga parem valida teine meetod, näiteks Van der Waerdeni test.

Ühefaktorilise dispersioonanalüüsi ülesanne püstitatakse tavaliselt järgmisel viisil. Oletatakse, et mingi väline mõjufaktor võiks muuta mingi tunnuse keskväärtust. Sisulise hüpoteesi kontrolliks tehakse katse, kus erinevaid katsealuste rühmi mõjutatakse faktoriga erineval määral. Tunnuse mõõtmistulemused faktori erinevatel mõjutasemetel moodustavad valimid, millest igaüks esindab iseseisvana vaadeldavat juhuslikku suurust. Kontrollitav formaalne nullhüpotees väidab: kõigi vaatlusaluste juhuslike suuruste keskväärtused on võrdsed. Nullhüpoteesi kummutamine tähendaks mõju reaalsust väitva alternatiivi vastuvõtmist.

Ühefaktoriline dispersioonanalüüs on Studenti testi üldistus mitme valimi üheaegse võrdlemise ülesandele. Ka eeldused on samad: normaaljaotus ning teoreetiliselt võrdsed dispersioonid. Valimite aritmeetiliste keskmiste suhtelise hajuvuse mõõduks on Fisheri statistik ehk dispersioonide suhe F, mida võrreldakse Fisheri-Snedecori jaotuse kriitiliste väärtustega. Täpsema arvutuseeskirja võib leida programmi ÜHEFAKTORILINE DISPERSIOONANALÜÜS kirjeldusest.

Ühefaktoriline dispersioonanalüüs kontrollib valimite erinevusi ainult komplektis ning üksikute paaride jaoks siit võrdlustulemusi ei saa. Erijuhul võib vaatluse alla võtta vaid kaks valimit, siis osutub dispersioonanalüüs täpselt võrdväärseks Studenti testiga kahepoolse alternatiiviga ülesande jaoks.

Kahefaktorilise dispersioonanalüüsi korral kontrollitakse korraga kahe faktori mõju ühele tunnusele. Olgu ühel faktoril (nimetame seda a-faktoriks) m erinevat taset ja teisel faktoril (nimetame seda b-faktoriks) n erinevat taset. Täisfaktorilise eksperimendi puhul sooritatakse vaatlused kõikvõimalikel faktoritasemete kombinatsioonidel, mida on kokku mn. Vaatluste arv tuleb suur juba siis, kui uurimisele võtta vaid üks objekt. NAIRII-2 statistikaprogrammide süsteemis ongi realiseeritud just kahefaktorilise dispersioonanalüüsi selline lihtsaim variant.

Vaatlustulemused korraldatakse tabelisse, kus a-faktori igale tasemele vastab rida ning b-faktori igale tasemele veerg. Lahtreid jääb parajasti mn. Nullhüpoteese on kaks:

1) a-faktor ei mõjuta tunnuse väärtust,

2) b-faktor ei mõjuta tunnuse väärtust.

Kummagi hüpoteesi jaoks arvutatakse eraldi Fisheri statistik (dispersioonide suhe), mida võrreldakse vastava kriitilise väärtusega.

Näide. Seitsmeliikmeline rühm komplekteeritakse ühtlaselt hea joonistamisoskusega õpilastest. Rühma ühtluse kontrollimiseks antakse kõigile ühesugune ülesanne, valminud seitset joonistust hindavad viis eksperti. Kümnepallisüsteemis olgu hinded järgmised:

Ekspert / Õpilane	V	W	X	Y	Z
A	5	6	6	7	5
B	7	6	6	8	5
C	6	7	5	7	6
D	5	7	6	7	5
E	5	7	6	6	5
F	6	4	6	7	4
G	6	6	6	7	4

Objektiks on joonistus, tunnuseks üldine tase, a-faktoriks õpi-

lase ja b-faktoriks eksperdi individuaalsus. Faktorite skaalad on vaid nimeskaalad, mis ei takista analüüsi. Hinnete skaalat käsitletakse siin kui meetrilist skaalat.

Programmi KAHEFAKTORILINE DISPERSIOONANALÜÜS abil teostatud arvutustest järeldub, et esimest nullhüpoteesi ei saa kummutada (usaldatavus on kõigest 54 %) ning õpilaste rühma ebahomogeensus pole märgatav. Ekspertide hindamissüsteemid osutuvad aga oluliselt erinevateks (teise nullhüpoteesi kummutamise usaldatavus on 99,97 %).

Mitteparameetrilised meetodid. Senikäsitletud keskmiste võrdlemise meetodeid nimetatakse parameetrilisteks, sest siin võrreldakse jaotusparameetreid: keskväärtusi. Keskväärtus on sisuliselt tõlgendatav vaid meetrilise skaala korral. Kui meetrilist skaalat teisendada, seda ühest otsast venitades ja teisest kokku surudes, siis keskväärtus nihkub ning eelkirjeldatud testide abil saadavad tulemused muutuvad. Parameetrilised meetodid on põhjendatud vaid siis, kui uuritaval suurusel on loomulik meetriline skaala.

Mitteparameetrilisteks nimetatakse niisuguseid statistikameetodeid, mille puhul parameetritele tähelepanu ei pöörata ning jaotusseaduse kohta eeldusi ei tehta (normaaljaotushüpotees pole tarvilik). Nende meetodite abil saadud järeldused ei olene skaala üksikute lõikude kokkusurumisest või venitamisest. Tunnuse kirjeldamisel piisab ka järjeskaalast.

Järjehinded ja normaalhinded on hõlpsaim vahend üleminemiseks mitteparameetrilisele töötlusele. Järjed ei olene mõõtskaala lõikude kokkusurumisest või venitamisest. Järjehinnete jaotus on alati lähedane ühtlasele jaotusele, normaalhinnete jaotus aga normaaljaotusele. Praktilistes rakendustes pakuvad huvi peaasjalikult normaalhinded.

Mõnel juhul õnnestub uurimistööd organiseerida niiviisi, et normaalhinded arvutatakse suure kogumi jaoks, võrreldavad valimid moodustavad aga kumbki vaid väikese osa sellest kogumist. Siis on normaalhinnete skaala võrdlusaluste valimite jaoks välise päritoluga, seda võib käsitada nagu tavalist katsest sõltumatut skaalat ning rakendada võrdlusülesande lahendamiseks Studenti testi või dispersioonanalüüsi.

Kui aga lisavaatlusi pole tehtud, osutuvad normaalhinded kohandatuks konkreetsele valimile, nende skaalat ei saa käsitada valimist sõltumata mõõtskaalana ja dispersioonanalüüs, eriti aga Studenti test, osutuvad vaid jämedateks lähendusmeetoditeks. Täpne meetod on sel juhul Van der Waerdeni test.

Van der Waerdeni testi statistik kirjeldab kahe valimi ühenduse järgi arvutatud valimitevahelist keskmist nihet normaalhinnetest lähtudes. Testi kasutamine on analoogiline Studenti testi kasutamisega. Erinevalt Studenti testist pole tarvis eeldada uuritavate suuruste normaaljaotust ning lõppjäreldus ei sõltu mõõtmisel kasutatud skaalast. Täpsemaid juhiseid võib leida programmi TASEMETE VÕRDLUS kirjeldusest.

1.13. ÜHERÜHMAEKSPERIMENT

Eksperimendi skeem. Pedagoogikaeksperimendi võimalikke skeeme on palju. Käesolevas ja järgmises punktis vaadeldakse kahte lihtsaimat tüüpskeemi.

Üherühmaeksperimendis on vaatluse all üksainus õpilaste rühm, tavaliselt klass. Eksperiment koosneb kolmest etapist:

1) teadmiste algkontroll,
2) katsealuste mõjutamine,
3) teadmiste lõppkontroll.

Rühma nimekirja jäetakse ainult need õpilased, kes on osa võtnud kõigist kolmest etapist. Eksperimendi tulemused võetakse kokku tabelisse, kus iga õpilase nime järel seisab alghinne ja lõpphinne.

Protsenthinded. Uurimistöös mõõdetakse teadmisi sageli testide abil, numbriliseks tulemuseks on pallide arv. Kui maksimaalne võimalik pallide arv on erinevate kontrollide korral erinev, pole tulemused otseselt võrreldavad. Järgnevas eeldatakse kõikjal, et hinded on väljendatud ühes ja samas mõõtskaalas. Tarvidüse korral kasutatakse protsenthindeid:

$$\text{protsenthinne} = 100\,\frac{\text{originaalhinne}}{\text{maksimaalhinne}}\,. \qquad (38)$$

Efekt. Individuaalseks formaalseks efektiks loetakse lõpphinde ja alghinde vahet. Kui hinneteskaala on ühtlane meetriline skaala, siis on ka efekt meetriline suurus. Kui hinneteskaala pole ühtlane või on vaid järjeskaala, siis omistatakse efektile üks kolmest võimalikust väärtusest:

- \- hinne langes,
- 0 hinne jäi samaks,
- \+ hinne tõusis.

Järgnevas eeldatakse, et algsele hinnetetabelile on lisatud veerg, milles on näidatud individuaalsed efektid (vt. tabel 7).

Tabel 7

Protsenthinnete ja efektide arvutamine (näide)

Õpilane	Originaalhinne		Protsenthinne		Efekt	
	Alghinne (max=25)	Lõpphinne (max=50)	Alg-hinne	Lõpp-hinne	Kvanti-tatiivne	Kvalita-tiivne
A	17	38	68	76	8	+
B	21	41	84	82	-2	-
C	13	33	52	66	14	+
D	16	32	64	64	0	0
.....	....	....	....	....	....	....
.....	....	....	....	....	....	....

Märgitest. Olgu positiivse efektiga õpilaste arv n_+, negatiivse efektiga õpilaste arv n_- ja nullist erineva efektiga õpilaste üldarv $n = n_+ + n_-$. Püstitame nullhüpoteesi "mõju puudub" ehk formaalsetes terminites "positiivse efekti tõenäosus nullist erinevate efektide hulgas on 50 %". Sama tõenäosuse eksperimendist saadud hinnang on suhteline sagedus n_+/n. Hüpoteesi kontrollimine on taandatud juhusliku sündmuse tõenäosuse võrdlemisele antud tõenäosusega (vt. p. 1.11). Ülesande lahendab programm TÕENÄOSUSTE VÕRDLUS. Programm annab vastuseks kahepoolse alternatiivi "mõju ei puudu" vastuvõtmise usaldatavuse p'. Ühepoolset alternatiivhüpoteesi "mõju on positiivne" võib vastu võtta ainult positiivsete efektide ülekaalu korral, usaldatavus on siis programmi abil leitust suurem:

$$p = \frac{p' + 1}{2} \qquad (39)$$

(võrdle p. 1.12, otsuse tegemine).

Normitabel ja ülenormiefekt. Katsetes, kus mõjutamine seisneb kontrollimisele kuuluva aine õpetamises, on mõju olemasolu ette teada ning küsitakse, kas mõju on piisav. Piisavaks arvatava efekti arvuline väärtus oleneb alghindest. Normitabel koosneb kahest veerust: esimesse veergu kirjutatakse järjekorras kõik võimalikud alghinded, teise - piisavale efektile vastavad lõpphinded. Normitabeli koostamine on komplitseeritud ülesanne, mille lahendamisel on otsustav sõna sisulistel kaalutlustel. Matemaatilistest meetoditest saab kasutada regressioonianalüüsi (vt. p. 1.19).

Kui normitabel olemas, võib pedagoogikaeksperimendi tulemuste tabelisse efektide asemel kanda ülenormiefektid. Ülenormiefekt on tegeliku lõpphinde ning alghinde ja normitabeli abil leitud normatiivse lõpphinde vahe.

Mõju piisavust kontrollitakse ülenormiefektide järgi täpselt samal viisil, kui kontrolliti mõju olemasolu efektide järgi.

Studenti test. Kui hinnete skaalat võib reaalsetele hinnetele vastavas vahemikus tõlgendada ühtlase meetrilise skaalana, siis tohib eelvaadeldud ülesannete lahendamiseks kasutada ka Studenti test, mis on tavaliselt märgitestist veidi tundlikum. Võrdlemisele võetakse ühelt poolt efekti või ülenormiefekti keskväärtus ja teiselt pool konstant 0. Studenti suhe arvutatakse valemi (32) kohaselt, kus a = 0.

Studenti testi on mõtet kasutada vaid juhul, kui märgitesti abil leitud usaldatavus osutub veidi väiksemaks vajalikust.

Probleemi püstitamise tüüpviga ilmneb sageli algajate uurijate töödes. Eksperimendi statistilise analüüsi abil püütakse tõestada õpetamise intensiivistamise mõju õpetatava materjali tundmisele. Hüpotees osutub triviaalseks ega vajagi katselist kontrolli: et õpetamisel teadmised kasvavad, on niigi teada. Võib küll juhtuda, et mõju on väike ja kontrollimisel tuleb usaldatavus madalavõitu. Eriti iseloomulik on taoline tulemus ebaratsio-

naalse töötlusskeemi korral (vt. järgmine alapunkt). See võib luua probleemi mõttekuse illusiooni.

Mõju olemasolu ning suund vajavad kontrollimist vaid siis, kui tulemus pole ette teada. Näiteks võib uurida, kuidas konkreetsetes tingimustes kehalise treeningu intensiivistamine mõjub õpilase matemaatiliste võimete arengule. Kui uuritavaks mõjufaktoriks on aga matemaatika õpetamine, siis väärivad tähelepanu vaid mõju suuruse hindamise, mõju piisavuse kontrollimise ja kahe erineva mõju võrdlemise ülesanded.

Töötlusmeetodi valiku tüüpviga ilmneb nende uurijate töödes, kes pole vaevunud probleemi matemaatilisse olemusse süvenema. Alghinnete ja lõpphinnete hulki võrreldakse kui kahte iseseisvat valimit Studenti testi, Van der Waerdeni testi või muu analoogilise statistilise testi abil. Niiviisi jääb kasutamata lõpphinnete ja alghinnete vahelises sõltuvuses peituv informatsioon, see aga tähendab praktikas mõju kontrolli tundlikkuse õige suurt langust.

1.14. KAHERÜHMAEKSPERIMENT

Eksperimendi skeem. Õpilased alluvad katse käigus mitmesugustele kõrvalmõjudele. Kõrvalmõjude kontrollimiseks võetakse vaatluse alla teine õpilaste rühm, mida nimetatakse kontrollrühmaks. Kontrollrühmas lastakse toimida ainult kõrvalmõjudel, esimeses rühmas, mida nimetatakse eksperimentaalrühmaks, aga samadel kõrvalmõjudel koos sihipärase mõjuga. Vaatlustulemused kantakse tabelis 8 näidatud vormiga tabelisse.

Tabel 8

Kaherühmaeksperimendi tabeli vorm

Eksperimentaalrühma õpilaste nimed	Alghinded	Lõpphinded
Kontrollrühma õpilaste nimed	Alghinded	Lõpphinded

Statistilise töötluse eesmärgiks on võrrelda efekte eksperimentaal- ja kontrollrühmas.

Mõne eksperimendi korral mõjutatakse kumbagi rühma sihipäraselt, kuid erineval viisil. See võib muuta terminoloogiat ja tulemuste tõlgendamist, statistilise töötluse skeem jääb aga samaks.

Märgitest. Sihipärase mõju kontrollimiseks varustatakse kaherühmaeksperimendi tabeli kumbki osa efektide ehk ülenormiefektide veeruga. Piisab kvalitatiivsest skaalast (-, 0, +). Nulleefektiga õpilased langevad vaatluse alt välja. Olgu eksperimentaalrühmas positiivse efektiga õpilasi n_+ ja nullist erineva efektiga õpilasi n ning kontrollrühmas positiivse efektiga õpilasi m_+ ning nullist erineva efektiga õpilasi m. Sihipärase mõju puudumise ehk mõjude võrdsuse kontroll taandub nüüd positiivse efekti tõenäosuste võrdlemisele suhteliste sageduste n_+/n ning m_+/m järgi. Ülesanne lahendatakse p. 1.11 kirjeldatud meetodil programmi TÕENÄOSUSTE VÕRDLUS abil. Tulemuste tõlgendamisel peab arvestama üherühmaeksperimendi käsitlemisel tehtud märkusi alternatiivi kohta.

Kaherühmaeksperimentide tulemuste vaatlemisel ilmneb, et sageli on nii eksperimentaal- kui kontrollrühmas pea kõik efektid positiivsed. Märgitest kaotab sel juhul tundlikkuse. Kirjeldatud olukorrast on kaks väljapääsu.

1) Lõpptaseme test koostatakse algtaseme testist niipalju raskem, et hinded tuleksid keskelt läbi samad ja pooled formaalsetest efektidest oleksid negatiivsed.

2) Efektide asemel võrreldakse ülenormiefekte. Sobivalt valitud normitabeli korral on positiivse ülenormiefekti tõenäosus ligikaudu 50 %, mis tagab märgitesti maksimaalse tundlikkuse.

Studenti test on kasutatav samadel tingimustel kui üherühmaeksperimendi korral. Nüüd on aga tarvis võrrelda kaht (eksperimentaalrühma ja kontrollrühma) efektide või ülenormiefektide keskväärtust omavahel. Selleks arvutatakse kummagi rühma jaoks efektide aritmeetiline keskmine ja standardhälve (programm ELEMENTAARSTATISTIKA) ja kasutatakse siis programmi STUDENTI TEST.

Algtasemete erinevuse mõju katsetulemusele. Kui üks ja sama mõju annab võrdse numbrilise efekti nii madala kui kõrge algta-

seme korral, siis pole rühmade algtasemete erinevus üldse häiriv. Tavaliselt aga on formaalne efekt madala algtaseme korral suurem kui kõrge algtaseme korral. Kui hinded on ülalt tõkestatud (näiteks protsenthinded), siis kaob maksimaalse alghinde korral positiivse formaalse efekti võimalus hoopiski. Niisugust mõõtskaalat ei saa käsitada kui ühtlast meetrilist skaalat. Oletame nüüd, et rühmad alluvad täpselt ühesugusele mõjule, eksperimentaalrühma algtase on aga kontrollrühma algtasemest kõrgem. Ilmselt tulevad kontrollrühmas numbrilised efektid keskelt läbi suuremad kui eksperimentaalrühmas ja formaalne võrdlus Studenti testi abil näitaks, et kontrollrühmas on efekt suurem. Siit on kerge tulema eksijäreldus, et kontrollrühmas on ka mõju suurem. Väär tulemus pole statistikameetodi süü, vaid on tingitud meetodi valimisel ja formaalse tulemuse tõlgendamisel tehtud veast.

Märgitesti korral eelkirjeldatud eksitust ei teki. Algtasemete erinevuse korrigeerivad siin hindamismeetod või normitabel. Õige normitabeli korral on algtasemete erinevuse mõju välistatud.

Algtasemete erinevust püütakse pedagoogikaeksperimendi korraldamisel vältida. Tavaliselt pole aga eksperimentaatoril võimalust rühmi vabalt koostada ning ta on sunnitud leppima ka ebasoodsa olukorraga.

Nivelleeriv skaalateisendus. Sageli ei saa märgitesti kasutada normitabeli puudumise tõttu. Kui mõõtskaalat võib siiski meetriliseks, olgugi ebaühtlaseks lugeda, jääb veel võimalus realiseerida selline skaalateisendus, mille järel üks ja sama mõju annaks alati ühesuguse numbrilise efekti, ja kasutada seejärel Studenti või Van der Waerdeni testi. Kirjeldatud võte on realiseeritud programmis MÕJUDE VÕRDLUS. Uue skaala saamiseks kasutatakse siin astmeteisendust (vt. p. 1.6), astmenäitaja valitakse katseandmete alusel niiviisi, et rühmade sees numbriline efekt ei sõltuks algtasemest. Sõltuvust kontrollitakse lineaarse korrelatsioonikordaja järgi. Meetod on ligikaudne ning õigustatud juhul, kui hinded pole maksimaalhinde lähedal. Ebasoodsas olukorras on tarvis hinded eelnevalt teisendada miinuspunktide süsteemi.

Veidi teistsuguse nivelleeriva skaalateisendustehnika kirjelduse koos rakendusnäidetega võib leida uurimistööst [11].

Näide. Kaherühmaeksperimendis kasutati kümnepallist hinneteskaalat. Eelnevate uurimiste alusel on koostatud efekti normitabel (tabel 9).

Tabel 9

Efekti normitabel (näide)

Alghinne	Normaallõpphinne
1	4
2	5
3	6
4	7
5	7
6	8
7	9
8	9
9	10
10	10

Hinded ja individuaalsed efektid on esitatud tabelis 10.

Märgitest lihtefektide järgi on ilmselt mõttetu. Positiivseid ülenormiefekte on nullist erinevate ülenormiefektide seas eksperimentaalrühmas 86 % ja kontrollrühmas 33 %. Tõenäosuste võrdlemise programmi järgi saame võrdsushüpoteesi kummutamise usaldatavuseks 94 %. See tulemus on põhjendatud niivõrd, kuivõrd on põhjendatud normitabel. Kvantitatiivsete efektide võrdlemine Studenti testi abil on alusetu. Programm MÕJUDE VÕRDLUS valib skaalateisenduse astmenäitajaks 1,2 ning leiab võrdsushüpoteesi kummutamise usaldatavuseks 43 %.

1.15. VAHEMIKHINNANGUD

Usaldusvahemik on niisugune juhuslike rajadega vahemik, mis katab vaadeldava parameetri tõelise väärtuse etteantud tõenäosusega. Etteantud tõenäosust nimetatakse usaldusnivooks ning tähistatakse p.

Tabel 10

Kaherühmaeksperimendi analüüs

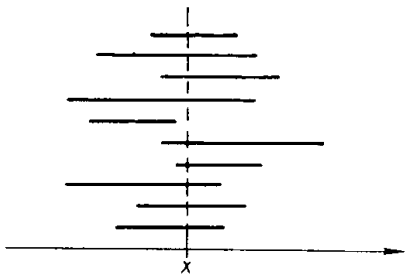
	Õpilane	Alg-hinne	Lõpp-hinne	Efekt	Ülenormi-efekt
Eksperimentaalrühm	EA	4	9	5	+2
	EB	6	6	0	-2
	EC	7	10	3	+1
	ED	2	5	3	0
	EE	5	8	3	+1
	EF	8	9	1	0
	EG	6	10	4	+2
	EH	7	9	2	0
	EI	6	9	3	+1
	EJ	5	8	3	+1
Kontrollrühm	KA	3	5	2	-1
	KB	2	5	3	0
	KC	6	6	0	-2
	KD	3	5	2	-1
	KE	5	8	3	+1
	KF	3	6	3	0
	KG	4	7	3	0
	KH	2	7	5	+2
	KI	4	6	2	-1

Usalduspiirid on usaldusvahemiku rajad. Parameetri x usalduspiire usaldusnivool p tähistatakse x_p^- ja x_p^+. Asjaolu, et usaldusvahemik katab parameetri tõelise väärtuse, võib üles kirjutada võrratusena $x_p^- \leqslant x < x_p^+$. See võrratus osutub ekslikuks tõenäosusega 1-p.

Vahemikhinnang on uuritava parameetri jaoks vaatlusandmete järgi koostatud usaldusvahemik. Vahemikhinnanguid kasutatakse paralleelselt punkthinnangutega. Parameetri x punkthinnangut tähistatakse allpool $\dot{x}$. Punkthinnang ei ütle midagi hindamistäpsuse kohta, vahemikhinnang aga kirjeldab ka hindamise täpsust.

Kõigi praktikas tarvitust leidvate hindamismeetodite korral jääb parameetri punkthinnang usaldusvahemiku sisse: võrratus $x_p^- \leq \dot{x} < x_p^+$ kehtib alati.

Näide. Joonisel 9 on graafiliselt kujutatud ühe ja sama parameetri x kümme vahemikhinnangut, igaüks leitud erineva valimi järgi. Joonisel näidatud parameetri tõeline väärtus jääb praktikas muidugi tundmatuks. Usaldusnivooks on valitud 90 %. Nagu oodata võib, on ka parajasti 9 vahemikhinnangut 10-st õiged.



Joon. 9. Kümme vahemikhinnangut.

Vahemikhinnanguid moodustatakse statistiliste hüpoteeside kontrollimistestide abil. Usaldusvahemikku võetakse kõik need väärtused, mille puhul hüpoteesi "parameetri tõeline väärtus võrdub vaatlusaluse väärtusega" pole võimalik valitud usaldusnivool kummutada. Mida kõrgem usaldusnivoo, seda laiem tuleb usaldusvahemik.

Piirvead on usalduspiiride ja punkthinnangu vahed: $\delta_p^- = \dot{x} - x_p^-$, $\delta_p^+ = x_p^+ - \dot{x}$. Vahemikhinnang $x_p^- \leq x < x_p^+$ kirjutatakse piirvigade kaudu üles järgmiselt: $x = \dot{x} \begin{matrix} +\delta_p^+ \\ -\delta_p^- \end{matrix}$.

Sümmeetrilise usaldusvahemiku korral on piirvead võrdsed ja neid tähistatakse δ_p. Vahemikhinnang kirjutatakse kujul

$$x = \dot{x} \pm \delta_p \ . \qquad (40)$$

Tõenäosuse vahemikhinnangut saab moodustada χ^2-testi abil (vt. p. 1.11). Usalduspiiride leidmiseks kirjutatakse valemi (31)

vasakule poole usaldusnivoole vastav χ^2 kriitiline väärtus ning vaadeldakse etteantud tõenäosust kui tundmatut. Niiviisi koostatud võrrandi lahendid ongi usalduspiirid. Usaldusvahemik tuleb ebasümmeetriline.

Praktiliste ülesannete lahendamisel kasutatakse programmi TÕENÄOSUSE USALDUSPIIRID.

Keskväärtuse vahemikhinnangut saab moodustada Studenti testi abil. Usaldusvahemik tuleb sümmeetriline, piirviga võrdub aritmeetilise keskmise standardhälbe hinnangu ja Studenti suhte kriitilise väärtuse korrutisega. Tulemuste esitamisel kasutatakse reeglina vormi (40).

Keskväärtuse piirviga arvutavad programmid ELEMENTAARSTATISTIKA, PÕHISTATISTIKUD ja TUNNUSTE ANALÜÜS.

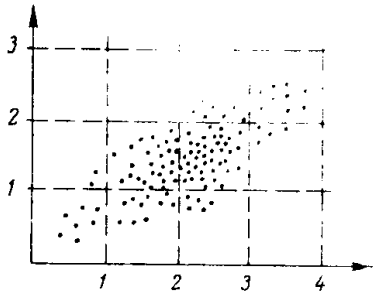
1.16. STATISTILINE SÕLTUVUS

Kahemõõtmeline sagedustabel. Võtame uurimisele mitme tunnuse abil kirjeldatavad objektid ning valime vaatlusalusteks kaks tunnust nimedega x-tunnus ja y-tunnus. Eespool kirjeldati ühe isoleeritud tunnuse ühemõõtmelist sagedustabelit ja histogrammi. Nüüd koostame kahemõõtmelise tabeli, kus iga rida vastab ühele x-tunnuse väärtusele või väärtuste klassile, iga veerg ühele y-tunnuse väärtusele või väärtuste klassile ja igas lahtris on näidatud nende objektide arv, mille tunnuste väärtused vastavad lahtri reale ja veerule.

Kahemõõtmelise sagedustabeli reasummad kirjutatakse lisaveeruna viimase veeru järele ning nad moodustavad tavalise ühemõõtmelise sagedustabeli x-tunnuse jaoks. Veerusummad kirjutatakse lisaritta ja nad kirjeldavad y-tunnuse jaotust. Kahemõõtmelise sagedustabeli näide on tabel 11.

Korrelatsiooniväli. Kahemõõtmelise sagedustabeli järgi saaks histogrammi ehitada vaid ruumilise mudelina. Sõltuvuse graafiliseks kujutamiseks kasutatakse korrelatsioonivälja (vt. joonis 10). Iga objekti kujutab punkt, mille koordinaatideks on vaatlusaluste tunnuste väärtused. Pideva mõõtskaala korral on korrelatsiooniväli üksikasjalikum kui sagedustabel,sest siin po-

le tarvis mõõtskaalat klassideks jagada. Sagedustabeli moodustamiseks korrelatsioonivälja järgi kantakse graafikule koordinaatvõrk (kriipsjooned joonisel 10). Ruutudesse jäävate punktide arvud kirjutatakse sagedustabeli lahtritesse.



Joon. 10. Korrelatsiooniväli (näide).

Tinglik tõenäosusjaotus on ühe tunnuse tõenäosusjaotus teise tunnuse fikseeritud väärtuse juures. Kahemõõtmelise sagedustabeli read on y-tunnuse tinglikku tõenäosusjaotust kirjeldavad ühemõõtmelised sagedustabelid rea ees näidatud x-tunnuse väärtuste jaoks. Niisamuti on veerud x-tunnuse tinglikud ühemõõtmelised sagedustabelid. Tavalist tõenäosusjaotust, mida kirjeldab summade rida või veerg, nimetatakse tingimuseta ehk marginaalseks tõenäosusjaotuseks.

Sõltumatuse tingimus. x-tunnust nimetatakse y-tunnusest sõltumatuks siis, kui x-tunnuse tinglik tõenäosusjaotus ei olene y-tunnuse väärtusest ning langeb alati kokku tingimuseta tõenäosusjaotusega. Kui x-tunnus on y-tunnusest sõltumatu, siis on paratamatult ka y-tunnus x-tunnusest sõltumatu. Seetõttu räägitakse tunnuste vastastikusest sõltumatusest.

Sõltumatute tunnuste sagedustabeli restaureerimine. Oletame, et x ja y on sõltumatud tunnused ning nende sagedustabelist on kõik arvud peale tunnuste väärtuste, reasummade ja veerusummade kustunud. Sõltumatuse eeldus võimaldab tabelit ligikaudselt täita ainult rea- ja veerusummade järgi. Olgu tabeli i-nda rea summa n_i, j-nda veeru summa n_j ning objektide üldarv n. Objekti i-ndasse ritta sattumise tõenäosuseks võtame hinnangu $\frac{n_i}{n}$, i-nda rea objekti j-ndasse veergu sattumise tõenäosuseks hinnangu $\frac{n_j}{n}$. Ükskõik millise rea objekt satub ühtaegu i-ndasse ritta ja

j-ndasse veergu tõenäosusega $\frac{n_i}{n} \cdot \frac{n_j}{n}$. Lahtrisse peab sattuma keskmiselt $\frac{n_i}{n} \cdot \frac{n_j}{n} \cdot n = \frac{n_i n_j}{n}$ objekti.

Restaureeritud tabel rahuldab absoluutselt täpselt sõltumatuse tingimust. Sõltumatute tunnuste tegelik sagedustabel erineb restaureeritud tabelist juhuslike hälvete võrra.

Näide. Tabel 11 esitab mingi katse järgi leitud kahemõõtmelist jaotust. Tabel 12 on moodustatud tunnuste sõltumatust eeldava restaureerimiseeskirja kohaselt eelmise tabeli summarea ja summaveeru järgi. Lahtritesse kantud väärtused on ümardatud täisarvudeks.

Tabel 11

Kahemõõtmeline sagedustabel (näide)

x \ y	1	2	3	Σ
1	0	3	2	5
2	7	9	3	19
3	10	24	17	51
4	1	9	15	25
Σ	18	45	37	100

Tabel 12

Restaureeritud sagedustabel (näide)

x \ y	1	2	3	Σ
1	1	2	2	5
2	3	9	7	19
3	9	23	19	51
4	5	11	9	25
Σ	18	45	37	100

Sõltumatuse kontroll. Tõeliselt sõltumatute tunnuste puhul ei saa eelkirjeldatud viisil restaureeritud ("teoreetiline") sagedustabel ja tegelik sagedustabel teineteisest oluliselt erineda. Kahe tabeli erinevuse mõõduks on teatavasti statistik χ^2 (valem 28). Tähistame originaaltabeli i-ndas reas ja j-ndas veerus seisva arvu n_{ij}. Restaureerimiseeskirja silmas pidades tuleneb valemist (28) järgmine valem

$$\chi^2 = n\left(\sum_{i=1}^{k_x}\sum_{j=1}^{k_y}\frac{n_{ij}^2}{n_i n_j} - 1\right), \qquad (41)$$

kus k_x on ridade ning k_y veergude arv. Vabadusastmete arv pole aga $k_x k_y - 1$, vaid $k_x + k_y$ võrra väiksem

$$f = (k_x - 1)\,(k_y - 1)\,, \qquad (42)$$

sest restaureeritud tabelisse on üle kantud $k_x + k_y$ summat originaaltabelist. Kui χ^2 ületab kriitilise väärtuse, siis kummutab χ^2-test tunnuste sõltumatuse hüpoteesi.

Sõltumatuse kontroll χ^2-testi abil on õigustatud kõigi mõõtskaalade korral. Praktiliste ülesannete lahendamiseks kasutatakse programmi SAGEDUSTABEL.

Kovariatsioon. Ühe tunnuse varieeruvuse mõõduna kasutatud dispersioon oli teatavasti tunnuse hälbe ruudu keskväärtus. Kahe tunnuse koosvarieeruvuse mõõduna kasutatakse kovariatsiooni, mis on defineeritud kui tunnuste hälvete korrutise keskväärtus

$$\mathrm{cov}(x,y) = \mu[(x - \mu_x)(y - \mu_y)]\,. \qquad (43)$$

Tunnuse kovariatsioon iseenesega osutub võrdseks dispersiooniga.

Sõltumatute tunnuste kovariatsioon on null. Vastupidine ei pruugi olla õige. Ka sõltuvate suuruste kovariatsioon võib olla erijuhul null. Sõltuvus pole aga sel juhul monotoonne.

Kahe juhusliku suuruse summa dispersioon võrdub liidetavate dispersioonide ja kahekordse kovariatsiooni summaga

$$\sigma^2_{x+y} = \sigma^2_x + \sigma^2_y + 2\,\mathrm{cov}(x,y). \qquad (44)$$

1.17. SEOSEKORDAJAD

Crameri kordaja. Kahe suuruse vahelise sõltuvuse tugevust iseloomustab üsna täielikult eespool defineeritud statistik χ^2 (valem 41). Suurima võimaliku väärtuse χ^2_{max} saavutab χ^2 juhul, kui ühe suuruse väärtus on teise suuruse väärtuse järgi üheselt määratav (funktsionaalne sõltuvus). Kuna χ^2_{max} sõltub nii objektide arvust kui tabeli mõõtmetest, siis osutub χ^2 vahetu väärtuse tõlgendamine tülikaks. Crameri kordaja arvutatakse väärtuse taandamise teel:

$$C = \sqrt{\frac{\chi^2}{\chi^2_{max}}} \qquad (45)$$

Kordaja vähim võimalik väärtus on null (täielik sõltumatus) ning suurim võimalik väärtus üks (üksühene sõltuvus). Juhuslike hälvete tõttu tulevad χ^2 ja Crameri kordaja veidi nullist erinevad ka siis, kui suurused on tõeliselt sõltumatud. Kordaja väärtust võib sõltuvuse tugevuse mõõduna interpreteerida vaid siis, kui sõltumatuse hüpotees on kummutatud.

Kirjandusest tuntud Tšuprovi kordaja erineb Crameri kordajast ainult sellepoolest, et siin kasutatakse χ^2_{max} täpse väärtuse asemel ligikaudset väärtust. Ruuttabeli korral on Tšuprovi kordaja täpselt võrdne Crameri kordajaga, ristküliktabeli korral aga veidi väiksem.

Crameri kordajat võib kasutada kõigi mõõtskaalade puhul. Põhirakenduseks on nimeskaalas mõõdetud tunnuste sõltuvuse kirjeldamine, sest nimeskaala korral pole teised seosekordajad, välja arvatud Pearsoni kordaja, kasutatavad. Crameri kordaja ei muutu tunnuste ümberskaleerimisel ega ka väärtuste ümberjärjestamisel. Arvutus lähtub alati lõplikust sagedustabelist, mistõttu meetriliste tunnuste puhul on tarvilik klassiteisendus. Meetrilise tunnuse sõltuvust iseloomustatakse Crameri kordajaga peaasjalikult sõltuvuse mitmese ehk oluliselt ebalineaarse (näiteks võnkuv sõltuvus) loomuse puhul.

Crameri kordaja arvutab programm SAGEDUSTABEL.

<u>Korrelatsioonisuhe</u>. Oletame nüüd, et y-tunnuse mõõtskaala on meetriline. Tunnuse väärtuste hajuvusel on kaks põhjust: hajuvus sagedustabeli ridade sees ning ridadevahelised erinevused. Ridadevaheliste erinevuste kirjeldamiseks võtame vaatluse alla y-tunnuse keskväärtused üksikutes ridades. Niisuguseid keskväärtusi $\mu_{y/x}$ nimetatakse tinglikeks keskväärtusteks antud x korral. Ridadevaheliste erinevuste mõõduks on tinglike keskväärtuste standardhälve üle kogu sagedustabeli $\sigma(\mu_{y/x})$. y-tunnuse korrelatsioonisuhe x-tunnusega

$$\eta_{y/x} = \frac{\sigma(\mu_{y/x})}{\sigma_y} \qquad (46)$$

näitab, kui suur osa y-tunnuse üldisest standardhälbest σ_y on põhjustatud ridade tinglike keskväärtuste erinevusest, teisiti öeldes, x-tunnuse muutumisest. Allpool nimetatakse x-tunnust argumenttunnuseks.

Sageli öeldakse, et korrelatsioonisuhte ruut $\eta^2_{y/x}$ näitab, kui suur osa y-tunnuse hajuvusest on põhjustatud x-tunnuse muutustest. Korrelatsioonisuhte ruut on dispersioonide suhe ja et dispersioone saab vahetult liita, siis on sõna "osa" siin enam õigustatud kui korrelatsioonisuhte tõlgendamisel standardhälvete kaudu.

Korrelatsioonisuhte vähim võimalik väärtus on 0 ja suurim võimalik väärtus 1. Et standardhälbeid saab vaatluste järgi vaid hinnata, jääb ka korrelatsioonisuhte täpne "teoreetiline" väärtus praktikas tundmatuks ning vaatlusandmete töötlus annab üksnes korrelatsioonisuhte hinnangu.

Kui x-tunnuse mõõtskaala on meetriline, saab analoogilisel viisil defineerida x-tunnuse korrelatsioonisuhte y-tunnusega $\eta_{x/y}$.

Korrelatsioonisuhe ei muutu argumenttunnuse ümberskaleerimisel ja selle väärtuste ümberjärjestamisel, küll aga muutub esimese tunnuse ümberskaleerimisel. Korrelatsioonisuhte põhirakenduseks on ülesanded, kus esimene tunnus on mõõdetud meetrilises, teine aga nimeskaalas. Üsna tihti aga arvutatakse korrelatsioonisuhteid ka kahe meetrilise tunnuse vahel, seda eriti oluliselt ebalineaarse sõltuvuse korral. Meetrilise argumenttunnuse korral peab kasutama klassiteisendust.

Korrelatsioonisuhteid arvutavad programmid SEOSEANALÜÜS ja SAGEDUSTABEL.

Näide. Sagedustabeliga (tabel 13)

Tabel 13

Sagedustabel (näide)

x \ y	1	2
1	100	0
2	0	100
3	100	0

kirjeldatud sõltuvust iseloomustavad järgmised seosekordajad

$$C = 1,$$
$$\eta_{y/x} = 1,$$
$$\eta_{x/y} = 0.$$

Korrelatsioonisuhe $\eta_{y/x}$ on üks seepärast, et x-tunnuse väärtus määrab üheselt y-tunnuse väärtuse. Korrelatsioonisuhe $\eta_{x/y}$ on null seepärast, et x-tunnuse tinglik keskväärtus on kummagi y-tunnuse väärtuse juures 2 ega olene y-tunnusest üldse. Nagu siit näha, ei pruugi vastupidised korrelatsioonisuhted teineteisega võrduda. Kui üks korrelatsioonisuhe on väike, aga teine suur, tähendab see, et üht tunnust saab teise väärtuse järgi hästi ennustada, teist esimese järgi aga mitte.

Regressiooniks nimetatakse funktsiooni, mis kirjeldab ühe tunnuse tingliku keskväärtuse sõltuvust teisest tunnusest:

$$\mu_{y/x} = f(x).$$

Regressiooni võib esitada tabelina, mille ühes veerus on x-tunnuse väärtused, teises y-tunnuse tinglikud keskväärtused. Eeltoodud näite järgi saaksime regressioonitabeli (tabel 14):

Tabel 14

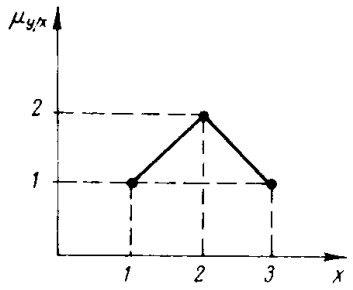
Regressioonitabel (näide)

x	$\mu_{y/x}$
1	1
2	2
3	1

Regressioonitabeli võib koostada ka x-tunnuse jaoks y-tunnuse järgi. Tavaliselt lisatakse andmetöötlusel mõlemad regressioonitabelid sagedustabelile kui täiendav lisaveerg ja lisarida. Seda teeb ka programm SAGEDUSTABEL.

Regressiooni võib kujutada graafiliselt. Eeltoodud näite korral saame graafikule kõigest 3 punkti (vt. joonis 11).

Regressiooni nimetatakse lineaarseks siis, kui graafiku punkte saab ühendada ühe sirgega. Regressioon alternatiivse tun-



Joon. 11. Regressioonigraafik (näide)

nuse järgi on alati lineaarne.

Lineaarseks korrelatsioonikordajaks ehk lühemalt korrelatsioonikordajaks tunnuste x ja y vahel nimetatakse nende tunnuste suhtelist kovariatsiooni

$$\varrho_{xy} = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y} . \qquad (47)$$

Lineaarse korrelatsioonikordaja vähim võimalik väärtus on -1 ja suurim võimalik väärtus +1. Sõltumatute suuruste vaheline korrelatsioonikordaja on parajasti 0. Kui regressioon on langev, tuleb korrelatsioonikordaja negatiivne, kui tõusev, siis positiivne.

On võimalik tõestada, et lineaarse korrelatsioonikordaja absoluutväärtus ei ületa kumbagi korrelatsioonisuhet $\eta_{y/x}$ ja $\eta_{x/y}$. Lineaarse regressiooni korral on mõlemad korrelatsioonisuhted ja korrelatsioonikordaja absoluutväärtus omavahel võrdsed. Siit tuleneb korrelatsioonikordaja üks tõlgendus: korrelatsioonikordaja ruut ϱ^2 näitab, kui suurt osa ühe tunnuse hajuvusest põhjustab lineaarne sõltuvus teisest tunnusest.

Lineaarset korrelatsioonikordajat kasutatakse kahe meetrilise tunnuse vahelise sõltuvuse tugevuse kirjeldamiseks juhul, kui võib eeldada lineaarset regressiooni või kui huvi pakub nimelt vaid sõltuvuse lineaarne komponent. Lineaarset korrelatsioonikordajat kasutatakse arvutustehnilistest kaalutlustest ajendatult sageli ka küsitava iseloomuga skaala ning järjeskaala korral. See tähendab mõõtskaala tõlgendamist ühtlase meetrilise skaalana ning tulemus on samavõrd tinglik, kuivõrd on tinglik

nimetatud tõlgendus. Kui aga üks skaaladest on nimeskaala, kaotab lineaarne korrelatsioonikordaja täielikult mõtte.

Lineaarse korrelatsioonikordaja arvutamisel pole klassiteisendus ega sagedustabeli koostamine tarvilik. Korrelatsioonikordajaid arvutavad pea kõik sõltuvuste analüüsi programmid.

Pearsoni kordaja arvutatakse sõltuvuse tugevust iseloomustava statistiku χ^2 järgi

$$P = \sqrt{\frac{\chi^2}{n + \chi^2}} . \qquad (48)$$

Kui regressioon on lineaarne, siis valimi ja sagedustabeli laiendamisel läheneb Pearsoni kordaja lineaarse korrelatsioonikordaja absoluutväärtusele. Crameri kordajal niisugust omadust ei ole. Pearsoni kordaja väärtus täieliku sõltuvuse korral pole aga täpselt üks, vaid veidi väiksem, olenevalt sagedustabeli mõõtmetest.

Pearsoni kordaja arvutab programm SAGEDUSTABEL ja teda kasutatakse Crameri kordaja asendamiseks juhul, kui mõõtskaalad on meetrilised või küsitava loomuga ning kordajat on tarvis kõrvutada lineaarse korrelatsioonikordajaga.

Normaalhinnete korrelatsioonikordaja. Järjeskaalas mõõdetud tunnuste korral pole lineaarne korrelatsioonikordaja sisuliselt tõlgendatav. Teised eelvaadeldud seosekordajad aga ei kasuta mõõtskaala järjestatuse omadust ega kirjelda sõltuvuse suunda ehk tendentsi.

Normaalhinnete korrelatsioonikordaja arvutatakse tunnuste normaalhinnete järgi täpselt samal viisil kui lineaarne korrelatsioonikordaja tunnuste originaalväärtuste järgi. Normaalhinnete korrelatsioonikordajat kasutatakse harva peaasjalikult arvutustehnilise tülikuse tõttu. Arvutamine NAIRII-2 abil on aga üsna hõlbus: kõigepealt kasutatakse programmi JÄRJETEISENDUS ja siis samu programme mis lineaarse korrelatsioonikordaja arvutamisel.

Normaalhinnete korrelatsioonikordaja omadused on lähedased lineaarse korrelatsioonikordaja omadustele: väärtused varieeruvad miinus ühest pluss üheni, sõltumatuse korral on väärtus null, negatiivne väärtus näitab langustendentsi, positiivne kasvutendentsi. Et järjed ei muutu mõõtskaala monotoonsel teisendamisel,

siis jääb muutumatuks ka normaalhinnete korrelatsioonikordaja. Seepärast kasutatakse normaalhinnete korrelatsioonikordajat ka meetrilise mõõtskaala korral, kui skaala pole uhtlane või ühtlus on küsitav.

Spearmani korrelatsioonikordaja arvutatakse järjehinnete järgi täpselt samal viisil kui normaalhinnete korrelatsioonikordaja normaalhinnete järgi. Elektronarvuti puudumise korral on Spearmani korrelatsioonikordaja eeliseks hoopis väiksem arvutustöö maht. Kuna järjehinnete jaotus on ühtlane, ei sobi Spearmani korrelatsioonikordajad normaaljaotushüpoteesile tuginevate meetoditega analüüsimiseks nii hästi kui normaalhinnete korrelatsioonikordajad. Otsesel tõlgendamisel on Spearmani korrelatsioonikordaja väga lähedane normaalhinnete korrelatsioonikordajale.

Spearmani korrelatsioonikordajat kasutatakse normaalhinnete korrelatsioonikordaja asendusena peaasjalikult siis, kui see soodustab uurimistöö tulemuste võrdlemist teiste uurimistööde tulemustega.

1.18. LINEAARNE KORRELATSIOONIANALÜÜS

Korrelatsioonianalüüsi eeltingimused. Lineaarne korrelatsioonikordaja omab teatavasti mõtet niivõrd, kuivõrd tunnuste mõõtskaalad on tõlgendatavad uhtlase meetrilise skaalana. Enamik korrelatsioonianalüüsi meetoditest tugineb normaaljaotushüpoteesile. Järjeskaala ja ebaühtlase meetrilise skaala korral on korrelatsioonianalüüs võimalik pärast eelnevat järjeteisendust, mis asendab tunnuse originaalväärtused normaalhinnetega.

Järgnevas eeldatakse kõikjal vaikides, et nimetatud tingimused on täidetud.

Korrelatsioonikordaja kriitilised väärtused. Vaatlustulemuste järgi arvutatud empiiriline korrelatsioonikordaja on tunnuste tõelist sõltuvust kirjeldava "teoreetilise" korrelatsioonikordaja hinnang. Ka siis, kui tunnused on sõltumatud, tuleb empiiriline korrelatsioonikordaja juhuse tõttu nullist erinev. Nullhüpoteesi "vaatlusalused tunnused on sõltumatud" võib kummutada alles siis, kui korrelatsioonikordaja absoluutväärtus üle-

tab kriitilise väärtuse. Kriitiline väärtus $\varrho_{p,n}$ sõltub usaldusnivoost p ja vaadeldud objektide arvust n. Korrelatsioonikordaja kriitilised väärtused on seotud Studenti suhte kriitiliste väärtustega

$$\varrho_{p,n} = \frac{t_{q,f}}{\sqrt{t^2_{q,f} + f}} , \qquad (49)$$

kus q = (1 - p)/2 on Studenti suhte kriitilise väärtuse olulisuse nivoo ühepoolse alternatiivi jaoks ning f = n-2 vabadusastmete arv (vt. [13, lk. 395]).

Korrelatsioonikordaja kriitilisi väärtusi arvutavad programmid KORRELATSIOONIANALÜÜS ja KORRELATSIOONIANALÜÜS 2. Valik kriitilisi väärtusi on esitatud tabelis 15.

Tabel 15

Korrelatsioonikordaja kriitilised väärtused (protsentides)

n	p= 90%	p= 95%	p= 99%	p= 99,9%
5	81	88	96	99,9
10	55	63	77	87
15	44	51	64	76
20	38	44	56	68
25	34	40	51	62
30	31	36	47	57
40	26	31	40	50
50	24	28	36	45
60	22	25	33	42
80	19	22	29	36
100	17	20	26	33
150	13	16	21	27
250	10	12	16	21
500	7	9	12	15
1000	5	6	8	10

Korrelatsioonikordaja usalduspiirid. Kui empiiriline korre-

latsioonikordaja pole parajasti null, siis paiknevad usalduspiirid punkthinnangu suhtes ebasümmeetriliselt. Praktika jaoks piisava täpsusega saab usalduspiire arvutada Fisheri z-teisenduse abil (vt. [5, II, lk. 85]). Korrelatsioonikordajaid koos usalduspiiridega arvutab programm LINEAARSÕLTUVUS.

Usalduspiiride abil saab kontrollida ka sõltuvuse olemasolu. Kui väärtus 0 ei mahu usalduspiiride vahele, võib sõltumatuse hüpoteesi kummutada valitud usaldusnivool.

Korrelatsioonimaatriks. n erinevast tunnusest on võimalik moodustada n^2 erinevat paari ning iga paari jaoks arvutada korrelatsioonikordaja. Kõik n^2 korrelatsioonikordajat paigutatakse ruudukujulisse tabelisse nii, et rida ja veerg näitavad korreleeritavate tunnuste järjenumbreid (vt. tabel 16). See tabel, mis vormilt meenutab turniiritabelit, kannabki korrelatsioonimaatriksi nime. Tunnuste korrelatsioonikordajad iseenesega seisavad tabeli peadiagonaalil ning võrduvad alati 100 %. Diagonaali suhtes sümmeetriliselt paiknevad arvud on võrdsed, sest korrelatsioonikordaja ei olene tunnuste järjekorrast paaris. Öeldakse, et maatriks on sümmeetriline. Kirjeldatud omaduste tõttu polegi tarvis tervet maatriksit üles kirjutada, piisab vaid peadiagonaali kohal või all seisvatest arvudest. NAIRII-2 programmides esitatakse korrelatsioonimaatriksist alumine kolmnurk (vt. tabel 17).

Tabel 16

Korrelatsioonimaatriks (näide). Korrelatsioonikordajad on avaldatud protsentides

	1	2	3	4	5
1	100	11	4	-32	28
2	11	100	-57	5	-26
3	4	-57	100	26	0
4	-32	5	26	100	54
5	28	-26	0	54	100

Tabel 17

Korrelatsioonimaatriks kolmnurksel kujul

	1	2	3	4
2	11			
3	4	-57		
4	-32	5	26	
5	28	-26	0	54

Mittenegatiivselt määratud maatriks. Tõeliste korrelatsioonikordajate maatriksil on eriline matemaatiline omadus, mida nimetatakse mittenegatiivseks määratuseks. Kui empiiriliste korrelatsioonikordajate maatriksit kavatsetakse kasutada lähtekohana edasisteks arvutusteks, peab see maatriks olema kindlasti mittenegatiivselt määratud. Seda peab tagama korrelatsioonimaatriksi koostamise meetod. Valmis korrelatsioonimaatriksi määratuse kontrollimine on väga tülikas. Ka vilunud matemaatik ei suuda ilma mahuka arvutustööta (sageli tuleb teha tuhandeid tehteid) ära tunda, kas korrelatsioonimaatriksina esitatud maatriks on mittenegatiivselt määratud või mitte.

Empiirilise korrelatsioonimaatriksi mittenegatiivset määratust võivad rikkuda arvutusvead või andmestiku hulgas esinevad küsimärgid ehk lüngad (määramata jäänud väärtused). Lünkliku andmestiku töötlemisel on mittenegatiivne määratus tagatud kahel juhul:

1) kõik objektid, millel kasvõi üks tunnus on määramata, jäetakse andmestikust välja,

2) arvutamise käigus lüngad täidetakse, näiteks võrdsustatakse määramata jäänud väärtus sama tunnuse määratud väärtuste aritmeetilise keskmisega.

Positiivne määratus on veidi rangem tingimus kui mittenegatiivne määratus. Et maatriks tuleks ka positiivselt määratud, ei tohi tunnuste hulgas olla üheaegselt kahe tunnuse summat ja mõlemaid liidetavaid või muid liitmise ja lahutamise abil teistest tunnustest tuletatud tunnuseid koos lähtetunnustega. Tuletatud tunnuste kasutamine pole keelatud, kuid siis peab osa lähtetunnuseid vaatluse alt välja jätma. Peale selle on veel nõutav, et maatriksi arvutamisel kasutatud objektide arv ületaks vaatlusaluste tunnuste arvu.

Korrelatsioonikordaja tõlgendamisel otsitakse reeglina formaalse näitaja taga seisvaid põhjuslikke seoseid. Siin peab arvesse võtma kolm võimalust:

1) esimene tunnus sõltub teisest,

2) teine tunnus sõltub esimesest,

3) mõlemad tunnused sõltuvad ühtaegu kolmandast või ka mitmest kõrvaltunnusest korraga.

Korrelatsioonikordaja ise ei võimalda teha mingit otsust ühegi versiooni kasuks.

Ühe tunnuse mõju elimineerimine. Kui on arvata, et x-tunnuse ja y-tunnuse korrelatsiooni põhjuseks on z-tunnuse ühine mõju, võib selguse saamiseks z-tunnuse mõju elimineerida. Selleks peab leidma x-tunnuse ja y-tunnuse korrelatsioonikordaja fikseeritud z väärtusel. Niisugust tinglikku korrelatsioonikordajat tähistatakse $\varrho_{xy/z}$ ja nimetatakse kahe tunnuse osakorrelatsioonikordajaks, milles kolmanda tunnuse mõju on elimineeritud. Olgu vaatluse all näiteks õpilaste matemaatilised võimed, füüsilised võimed ja vanus. Erineva vanusega õpilastest koosnevas kollektiivis osutuvad matemaatilised ja füüsilised võimed tugevasti positiivselt korreleerituiks. On alust arvata, et seda põhjustab ühine sõltuvus vanusest. Selguse saamiseks on tarvis leida matemaatiliste võimete ja füüsiliste võimete korrelatsioonikordaja ühevanuste õpilaste seas. Erineva vanusega õpilaste kollektiivi jaoks on see osakorrelatsioonikordaja, milles vanuse mõju on elimineeritud.

Osakorrelatsioonikordaja otsene määramine on tavaliselt andmete nappuse tõttu takistatud. Kui vaatlusaluses kollektiivis on sada õpilast, võib ühekuulisse vanuseintervalli sattuda vaid mõni õpilane. Juhul kui uuritavate muutumisvahemike piires võib eeldada lineaarseid regressioone, pole osakorrelatsioonikordaja otsene määramine tarvilik, ja kui see olekski võimalik, siis mitte otstarbekas. Täpsema ja kokkuvõtliku tulemuse saab arvutuslikul teel:

$$\varrho_{xy/z} = \frac{\varrho_{xy} - \varrho_{xz}\,\varrho_{yz}}{\sqrt{1-\varrho_{xz}^2}\,\sqrt{1-\varrho_{yz}^2}} . \qquad (50)$$

Esitatud valem võimaldab hõlpsalt ümber töötada ka terve korrelatsioonimaatriksi, elimineerides alati ühe ja sama tunnuse ning jättes selle lõpuks vaatluse alt üldse välja. Niisuguse arvutuse realiseerib programm ÜHE TUNNUSE MÕJU ELIMINEERIMINE.

Täielikud osakorrelatsioonikordajad. Valemit (50) võib rakendada mitu korda järjest, elimineerides iga kord korrelatsioonimaatriksist uue tunnuse. Lõpuks jääb tunnuste hulgast järele ainult kaks ning viimane osakorrelatsioonikordaja on vabastatud

kõigi esialgu vaatluse all olnud tunnuste kõrvalmõjust. Sellist osakorrelatsioonikordajat nimetame täielikuks.

Keda arvutustöö ei kohuta, see võib kirjeldatud protseduuri mitu korda läbi teha, jättes iga kord järele uue tunnusepaari. Lõpuks saame täielikud osakorrelatsioonikordajad kõigi tunnusepaaride jaoks ja nendest moodustub täielik osakorrelatsioonimaatriks. Ülesande lahendab veidi ratsionaalsemat arvutusalgoritmi kasutades programm MITMENE KORRELATSIOON.

Täieliku osakorrelatsioonikordaja tõlgendamisel ei tohi unustada, et elimineeritud on ainult nende tunnuste mõju, mis olid esialgses korrelatsioonimaatriksis arvele võetud. Arvele võtmata kõrvaltunnused võivad ikkagi põhjustada korrelatsiooni nende tunnuste vahel, mis ei ole omavahel otseses põhjuslikus seoses.

Kanooniline korrelatsioonikordaja. Tavaline korrelatsioonikordaja kirjeldab kahe üksiktunnuse seost. Nii mõnegi probleemi juures küsitakse aga, kui tugev on ühe tunnuserühma seos teise tunnuserühmaga. Näiteks võib küsida, kui tugev on seos õpilase õppeedukuse ja tervisliku seisundi vahel. Nii õppeedukust kui tervislikku seisundit kirjeldab mitu üksiktunnust.

Nimetame üht tunnuserühma vasakuks ja teist paremaks (siin peetakse silmas veergude paigutust andmetabelis) ja tähistame vasakpoolsed tunnused x_1, x_2, ..., x_v ning parempoolsed x_{v+1}, x_{v+2}, ..., x_n. Vasakpoolseteks tunnusteks võivad olla näiteks ainehinded, parempoolseteks meditsiinilised näitajad. Moodustame vasakpoolsetest tunnustest kombineeritud tunnuse

$$y = a_1x_1 + a_2x_2 + ... + a_vx_v$$

ning parempoolsetest tunnustest teise kombineeritud tunnuse

$$z = a_{v+1}x_{v+1} + a_{v+2}x_{v+2} + ... + a_nx_n .$$

Kombineeritud tunnused on korreleeritud, kusjuures korrelatsioonikordaja ϱ_{yz} oleneb tegurite a_1 ... a_n valikust. Kui valida need tegurid niiviisi, et ϱ_{yz} saaks suurima võimaliku väärtuse, olemegi leidnud vasakpoolse ja parempoolse tunnuserühma kanoonilise korrelatsioonikordaja.

Programm KANOONILINE KORRELATSIOON lähtub tavalisest kor-

relatsioonimaatriksist ning leiab sobivad tegurid $a_1 \ldots a_n$ ja kanoonilise korrelatsioonikordaja kahe tunnuserühma jaoks.

Mitmene korrelatsioonikordaja. Erijuhul võib kanoonilise korrelatsioonikordaja arvutamisel üks tunnuserühm koosneda ühestainsast tunnusest. Ühe tunnuse ja tunnuserühma korrelatsioonikordajat nimetatakse mitmeseks korrelatsioonikordajaks. Mitmene korrelatsioonikordaja ei saa olla väiksem ühestki lihtsast korrelatsioonikordajast vaatlusaluse üksiktunnuse ja rühma kuuluvate tunnuste vahel. Ühe tunnuse ja kõigi ülejäänud tunnuste mitmest korrelatsioonikordajat võib nimetada täielikuks mitmeseks korrelatsioonikordajaks. Täielik mitmene korrelatsioonikordaja iseloomustab tunnuse iseseisvust. Suure mitmese korrelatsioonikordajaga tunnuse väärtus on teiste tunnuste väärtuste järgi hästi ennustatav (vt. järgmine punkt). Väikese mitmese korrelatsioonikordajaga tunnus on suhteliselt iseseisev.

Kõigi tunnuste täielikud mitmesed korrelatsioonikordajad arvutab programm MITMENE KORRELATSIOON.

1.19. REGRESSIOONIANALÜÜS

Lineaarne regressioonivalem. Kahe juhusliku tunnuse x ja y vaheline sõltuvus tähendab seda, et ühe tunnuse väärtust saab teise tunnuse väärtuse järgi enam või vähem täpselt ennustada ehk prognoosida. Korrelatsioonikordaja ϱ_{xy} iseloomustab sõltuvuse tugevust, ei kirjelda aga sõltuvust ennast. Sõltuvust kirjeldab regressioon, mis teatavasti näitab, kuidas oleneb ühe tunnuse tinglik keskväärtus teise tunnuse väärtusest. Lineaarse regressiooni korral esitab y-tunnuse sõltuvust x-tunnusest regressioonivalem $\mu_{y/x} = ax+b$. y-tunnuse prognoosiväärtuseks y^* valitaksegi $\mu_{y/x}$. Regressioonivalem kirjutatakse tavaliselt prognoosivalemi vormis

$$y^* = ax + b \ . \tag{51}$$

Regressioonivalemi konstante saab hinnata korrelatsioonianalüüsi tulemuste järgi

$$\begin{aligned} a &= \varrho_{xy} \frac{s_y}{s_x} \ , \\ b &= \bar{y} - a\bar{x} \ . \end{aligned} \tag{52}$$

Lineaarse regressiooni korral sõltub x-tunnus y-tunnusest niisama tugevalt kui y-tunnus x-tunnusest. Vastupidise regressioonivalemi $x^* = a'y + b'$ konstandid leitakse analoogilisel viisil, valemites (52) tuleb vaid tähed x ja y omavahel vahetada.

Regressioonivalemi graafik on regressioonisirge. Kaks vastupidiste regressioonivalemite $y^* = ax+b$ ja $x^* = a'y+b'$ järgi joonistatud sirget langevad kokku vaid siis, kui $\varrho_{xy} = 1$. Vastasel korral $a' \neq 1/a$ ning sirged lõikuvad punktis, mille koordinaadid on $\bar{x}$ ja $\bar{y}$.

Lineaarse regressioonivalemi konstante saab arvutada programmidega LINEAARREGRESSIOON ja LINEAARSÕLTUVUS.

Jääkhälve. Kui mõõtmisi üldse mitte sooritada, on tunnuse y parimaks prognoosiks keskväärtus μ_y ning prognoosi ruutkeskmiseks veaks standardhälve σ_y. Pärast x-tunnuse mõõtmist saab y-tunnuse jaoks regressioonivalemi abil täpsema prognoosi. Selle prognoosi ruutkeskmist viga nimetamegi jääkhälbeks. Tõeliselt lineaarse regressiooni korral on jääkhälbeks parajasti tinglik standardhälve $\sigma_{y/x}$. Dispersioonide liitmise reegli järgi $\sigma_y^2 = \sigma_{y/x}^2 + \sigma^2(\mu_{y/x})$. Korrelatsioonisuhte definitsioonist (46) tuleneb

$$\eta_{y/x}^2 = 1 - \frac{\sigma_{y/x}^2}{\sigma_x^2} . \qquad (53)$$

Lineaarse regressiooni korral $\varrho_{xy}^2 = \eta_{y/x}^2$ ning jääkhälve

$$\sigma_{y/x} = \sigma_x \sqrt{1 - \varrho_{xy}^2} . \qquad (54)$$

Informatsiooni hulk ja korrelatsioonikordaja. Mõõteinformatsiooni hulk näitas teatavasti, kuidas suhtuvad mõõdetava suuruse hinnangute standardhälbed enne ja pärast mõõtmist. Kui me y asemel mõõdame teist tunnust x, on y väärtuse hindamisel see suhe parajasti $\sigma_y / \sigma_{y/x}$. Valemist (54) järeldub

$$\frac{\sigma_y}{\sigma_{y/x}} = \frac{1}{\sqrt{1 - \varrho_{xy}^2}} . \qquad (55)$$

Ühes tunnuses teise tunnuse kohta sisalduva informatsiooni hulk osutub üheselt seotuks korrelatsioonikordajaga. Siit tuleneb korrelatsioonikordaja üks võimalik tõlgendus: korrelatsioonikordaja (üldisemal juhul aga korrelatsioonisuhe) mõõdab ühes tunnuses teise tunnuse kohta sisalduva informatsiooni hulka.

Ebalineaarne regressioonivalem. Kahe suuruse seose kirjeldamiseks ebalineaarse valemi abil on lõpmata palju erinevaid võimalusi. Ülesande konkretiseerimisel määratakse ette nende funktsioonide hulk, mida valemis on lubatud kasutada. Lubatud funktsioone $f_1(x)$, $f_2(x)$, ..., $f_n(x)$ nimetatakse koordinaatfunktsioonideks. Regressioonivalem avaldatakse koordinaatfunktsioonide lineaarkombinatsioonina

$$y^* = c_1 f_1(x) + c_2 f_2(x) + \ldots + c_n f_n(x) \; . \qquad (56)$$

Konstandid $c_1 \ldots c_n$ peab leidma nii, et jääkhälve tuleks võimalikult väike.

Erijuhul $f_1(x) = 1$ ja $f_2(x) = x$ taandub valem (56) tavaliseks lineaarseks regressioonivalemiks.

Regressioonivalemi koostamine jaguneb kaheks alaülesandeks:

1) koordinaatfunktsioonide arvu ja kuju valimine,

2) kordajate $c_1 \ldots c_n$ määramine.

Esimese alaülesande lahendamine ei ole algoritmeeritav. Siin osutuvad määravateks sisulised kaalutlused sõltuvuse iseloomu ja koordinaatfunktsioonide lihtsuse kohta. Teine alaülesanne on lahendatav vähimruutude meetodi nime all tuntud arvutusmeetodi abil.

Praktilisi ülesandeid lahendatakse arvutiga dialoogirežiimis töötades programmi REGRESSIOONIVALEM abil. Programmi juhendist võib leida ka arvutusmeetodi täpsema kirjelduse.

Mitmene lineaarregressioon kirjeldab ühe tunnuse sõltuvust ühtaegu mitmest tunnusest. Tähistame tunnused x_1, x_2, ..., x_n ja valime prognoositavaks tunnuseks x_i. Mitmese lineaarse regressioonivalemi üldkuju prognoosivalemina kirjutatult on

$$x_i^* = c_{i1}x_1 + c_{i2}x_2 + \ldots + c_{i,i-1}x_{i-1} + c_{i,i+1}x_{i+1} + \ldots + c_{in}x_n + c_{i0} \; . \qquad (57)$$

Näiteks viie tunnusega kirjeldatud objektide korral võib kolmanda tunnuse prognoosivalemiks olla

$$x_3^* = 2{,}1x_1 - 0{,}3x_2 - 0{,}7x_4 + 1{,}6x_5 + 6{,}8.$$

Mitmese regressioonivalemi konstandid leitakse niiviisi, et prognoosi täpsus oleks maksimaalne. Tunnuse täpseim prognoos x_i^* on sama tunnuse tõelise väärtusega x_i maksimaalselt korreleeritud, mis tähendab, et x_i^* ja x_i vaheline korrelatsioonikordaja on x_i täielik mitmene korrelatsioonikordaja R_i. Valemist (54) järeldub, et prognoosi ruutkeskmine viga ehk jääkhälve on

$$\sigma^*_{x_i} = \sigma_{x_i} \sqrt{1 - R_i^2}. \qquad (58)$$

Mitmese lineaarse regressioonivalemi konstandid arvutab programm MITMENE LINEAARREGRESSIOON, mitmesed korrelatsioonikordajad ja jääkhälbed aga programm MITMENE KORRELATSIOON.

Prognoosiülesande rakendused. Prognoosiülesande primitiivseim rakendus on andmestikku jäänud lünkade täitmine prognoositud väärtustega. Lünkade täitmise võimalust ei tohi kuritarvitada. Kui prognoos põhineb sama andmestiku analüüsi tulemustel, ei lisa lünkade täitmine mingit informatsiooni, osutudes tihti mõttetuks ning isegi eksitavaks.

Tõsisemat huvi pakub tunnuste väärtuste prognoosimine ajas ette. Olgu näiteks uurimisel õpilaste kehaliste võimete ja sportlike tulemuste dünaamika. Registreerime tunnustena kehaliste ja vaimsete võimete näitajad ning sportlikud tulemused iga uurimisaluse jaoks laias vanusevahemikus. Pika aja jooksul kogutud andmete alusel on võimalik eelkirjeldatud meetodeid kasutades koostada prognoosivalemid, mis kirjeldavad maksimaalsete sportlike tulemuste sõltuvust kuni 12 aasta vanuseni mõõdetud näitajatest. Need valemid arvestavad täielikult näitajate kõikvõimalikke lineaarseid seoseid ning dünaamikat. Prognoosiväärtusega koos saame ka jääkhälbe hinnangu, mis lubab ehitada prognoositava suuruse jaoks usaldusvahemiku, kuhu tegelik väärtus satub uurija poolt valitud tõenäosusega. Prognoos võimaldab anda noortele objektiivseid soovitusi spordiala valikuks. Analoogiliselt võib prognoosida ka jõudlust õppeainetes jne.

Regressioonihinnete meetod. Eelkirjeldatud prognoosimeetod nõuab, et kõigi argumenttunnuste mõõtskaalad oleksid meetrilised. Kuidas talitada siis, kui mõni oluline argumenttunnus on mõõdetud nimeskaalas?

Programmid SEOSEANALÜÜS ja SAGEDUSTABEL võimaldavad leida regressioonitabeleid (vt. p. 1.17), mis näitavad, kuidas sõltub prognoositava tunnuse y tinglik keskväärtus $\mu_{y/x}$ üksiku nimeskaalas mõõdetud argumenttunnuse x väärtustest. Kui nüüd teisendada argumenttunnuse mõõtskaalat, võttes uuteks väärtusteks prognoositava tunnuse tinglikud keskväärtused $\mu_{y/x}$, osutub vaatlusalune osaregressioon lineaarseks. Uus skaala ei olene üldse vanast. Tulemust ei riku uue skaala suvaline lineaarteisendus. Valime kaks konstanti a, b ning nimetame arvud

$$x' = a\mu_{y/x} + b \qquad (59)$$

argumenttunnuse x regressioonihinneteks prognoositava tunnuse y suhtes.

Eeltoodud arutluse abil saab püstitatud küsimusele lihtsaima vastuse: nimeskaalas mõõdetud argumenttunnuste mõõtskaalad tuleb metriseerida, asendades algsed väärtused regressioonihinnetega prognoositava tunnuse suhtes. Samaviisi talitatakse ka järjeskaalas mõõdetud argumenttunnustega ja meetrilises skaalas mõõdetud argumenttunnustega juhul, kui vastav osaregressioon on oluliselt ebalineaarne.

1.20. TUNNUSEHULGA STRUKTUUR

Suurima korrelatsiooni tee. Tunnuste seoseid kirjeldava korrelatsioonimaatriksi kõiki elemente pole nende suure hulga tõttu praktiliselt võimalik ühekaupa interpreteerida, seepärast kontsentreeritakse uurimistöös tähelepanu vähesele arvule valitud seostele. Osa seoseid valitakse erilise tähelepanu objektiks sisulistel kaalutlustel, osa aga vaatlustulemuste analüüsimisel selgunud seosetugevuse järgi. Üldise objektiivse ülevaate saamiseks korrelatsioonimaatriksist tugevamate seoste kaudu kasutatakse suurima korrelatsiooni teed.

Suurima korrelatsiooni tee moodustamisel võib lähtuda tavalisest korrelatsioonimaatriksist, ühe tunnuse mõjust vabast osakorrelatsioonimaatriksist või ka täielikust osakorrelatsioonimaatriksist. Lähtemaatriksi valikust oleneb tulemuste tõlgendamine. Kõige ilmekama ja lihtsalt tõlgendatava pildi tunnuste seostest annab täieliku osakorrelatsioonimaatriksi järgi moodustatud suurima korrelatsiooni tee.

Suurima korrelatsiooni tee on kõiki tunnuseid ühendav, sageli hargnev, kuid silmusteta ahel, mille lülid ühendavad kõige tugevamini seotud tunnuseid. Seose tugevuse mõõduks võetakse korrelatsioonikordaja absoluutväärtus.

Suurima korrelatsiooni tee koostamise algul otsitakse korrelatsioonimaatriksi järgi kõige tugevamini seotud tunnusepaar. Need tunnused arvatakse esimestena korrelatsiooniahelasse. Järgnevalt lisatakse ahelale ühekaupa uusi tunnuseid, kuni kõik nad on ahelas. Seejuures jagatakse tunnused igal sammul kahte rühma: ahelasse võetud ja ahelasse võtmata tunnusteks. Järgmise tunnuse valimiseks otsitakse kõigi selliste tunnusepaaride hulgast, kus üks tunnus kuulub ühte ja teine teise rühma, kõige tugevamini seotud paar ja lisatakse ahelale selle paari uus tunnus.

Suurima korrelatsiooni tee tabeli koostab programm KORRELATSIOONITEE. Tabeli järgi joonise tegemisel on heaks töövahendiks nummerdatud nuppude või pappkettakeste komplekt. Joonis on soovitatav teha nii, et ahela hargnemata osad oleksid sirged. Korrelatsioonitee tabeli ja joonise näite võib leida lisas 34.

<u>Tunnusehulga lineaarteisendus</u>. Vaatleme objekte, mida kirjeldatakse n meetrilise tunnuse x_1, x_2, ..., x_n väärtustega ja eeldame, et tunnuste väärtuste summasid või vahesid on võimalik sisuliselt tõlgendada. Paljud autorid väidavad, et liita ja lahutada tohib ainult samades mõõtühikutes mõõdetud tunnuseid. Kuna aga juttu tuleb vaid tunnuste arvväärtuste liitmisest ja lahutamisest, osutub nimetatud kitsendus ülearuseks. Näiteks keha täidluse iseloomustamisel lahutatakse inimese kaalu arvväärtusest inimese pikkuse arvväärtus, olgugi et kaalu ja pikkust mõõdetakse erinevates mõõtuhikutes.

Tunnuste lineaarkombinatsioonideks nimetati teatavasti suurusi $y_i = \sum_{j=1}^{n} a_{ij}x_j + b_i$, kus a_{ij} on kombinatsiooni kordajad ja

b_i kombinatsiooni vabaliige. Näiteks kolme tunnuse lineaarkombinatsioon võib olla $y_2 = 3,7x_1 - 2,3x_2 + 0,8x_3$. Selles näites on $a_{2,1} = 3,7$, $a_{2,2} = -2,3$, $a_{2,3} = 0,8$ ja $b_2 = 0$. Lineaarkombinatsioone võib käsitada kui uusi, originaaltunnustest tuletatud tunnuseid.

Teatud tingimuste täidetuse korral on tuletatud tunnuste $y_1, y_2, \ldots, y_m$ järgi võimalik tagasi arvutada kõiki originaaltunnuseid kui tuletatud tunnuste lineaarkombinatsioone. Siis öeldakse, et üleminek objektide kirjeldamiselt originaaltunnuste x abil kirjeldamisele tuletatud tunnuste y abil on tunnusehulga lineaarteisendus, tagasipöördumine kirjeldamisele originaaltunnuste abil aga eelmise teisenduse pöördteisendus.

Tunnusehulga lineaarteisendus pakub huvi siis, kui tuletatud tunnuseid saab otseselt tõlgendada. Olgu näiteks originaaltunnusteks matemaatikahinne x_1 ja keelehinne x_2. Tuletatud tunnust $y_1 = x_1+x_2$ võib tõlgendada kui üldtaset ja tunnust $y_2=x_1-x_2$ kui kalduvust. Pöördteisendus on siin $x_1 = (y_1+y_2)/2$ ning $x_2 = (y_1-y_2)/2$.

Ortogonaalpööre. Valime kahe objekti vahelise erinevuse mõõduks eukleidilise kauguse

$$l_x = \sqrt{(x_1'-x_1'')^2 + (x_2'-x_2'')^2 + \ldots + (x_n'-x_n'')^2} . \qquad (60)$$

x_i' ja x_i'' tähistavad i-nda originaaltunnuse väärtust ühel ja teisel objektil. Samade objektide vahekaugus tuletatud tunnuste järgi oleks

$$l_y = \sqrt{(y_1'-y_1'')^2 + (y_2'-y_2'')^2 + \ldots + (y_n'-y_n'')^2} . \qquad (61)$$

Lineaarteisendust, mille vabaliikmed b_i on nullid, nimetatakse tunnusehulga pöördeks. Kui pööre jätab kõikvõimalike objektipaaride kaugused muutumatuks ($l_y=l_x$), öeldakse, et pööre on ortonormeeritud ehk lühemalt ortogonaalpööre. Ortogonaalpööret võib lugeda kõige väiksema "vägivallaga" seotud lineaarteisenduseks. Eelmise alapunkti lõpul näitena esitatud teisendus ei ole ortogonaalpööre. Ortogonaalpöördeks osutub aga teisendus

$$y_1 = \frac{\sqrt{2}}{2} x_1 + \frac{\sqrt{2}}{2} x_2 ,$$

$$y_2 = \frac{\sqrt{2}}{2} x_1 - \frac{\sqrt{2}}{2} x_2 \quad .$$

Üldjuhul õige keeruline pöördteisenduse leidmise ülesanne osutub ortogonaalpöörde korral üllatavalt lihtsaks: teisenduse

$$y_i = \sum_{j=1}^{n} a_{ij} x_j \qquad (62)$$

pöördteisendus on

$$x_j = \sum_{i=1}^{n} a_{ij} y_i \ , \qquad (63)$$

arvud a_{ij} võetakse samast tabelist, kust päriteisenduse kordajad.

Tunnusehulga lihtsuse printsiibid. Tunnusehulka võib nimetada maksimaalselt lihtsaks siis, kui

1) kõik tunnused on iseseisvad, s.t. ei korreleeru;

2) objekti piisavaks iseloomustamiseks tarvisminevate tunnuste arv on minimaalne.

Tsentreeritud tunnuseks nimetatakse tunnuse ja selle keskväärtuse vahet $x - \mu_x$. Tsentreeritud tunnuse väärtusteks on tsentreerimata tunnuse hälbed.

Tsentreerimine tähendab vaid mõõtskaala alguspunkti nihutamist ega muuda tunnuse sisulist tähendust.

Komponentanalüüs. Võrdleme mitmesuguseid tunnusehulki, mis saadakse tsentreeritud originaaltunnustest erinevate ortogonaalpöörete teel, ja nimetame lihtsaimaks ehk loomulikuks tunnusehulgaks seda tunnusehulka, mis kõikvõimalike ortogonaalpöörete abil tuletatavate tunnusehulkade seas kõige paremini rahuldab eelkirjeldatud lihtsuse printsiipe. Loomulik tunnusehulk võimaldab objekte kirjeldada matemaatiliselt kõige lihtsamal viisil. Võib arvata, et loomuliku tunnusehulga tunnused kirjeldavad vahetult objekti sisemisi omadusi.

Komponentanalüüs on matemaatiline meetod loomuliku tunnusehulga leidmiseks. Uusi tunnuseid nimetatakse peakomponentideks. Esimene peakomponent valitakse niiviisi, et see võimaldaks maksimaalselt täielikult kirjeldada objektidevahelisi erinevusi. Erinevuse mõõduks loetakse eukleidilist kaugust (60). Teine pea-

komponent valitakse esimese peakomponendiga korreleerimata lineaarkombinatsioonide hulgast niiviisi, et see võimaldaks maksimaalselt täielikult kirjeldada neid erinevusi, mis jääksid ainult esimese peakomponendi kasutamisel kirjeldamata. Järk-järgult leitakse kokku n peakomponenti, millest igaüks on sõltumatu kõigist eelnevatest ning kirjeldab maksimaalselt täielikult ainult eelmiste peakomponentide kasutamisel kirjeldamata jäänud erinevusi.

Komponentanalüüsi tulemuseks on ortogonaalpöörde kordajate a_{ij} tabel. Originaaltunnused x avalduvad peakomponentide g kaudu järgmiselt

$$x_i = \sum_{j=1}^{n} a_{ij} g_j + \mu_i \qquad (64)$$

ja peakomponendid originaaltunnuste kaudu

$$g_j = \sum_{i=1}^{n} a_{ij} (x_i - \mu_i). \qquad (65)$$

Valemite (64, 65) kirjutamisel on eeldatud, et indeks i numereerib originaaltunnuseid, j peakomponente ning kordajad on tabelisse paigutatud niiviisi, nagu seda teeb programm KOMPONENTANALÜÜS (originaaltunnustele vastavad read ja peakomponentidele veerud). Peakomponentide arv on alati niisama suur kui tunnuste arv.

Peakomponendi olulisuse mõõduks on komponendi dispersioon. Komponendi dispersiooni suhet kõigi komponentide dispersioonide summasse võib nimetada komponendi osatähtsuseks. Esimeste peakomponentide osatähtsuste summa näitab, millisel määral on objektidevahelised erinevused kirjeldatavad ainult nende komponentide vahendusel. Osatähtsuste ja summeeritud osatähtsuste tabel kuulub komponentanalüüsi tulemuste hulka.

Komponentanalüüsi kasutamisel on sõlmprobleem formaalsete arvutuste teel leitud peakomponentide sisuline tõlgendamine. Tavaliselt tõlgendatakse vaid esimesi peakomponente. Viimastel peakomponentidel puudub objektide diferentseerimise võime ning neid võib tõlgendada vaid kui invariante.

Peakomponentide individuaalsete väärtuste arvutamiseks saab kasutada programmi OBJEKTIDE ANALÜÜS ehk programmi TOIMETAJA.

Standardiseeritud tunnused. Standardiseerimisel tunnus

tsentreeritakse ning siis valitakse uus mõõtühik niiviisi, et standardhälve saaks väärtuseks üks. Kui tähistada standardiseerimata tunnus x ja standardiseeritud tunnus z, siis

$$z = \frac{x - \mu_x}{\sigma_x}, \qquad x = \mu_x - \sigma_x z . \tag{66}$$

Standardiseerimist kasutatakse erinevate tunnuste mõõtskaalade ühtlustamiseks.

Komponentanalüüsi ülesannetes võib lähtuda nii standardiseerimata kui standardiseeritud tunnustest. Kumba valida, oleneb sellest, kummal juhul valemi (60) järgi arvutatud objektide kaugused vastavad paremini sisulistele kaalutlustele.

Lineaarne faktormudel. Traditsioonikohaselt kasutatakse faktormudelis standardiseeritud tunnuseid $z_1 \ldots z_n$. Kui tunnused ei osutu sõltumatuteks, tekib küsimus sõltuvuse mehanismist. Teatavasti võis kahe tunnuse korreleerituse põhjuseks olla hoopiski kolmas tunnus, mis ei pruugi olla otseselt mõõdetav. Faktormudel rajaneb hüpoteesil, et objektid on põhimõtteliselt kirjeldatavad k varjatud tunnusega $f_1, f_2, \ldots, f_k$, mida nimetatakse ühisfaktoriteks, lühemalt - faktoriteks. Faktorid kirjeldavad igaüks üht objekti sisemist, otsesele mõõtmisele kättesaamatut omadust, ning mõõdetavad tunnused on vaid faktorite funktsioonid. Faktorite arv pole tavaliselt kuigi suur, ulatudes vaid üksikute rakenduste korral kaheksani. Üldkasutatavat kokkulepet jälgides eeldame, et faktorid on samal viisil standardiseeritud kui tunnused. Teiseks eeldame, et faktorid on teineteisest sõltumatud (sõltuvate faktoritega mudeleid kasutatakse harva).

Lineaarse faktormudeli matemaatiline väljendus on võrrand tunnuste arvutamiseks faktorite järgi

$$z_i = \sum_{j=1}^{k} a_{ij} f_j + \Delta_i . \tag{67}$$

Kordaja a_{ij} kirjeldab j-nda faktori osa i-ndas tunnuses ja seda nimetatakse faktori koormuseks tunnusele. Ühtaegu on a_{ij} võrdne i-nda tunnuse ja j-nda faktori korrelatsioonikordajaga. Δ_i on mõõtmisvigadest ja tunnuse spetsiifikast tingitud juhuslik jääk-

liige. Mida väiksemad on mõõtmisvead ja mida adekvaatsem mudel, seda väiksemad tulevad jäägid Δ .

Tunnuse kommunaliteet

$$h_i^2 = \sum_{j=1}^{k} a_{ij}^2 \qquad (68)$$

näitab, kui suur osa i-nda tunnuse dispersioonist on seletatav faktorite väärtuste varieerumisega. Ülejäänud osa $1 - h_i^2$ põhjuseks on mõõtmisvead ja faktorite mõjuga seletamatu spetsiifika.

Faktorile vastav dispersioon

$$s_j^2 = \sum_{i=1}^{n} a_{ij}^2 \qquad (69)$$

näitab, mitme keskmise tunnusega võrdväärselt suudab vaadeldav faktor iseloomustada objekti.

Faktormudel seletab tunnuste korrelatsioonikordajaid järgmiselt:

$$\varrho_{il} = \sum_{j=1}^{k} a_{ij}a_{lj} \cdot \qquad (70)$$

Kuna mudelit ei õnnestu täpselt koostada, erinevad niiviisi arvutatud väärtused veidi vaatlusandmete järgi hinnatud väärtustest. Erinevused iseloomustavad konkreetse faktormudeli täpsust ja peaksid rahuldava mudeli korral umbes p-protsendise tõenäosusega jääma p-protsendisel usaldusnivool määratud usalduspiiridesse.

Faktorite pööramine. Täpselt niisama hästi kui mingi konkreetne faktorihulk $f_1 \ldots f_n$, sobib tunnuste formaalseks seletamiseks ükskõik milline teine faktorihulk, mis on saadud esimesest ortogonaalpöörde teel. Öeldakse, et tunnuste seletamise nõue määrab faktorid vaid suvalise ortogonaalpöörde täpsuseni. Faktorihulga teisendamist teiseks võrdväärseks faktorihulgaks nimetatakse faktorite pööramiseks. Kuidas faktoreid pöörata, seda otsustavad vaid faktorite tõlgendatavuse kaalutlused. Üldisema iseloomuga kaalutlustest väärivad erilist tähelepanu Kelley ja Hotellingi maksimumdispersiooni printsiip ning Thurstone'i lihtstruktuuri printsiip.

Peafaktoriteks nimetatakse maksimumdispersiooni printsiipi

rahuldavaid faktoreid. Peafaktorid määratakse niiviisi, et igale järgmisele faktorile vastab maksimaalne võimalik dispersioon (valem 69). Objektide kirjeldamisel on kõige olulisem esimene peafaktor, järgnevate faktorite tähtsus langeb. Pedagoogikas on esimest peafaktorit sageli võimalik tõlgendada kui õpilase üldtaset, järgmisi faktoreid aga kui kalduvusi teineteisest sõltumatute diferentseerimisviiside järgi.

Peafaktorite kasutamisel on kõige hõlpsam teha otsust sobiva faktorite arvu kohta. Sageli valitakse peafaktoreid parajasti niipalju, et viimane neist iseloomustaks objekti võrdväärselt vähemalt ühe keskmise tunnusega ($s_j^2 \geqslant 1$).

Esimeses lähenduses võib peafaktoriteks valida standardiseeritud esimesed peakomponendid. Selle lähendusega piirdutaksegi üldist tarvitamist leidnud IBM statistikaprogrammide süsteemis (meil tarvitusel ühtsusseeria arvutitel) ja ka NAIRII-2 programmis FAKTORANALÜÜS.

Veel jämedamaks lähenduseks peafaktoritele on tsentroidimeetodi abil leitavad faktorid. Tsentroidimeetod (vt. [16]) oli omal ajal populaarne tänu arvutustehnilisele lihtsusele.

Varimaks-faktoriteks nimetatakse lihtstruktuuri printsiipi (täpsemini, selle printsiibi üht formaalset ekvivalenti) rahuldavaid faktoreid. Psühholoogiaalastes uurimistöödes populaarne lihtstruktuuri printsiip nõuab, et iga üksiktunnus oleks tugevalt seotud võimalikult väheste faktoritega. Varimaks-faktoreid saab leida programmi FAKTORANALÜÜS abil.

Faktorite individuaalväärtuste arvutamiseks on tarvilik teisenduse (67) täpseima pöördteisenduse

$$f_j \approx \sum_{i=1}^{n} b_{ji} z_i \qquad (71)$$

kordajate b_{ji} tabel. Päris täpset pöördteisendust teisendusel (67) ei ole. Varimaks-faktorite jaoks leiab kordajate b tabeli vähimruutude meetodi abil programm FAKTORANALÜÜS. Peafaktorite arvutamiseks tarvilikke kordajaid aitab leida programm KOMPONENTANALÜÜS. Faktorite individuaalväärtused tabuleeritakse hiljem programmi OBJEKTIDE ANALÜÜS abil.

1.21. OBJEKTIHULGA STRUKTUUR

Kauguste maatriks. Kahe objekti vaheline kaugus arvutatakse tavaliselt valemi (60) järgi. Kui objekte on mitu, võib arvutada kauguse iga objektipaari jaoks. Kõikvõimalikud kaugused kirjutatakse ruudukujulisse tabelisse, kus rida vastab paari esimesele objektile, veerg teisele objektile, lõikumiskohta aga kirjutatakse paarisisene kaugus. See tabel ongi kauguste maatriks. Vormilt on kauguste maatriks üsna sarnane korrelatsioonimaatriksiga. Tunnustevaheliste suhete asemel kirjeldatakse siin objektidevahelisi suhteid. Kui korrelatsioonikordaja oli läheduse mõõt, siis kauguste maatriksi elemendid on erinevuse mõõduks. Peadiagonaalil seisvad arvud näitavad kaugusi objektist iseeneseni ja on nullid. Erinevalt korrelatsioonikordajast ei ole kaugusel ülemmäära. Peadiagonaali suhtes sümmeetriliselt paiknevad arvud on kauguste maatriksis teineteisega võrdsed samal põhjusel kui korrelatsioonimaatriksiski, seepärast võib ka kauguste maatriksi esitamisel piirduda vaid alumise kolmnurgaga.

Kauguste maatriksi koostab programm OBJEKTIDE KAUGUSED.

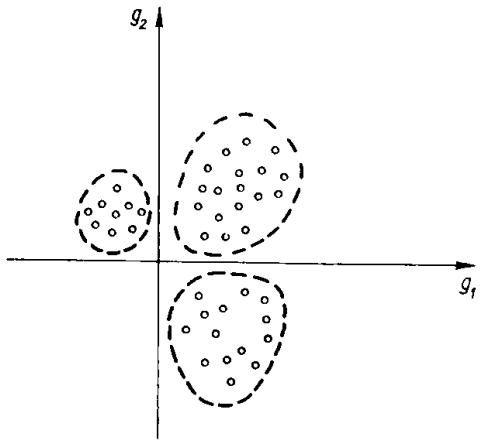
Vähima kauguse tee ühendab kõik objektid selliselt, et vahetult ühendatud objektide kaugused on võimalikult väikesed. Vähima kauguse tee tabeli ja joonise koostamine on analoogiline suurima korrelatsiooni tee tabeli ja joonise koostamisega. Ülesande lahendab programm OBJEKTIDE KAUGUSED.

Vähima kauguse teed võib kasutada objektide rühmitamiseks. Rühmad eraldatakse vähima kauguse tee nõrgemate lülide (suuremad kaugused) katkestamise teel, ühte rühma jäävad teineteisele lähedased objektid.

Rühmitamine peakomponentide järgi. Vähima kauguse tee abil saab NAIRII-2 kasutaja rühmitada kuni 50-liikmelisi objektihulki. Suuremate, aga sageli ka väikeste hulkade korral, alustatakse rühmitamist tunnusehulga lihtsustamisest. Programmi KOMPONENT-ANALÜÜS abil leitakse kahe esimese peakomponendi kordajad ja siis tabuleeritakse need komponendid programmi OBJEKTIDE ANALÜÜS või TOIMETAJA abil kõigi objektide jaoks. Iga objekti järgi kantakse teljestikuga varustatud joonisele väike ringike, mille sees on objekti number ja mille koordinaadid vastavad peakomponentide

väärtustele. Kui objekte on palju, võib objekti numbrite märkimisest ka loobuda. Niiviisi koostatud hajumisvälja näide on joonisel 12. Sellega on rühmitamisülesande matemaatiline osa lõpetatud. Objektihulga rühmadesse jaotamine jääb uurija enese hooleks. Joonisel 12 on punktiiri abil eraldatud kolm rühma.

On olemas ka täielikult formaliseeritud rühmitamiseeskirju, kuid praktilise efektiivsuse poolest ei ületa ükski nendest eelkirjeldatud poolintuitiivset meetodit.



Joon. 12. Peakomponentide hajumisväli.

Transponeerimine tähendab andmetabeli sellist ümberkorraldamist, mille järel varem ühes reas paiknenud arvud (erinevate tunnuste väärtused ühe objekti jaoks) satuvad ühte veergu ning varem ühes reas paiknenud arvud (ühe tunnuse väärtused erinevate objektide jaoks) satuvad ühte ritta. Tavalise andmetabeliga võrreldes on transponeeritud andmetabel "küljeli pööratud".

Transponeerimiseks pole tarvis aritmeetilisi arvutusi, küll aga peab andmestiku uuesti perforeerima. Kui andmestik pole liiga suur, saab käsitsi perforeerimist asendada automaatse perforeerimisega programmi TRANSPONEERIMINE abil.

Q-tehnika korrelatsioonianalüüs. Kahe objekti sarnasust võib iseloomustada kordajaga, mis arvutatakse kõigi tunnuste väärtuste järgi täpselt samaviisi, kui kahe tunnuse lähedust iseloomustav korrelatsioonikordaja arvutatakse kõigi objektide järgi. Niiviisi leitud sarnasuse mõõtu nimetatakse Q-tehnika korrelatsioonikordajaks. Q-tehnika korrelatsioonikordaja ei ole ühe-

selt seotud eukleidilise kaugusega. Q-korrelatsioon võib olla 100 % ka siis, kui kaugus on suur, ent ühe objekti kõiki tunnuseid saab arvutada teise objekti tunnuste järgi lineaarse teisendusvalemi abil.

Q-tehnika korrelatsioonianalüüsi korral kasutatakse kauguste maatriksi asemel Q-tehnika korrelatsioonikordajate maatriksit, mis vormilt ja matemaatilistelt omadustelt on õige sarnane tavalise korrelatsioonimaatriksiga. Q-tehnika korrelatsioonimaatriksi koostamiseks transponeeritakse andmestik ja rakendatakse seejärel tavalisi korrelatsioonianalüüsi programme. Objektihulga struktuuri uurimisel kasutatakse formaalselt samu meetodeid ja programme (KORRELATSIOONITEE, FAKTORANALUUS) mis tavaliselt tunnusehulga struktuuri uurimisel.

Kui on tarvis rõhutada erinevust Q-tehnikast, nimetatakse tavalist tunnuste korrelatsioonianalüüsi R-tehnika korrelatsioonianalüüsiks.

Q-tehnika korrelatsioonianalüüsi peab käsitama kui teatud kunstlikku võtet, mida võivad õigustada vaid konkreetsed sisulised kaalutlused. Põgusa kirjelduse Q-tehnika ja siin nimetamata P-tehnika ning O-tehnika korrelatsioonianalüüsi kasutamisest psühholoogias võib leida raamatust [16, lk. 139-149].

Tunnustevahelise sõltuvuse arvestamine kauguse määramisel. Tavaline kauguse valem (59) ignoreerib tunnustevahelist sõltuvust ja on tingimusteta kasutatav vaid sõltumatute tunnuste korral. Tunnuste sõltuvus mõjutab intuitiivset kauguse mõistet, milles võib veenduda näite varal. Vaatleme kahe sõltumatu tunnuse x_1 ja x_2 abil kirjeldatud objekte. Kaugus avaldub $\sqrt{(x_1' - x_1'')^2 + (x_2' - x_2'')^2}$. Defineerime kolmanda tunnuse valemiga $x_3 = 2x_2$. See tunnus on 100 % korreleeritud teise tunnusega ega lisa objekti kirjeldusele midagi uut. Ometi muudab kolmanda tunnuse lisamine oluliselt valemi (60) järgi arvutatud kaugusi ja nende suhteid. Ilmselt oleks õigem kolmandat tunnust lihtsalt ignoreerida. Probleem muutub aga keeruliseks, kui korrelatsioon pole 100 %, vaid 99 % või 80 %.

Tõenäosusteoorias näidatakse, et normaaljaotuse korral eksisteerib kõiki sõltuvusi arvesse võttev ratsionaalne kauguse mõõt, mis sõltumatute suuruste korral taandub valemi (60) järgi

arvutatud kauguseks standardiseeritud tunnuste jaoks. See kauguse mõõt leiab NAIRII-2 statistikaprogrammide süsteemis rakendamist ainult programmis OBJEKTIDE ANALÜÜS. Laiemat kasutamist takistab sõltuvuste arvesse võtmisega kaasnev mahukas arvutustöö.

K I R J A N D U S

1. Mereste, U. Statistika üldteooria. Tallinn, 1975, 496 lk.

Majandusüliõpilastele adresseeritud õpik, milles klassikalise statistika probleeme käsitletakse üksikasjalikult, matemaatilise statistika meetodid jäävad aga tagaplaanile. Raamat sisaldab rikkalikult näiteid ega nõua lugejalt matemaatilist ettevalmistust.

2. Noorma, R., Kaasik, Ü. Elektronarvuti "Nairi-2". - "Programme kõigile", 1972, nr. 3, 102 lk. (Tartu Riiklik Ülikool.)

Arvuti NAIRII-2 matemaatiliste omaduste kirjeldus ning programmeerimisjuhend. Lugemisel on tarvis algteadmisi elektronarvutitest ja programmeerimisest.

3. Petersen, I. Katsete planeerimine. Tallinn, 1966, 90 lk.

Konspektiivne sissejuhatus pedagoogikauurimistöö jaoks suure väärtusega matemaatilisse teooriasse. Matemaatiline ettevalmistus hõlbustab lugemist, tõsise huvi ja püsivuse korral saab aga esitatavast aru ka humanitaarharidusega uurija.

4. Tiit, E. Matemaatiline statistika I. Tartu, 1971, 298 lk. + 17 tabelit. (Tartu Riiklik Ülikool.)

Raamat sobib matemaatilise statistika iseseisvaks õppimiseks. Siin käsitletakse ka tarvilikke tõenäosusteooria mõisteid. Matemaatilist eelharidust lugejalt ei nõuta.

5. Tiit, E. Matemaatilise statistika tabelid. Tartu, I osa 1971, 223 lk., II osa 1972, 252 lk. (Tartu Riiklik Ülikool.)

Suurema osa raamatust täidavad statistikameetodite kirjeldused, arvutusjuhendid ja rakendusnäited. Lugeja peab tundma matemaatilise statistika algmõisteid, mille osas pole matemaatiline ettevalmistus nõutav. Arvutusjuhendid ja näited on orienteeritud lugejale, kellel pole võimalik kasutada elektronarvutit.

6. Koppel, E., Tiit, E. Sotsioloogilise uurimise statistikast I. - "Matemaatika ja kaasaeg", 1972, nr. 18, lk. 3-12.

7. Tiit, E. Sotsioloogilise uurimise statistikast II. - "Matemaatika ja kaasaeg", 1973, nr. 19, lk. 48-69.

Kahest viimasest artiklist võib leida matemaatilise statistika ülesannete ja algmõistete kirjelduse matemaatilise ettevalmistuseta lugejale kõige kergemini mõistetavas vormis.

8. Tiit, E. Statistilise andmetöötluse üldprintsiibid. - "Programme kõigile", 1975, nr. 9, lk. 3-27. (Tartu Riiklik Ülikool.)

Sissejuhatus ja näpunäited TRÜ Arvutuskeskuse statistilise andmetöötlussüsteemi (arvuti MINSK-32 jaoks) kasutajale.

9. Clauss, G., Ebner, H. Grundlagen der Statistik für Psychologen, Pädagogen und Soziologen. Berlin, 1970, 367 S.

Lühikese aja jooksul populaarseks saanud käsiraamat. Lõviosa raamatust on pühendatud võrdlusülesannetele. Kirjeldatakse mitmeid huvipakkuvaid meetodeid, mis NAIRII-2 programmisüsteemis on jäänud realiseerimata. Regressioonianalüüsi ja mitmemõõtmelise analüüsi (faktoranalüüs jne.) käsitlus puudub.

10. Göttner, R., Fischer, P., Krieg, R. Was ist - was kann Statistik? Leipzig, Jena, Berlin, 1975, 255 S.

Tähelepanuväärselt õnnestunud populaarne raamat esmaseks tutvumiseks statistikaga.

11. Liiv, H. Instructional effectiveness of programmed learning of English tense forms I: methodology of a pedagogical experiment. - "Linguistica", nr. 3, p. 125-141. (Tartu Riiklik Ülikool.)

Artikkel sisaldab kaherühmaeksperimendis rühmade algtasemete erinevust nivelleeriva töötlusmeetodi kirjeldust koos rakendusnäidetega.

12. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. Москва, 1968, 474 с.

Täiuslik statistikatabelite kogu. Kolmandiku raamatust võtavad enda alla statistikameetodite tarvitamisjuhendid, mille kasutamine eeldab aga head matemaatilist ettevalmistust.

13. Кендалл М.Дж., Стьюарт А. Статистические выводы и связи. Москва, 1973, 899 с.

Põhjalik statistikakäsiraamat. Eeldatakse lugeja head mate-

maatilist ettevalmistust.

14. Кыверялг А.А. Вопросы методики педагогических исследований. Таллин, 1971, 134+227 с.

Käsiraamat algajale uurijale. Kirjeldatakse ka lihtsamaid statistikameetodeid, orienteerudes lugejale, kel puudub elektronarvuti kasutamise võimalus.

15. Налимов В.В. Теория эксперимента. Москва, 1971, 208 с.

Autor käsitleb matemaatilist statistikat kui üldist eksperimendi teooriat. Raamatust võib leida lühiülevaate paljudest statistikameetoditest. Huvipakkuvaim osa on eksperimendi ja statistilise töötluse metodoloogiliste kontseptsioonide analüüs. Raamatu lugemiseks piisab keskpärasest matemaatilisest ettevalmistusest.

16. Окунь Я. Факторный анализ. Москва, 1974, 200 с.

Valdav osa raamatust on pühendatud faktoranalüüsi tsentroidimeetodi tehnikale.

17. Сборник научных программ на Фортране, выпуск 1, статистика. Москва, 1974, 316 с.

Süsteemi IBM/360 statistikaprogrammide tekstid ja kirjeldused programmeerimist tundva lugeja jaoks.

18. Смирнов Н.В., Дунин-Барковский И.В. Курс теории вероятностей и математической статистики для технических приложений. Москва, 1969, 512 с.

Tehnikaüliõpilastele adresseeritud õpik, mis aga tänu üldisele käsituslaadile on kasutatav sõltumatult lugeja erialast.

19. Суходольский Г.В. Основы математической статистики для психологов. Ленинград, 1972, 430 с.

Pedagoogikauurija jaoks peaaegu kõige sobivamalt orienteeritud põhjalik statistikaõpik. Lugejalt ei nõuta matemaatilist eelharidust, küll aga visa ja tõsist tööd. Claussi ja Ebneri raamatuga võrreldes on ainevalik laiem.

20. Харман Г. Современный факторный анализ. Москва, 1972, 486 с.

Tunnustatuim praktilise kallakuga faktoranalüüsi käsiraamat. Sisaldab häid arvutusnäiteid. Lugejalt nõutakse keskpärast matemaatilist ettevalmistust.

Kaane kujundanud V. Smirnov

Ханнес Феликсович Т а м м е т. Статистические методы при пользовании ЭВМ НАИРИИ - 2. На эстонском языке. Художник-оформитель В. Смирнов. Издательство "Валгус", Таллин.

Toimetaja H. Heinoja. Kunstiline toimetaja O. Herodes. Trükkida antud 8.VII 1976. Paber 60x84/16. Trükipoognaid 14,5. Tingtrükipoognaid 13,49. Arvestuspoognaid 11,37. Trükiarv 500. MB- 07314. Kirjastus "Valgus", Tallinn, Pärnu mnt. 10. Eksperimentaalkombinaat "Bit" rotaprint, Tallinn, Pikk 68.

Tell.nr. 1822-2230. Hind 1.14

T 20204 - 259
M 902(16) - 76