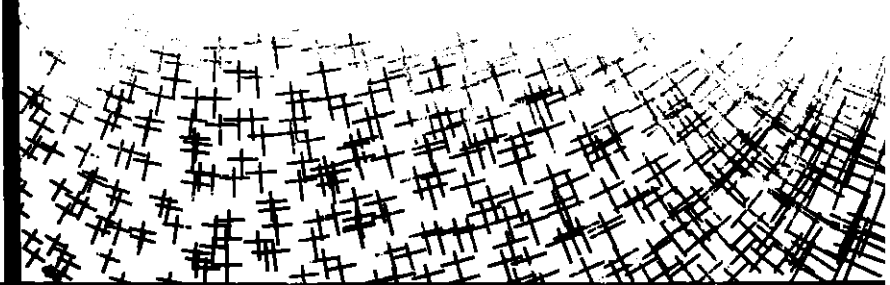
Rakendusstatistika algkursus



Ene-Margit Tilt, Märt Möls

Rakendusstatistika lühikursus

RAKENDUSSTATISTIKA LÜHIKURSUS

Ene-Margit Tiit

Märt Möls

Tartu, 1997

Retsenseerinud:
Dots. Anne-Mai Parring,
Lektor Ene Käärik,
Lektor Martin Viil

Keeletoimetaja:
Tiina Viil

Kaane kujundanud:
Michael Walsh

SISUKORD

SISSEJUHATUS

Käesoleva lühikursuse aluseks on 1996. aasta oktoobris Eesti Statistikaametis peetud loengutesari, sealt pärinevad ka raamatus esitatavad näited. Osa näiteid on laenatud ka selle kursuse kuulajate kodutöödest. Raamatu autorid on nende eest väga tänulikud. Üldse on näited käesolevas kursuses väga olulised, nende varal esitatakse suur osa selgitusi.

Kursus on rakendusliku suunitlusega ega sisalda matemaatilisi tõestusi, piirdudes selle asemel intuitiivsete selgitustega. Ka ei eelda kursus kuulajatelt varasemaid teadmisi tõenäosusteooriast ja matemaatilisest statistikast, sisaldades teises peatükis hädapärase teadmiste miinimumi. Kõigile neile, kes on huvitatud raamatus puudutatud probleemide sügavamast käsitlusest, soovitame A.-M. Parringu, E. Käärikü ja M. Vähi õpikut "Statistilise andmetöötluse algkursus".

Loomulikult ei saa nii väikesemahuline raamatuke sisaldada väga suurt osa statistikameetodite rikkalikust varamust. Esitatava materjali osas on tehtud väga range valik, piirdudes ühelt poolt ainult kahe tunnusega ja teiselt poolt põhiliselt klassikaliste meetoditega. Seega on välja jäetud mitteparameetrilised statistikameetodid (peale üksikute erandite), samuti on mudelite käsitluse juures jäädud lihtregressiooni ja ühefaktorilise dispersioonanalüüsi tasemele. See on autoritel võimaldanud esitada ideid ja lihtsaid rakenduslikke näiteid ilma tehnilistesse üksikasjadesse laskumata.

Kuigi tänapäeval tehakse statistilist analüüsi valdavalt spetsiaalse tarkvara abil, ei ole see raamat orienteeritud tarkvara kasutamise õpetamisele. Otse vastupidi, olulisemate arvutuste juures on esitatud ka arvutusskeemid ning raamatu lõppu on lisatud olulisemad statistikatabelid. Nende abil saab kõiki järeldusi vastu võtta ka ilma arvutit kasutamata. Selleks, et raamatut saaks ka käsiraamatuna kasutada, on raamatu lõppu lisatud indeks.

Oleme tänulikud kolleegidele Anne-Mai Parringule, Ene Käärikule ja Martin Viilile väärtuslike soovituste ja näpunäidete eest.

E.-M. Tiit,
M. Möls.

November 1996.

1. Põhimõisted. Kirjeldav ja illustreeriv statistika

1.1. Uurimisobjekt, üldkogum ja valim

1.1.1. Üldkogum. Kõikne uuring

Statistilist uuringut planeerides alustatakse *uurimisobjekti* määratlemisest. See defineeritakse vastavalt *uurimisülesandele* kui nähtus või protsess, mille kohta soovitakse teha järeldusi. Statistikas nimetatakse uurimisobjekti *üldkogumiks*. Mõningates konkreetsetes uurimisvaldkondades (bioloogia) kasutatakse üldkogumi sünonüümina ka *populatsiooni* mõistet. Üldkogumi määratlemisel fikseeritakse mõõtmishetk ja piiritletakse see ka ruumis. Näiteks rahvaloenduse korral määratakse rahvaloenduse nn *kriitiline päev* ning *loendushetkeks* loetakse selle päeva algus kell 00.00.

Uurimisobjekt kui tervik koosneb *üksikobjektidest* ehk *punktidest* (ka *indiviididest*). Ajas ja ruumis piiritletud üldkogumi korral on sellesse kuuluvate punktide hulk tavaliselt lõplik, kuid matemaatiline teooria on arendatud välja ka lõpmatu üldkogumi juhuks. *Kõikse* uuringu korral mõõdetakse kõiki üldkogumi punkte. Lõpliku üldkogumi korral tähistatakse mõõdetavate punktide arvu ehk *üldkogumi mahtu* tähega *N*.

Näide 1

Üldkogumiks on Eesti ettevõtted 1996. aasta 1. juulil. Siis oli nende arv N Eesti ettevõtteregistri alusel 93 167. Kõikse uuringu korral kogutakse andmeid kõigi Eesti ettevõtete kohta.

Üldtuntuim kõikne uuring on rahvaloendus, kuid samuti võimaldavad täielikud, kogu populatsiooni haaravad registrid teha kõikseid järeldusi. Suur osa ametlikust statistikast on kõikne, st andmeid kogutakse kõigi uurimisobjekti punktide kohta (kõik koolid, kõik raviasutused, jm). Kõikse uuringuga on seotud alljärgnevad probleemid:

- objekti täpne (kontseptuaalne) piiritlemine (näit missugust tüüpi õppeasutus on kool?);
- üldkogumi punktide hulga võimalikult täielik haaramine, st puuduvate punktide ehk puuduvate andmete arvu minimeerimine.

1.1.2. Valikuuring

Valikuuring on palju odavam kui kõikne uuring, sest valikuuringu korral ei uurita mitte kõiki üldkogumi objekte, vaid küllaltki väikest osa sellest. Valikuuringu kavandamisel määratletakse uurimisobjekt (üldkogum) ja kirjeldatakse selle punktide hulk. Mõõdetakse osa sellesse kuuluvatest punktidest. Mõõdetavate punktide hulka nimetatakse *valimiks*, valim moodustatakse *valikueeskirja* (*disaini*) alusel. Valimi põhjal tehakse järeldusi üldkogumi kohta. Järelduste õigsuse tagavad:

- valimi *esindavus* ehk *representatiivsus*;
- matemaatilise statistika poolt välja töötatud protseduurireeglid otsustuste tegemiseks.

Igasugune valimi põhjal tehtud otsustus sisaldab põhimõtteliselt alati valimi juhuslikkusest tingitud vea võimalust, kuid sellise otsustusvea tõenäosus on hinnatav. Praktikas kasutatakse enamasti niisuguseid otsustusreegleid, mille puhul otsustusvea tõenäosus on küllalt väike.

Näide 2

Pere-eelarve valikuuringu jaoks valiti rahvastikuregistrist 6000 aadressisikut, kelle kaudu lülitati uuringusse 6000 peret. Pered valiti juhusliku protseduuri alusel (näiteks genereerides 6000 juhuslikku arvu vahemikus 1...1 500 000, kui registris on 1 500 000 isikukirjet).

1.1.3. Valimi esindavus

Kõige mugavam on järeldusi teha niisuguse valimi põhjal, mis on oma struktuuri poolest küllalt sarnane üldkogumi jaotusega. Valimit, mis on lisaks sellele ka piisavalt arvukas, nimetatakse *esindavaks*. Valikuteoorias kasutatakse ka selliseid valimeid, mille struktuur erineb üldkogumi struktuurist[1]. Otsuste tegemiseks nende põhjal tuleb kasutada keerukamaid reegleid. Lihtsaim esindav valim saadakse *lihtsa juhusliku valiku* teel, mille korral igal üldkogumi punktil on võrdne võimalus valimisse sattuda.

- Lihtne juhuslik valik tehakse tavaliselt juhuslike arvude generaatori või tabeli abil, kuid on võimalik kasutada ka uuritavatest tunnustest sõltumatut tunnust. Valimist kõneldes peetakse üldjuhul silmas lihtsat juhuslikku valimit.

---

[1] Niisuguste valimite moodustamise ja kasutamise kohta soovitame lugeda I. Traadi ja J. Inno raamatut "Tõenäosuslik valikuuring" (TÜ Kirjastus, Tartu, 1997).

Näide 3

1968. aastal küsitleti Tartu elanikkonna uurimisel inimesi, kes elasid majades, mille number oli 7 või lõppes 7-ga.

Näide 4

Et saada 50-inimeselist valimit Harjumaa puudega inimeste registrist, milles oli 447 inimest, leiti, et iga valimi punkt peaks esindama 8,94 üldkogumi punkti. Inimesed registris olid tähestikulises järjestuses. Juhuslikult valiti alguspunkt vahemikust 1 kuni 1+8,94 (see alguspunkt oli 2,59) ja seejärel moodustati arvude jada sammuga 8,94. Saadud arvudest võeti täisosa, ning saadi sellega järjekorranumbrite jada 2, 11, 20, ... Valimisse võeti nende järjekorranumbritega inimesed.

Valimit iseloomustab *valimi maht n*, s.o. mõõdetavate objektide arv ning *esindavuse suhe d*, viimase määrab jagatis *n/N*. Esindavuse suhet väljendatakse sageli ka protsentides. Suhet *N/n* nimetatakse *laiendusteguriks*.

- Sageli on otstarbekas üldkogum mingi tuntud tunnuse järgi enne osadeks (*kihtideks*) jaotada ja valida igast kihist esindav valim. Siis saab koguvalimi, ühendades üksikutest kihtidest tehtud valimid. Kui igast kihist on võetud sama esindavuse suhtega valim, siis koosneb ühendvalim lihtsalt kihtide valimitest, ning ühendvalimi esindavuse suhe on sama. Kui erinevates kihtides on erinevad esindavuse suhted, siis tuleb ühendvalimi moodustamiseks kasutada kaalusid.

Näide 5

Eesti elanikkonnast eelistab kõnelda eesti keeles 66% ja vene keeles 34%. Selleks, et saada 3000-list valimit, valiti juhuslikult 1980 eesti keelt ja 1020 vene keelt kõnelevat vastajat. Saadi esindav valim nii eestikeelsete kui venekeelsete elanike jaoks, kusjuures summaarne valim on esindav kogu elanikkonnna jaoks.

Näide 6

Võrreldes Eesti saarte elanikke, küsitleti 200 saarlast ja 200 hiidlast. Seejärel sooviti moodustada saarlaste-hiidlaste ühisvalim. Et Hiiumaa elanike arv on 11 000 ja Saaremaal – 39 500, siis esindas iga küsitletud hiidlane 55 kaashiidlast ja iga saarlane 197 kaassaarlast. Arvud 55 ja 197 on vastavad laiendustegurid. Seega on saarlase vastusel ühendatud valimis 3,59 korda suurem kaal kui hiidlase vastusel.

1.1.4. Statistiline kogum

Niihästi üldkogum (kõikse uuringu puhul) kui valim (valikuuringu korral) moodustab teatavate punktide (objektide, indiviidide) hulga, nn *statistilise kogumi*, millesse kuuluvaid punkte uuritakse. Uurimisobjektiks, mille kohta järeldusi soovitakse teha, on aga alati üldkogum.

- Kõikse uuringu puhul saadakse mõõtmistulemustest teave vahetult uuritava objekti kohta.
- Valikuuringu puhul tuleb valimi mõõtmistulemustest teha järeldusi üldkogumi kohta, st lisandub üldistamisprotsess, mille puhul tuleb arvestada juhusliku vea võimalusega.

Kirjeldava statistika protseduuride puhul pole olulist vahet, kas on tegemist valik- või kõikse uuringuga. Küll aga tuleb olla tähelepanelik tulemuste tõlgendamise juures. Kui kõikse uuringu tulemused on vahetult tõlgendatavad, siis valikuuringu puhul pole esialgu selge see, missugused leitud tulemused kehtivad ka üldkogumi jaoks (mis ju uurijat tegelikult huvitab) ja missugused kehtivad ainult vaadeldavas konkreetses valimis, moodustades nö valimi eripära.

1.2. Tunnus, tunnuse tüübid

1.2.1. Tunnus, tunnuse väärtused

Tunnuseks nimetatakse üldkogumisse kuuluvate objektide mingit (arvulist või kvalitatiivselt kirjeldatavat) näitajat, mida on põhimõtteliselt võimalik mõõta. *Mõõtmine* võib olla kas füüsikaline (mingite füüsikaliste instrumentide, näiteks termomeetri abil) või psühholoogiline (mingi testi abil). Mõõtmiseks võib olla vaatlus, küsitlus, loendamine, mingi näidu leidmine dokumendist jmt.

Mõõtmise protsessis omistatakse igale mõõdetavale objektile mõõdetava *tunnuse väärtus*. Üldiselt eeldatakse, et tunnuse väärtus ei ole üldkogumis konstantne, st et üldkogumis leidub objekte, mille puhul tunnuse väärtused on erinevad.

Näide 7

Uurimisobjektiks on mingi koolilaste hulk, mõõdetavaks tunnuseks koolilapse vanus (aastates). Selle väärtus saadakse teada kas last küsitledes või tema sünnitunnistust vaadates.

Näide 8

Uuritakse sama koolilaste hulka, tehakse kindlaks nende pikkus (kasutades sentimeetrimõõtu) ja kaal (kilogrammides).

Näide 9

Küsitletakse linnaelanikke, paludes neil märkida oma haridus, valimiseelistus (erakond või valimisliit) ja hinnang linnavolikogu tööle (vastusevariandid: suurepärane, hea, rahuldav, kehvapoolne, viletsi).

Näide 10

Sotsioloogiline ankeet algab sõnadega:'Palun märkige oma sugu, rahvus, kodakondsus ja pereliikmete arv.'

1.2.2. Tunnuse tüübid

Näidetes 7–10 esineb üsna mitmesuguseid tunnuseid. Põhilised *tunnuse tüübid* on alljärgnevad.

- *Pidev arvtunnus* (enamasti füüsikalise mõõtmise tulemus), näiteks koolilapse kasv ja kaal.
- *Diskreetne arvtunnus* (enamasti loendamise tulemus), näiteks pereliikmete arv. Vahel käsitletakse ka põhimõtteliselt pidevat arvtunnust diskreetsena, näiteks vanus (aastates).
- *Järjestustunnus* (enamasti psühholoogilise mõõtmise tulemus). Tavaliselt on järjestustunnuse puhul vastusevariandid ette antud, kusjuures need variandid on sisuliselt järjestatavad, näiteks hinnang linnavolikogu tööle (vastusevariandid muutuvad järjest kriitilisemaks). Põhimõtteliselt samasugune tunnus on haridus, kui vastusevariandid grupeeritakse ja järjestatakse üldtunnustatud haridustasemete järgi.
- *Nominaaltunnus*, so mittearvuline tunnus, mille vastusevariantide jaoks ei leidu sisulist (täielikku) järjestust: valimiseelistus, rahvus.
- *Kahe väärtusega* (*binaarne* ehk *dihhotoomne*) tunnus, s.o tunnus, millel on ainult kaks võimalikku väärtust: sugu. Kahe väärtusega tunnus kuulub erijuhuna järjestustunnuste hulka, sest tema väärtused on alati sisuliselt järjestatavad. Et aga binaarsete tunnuste jaoks on välja töötatud rida erimeetodeid, on sobiv neid vaadelda eraldi tüübina. Mõnikord käsitletakse dihhotoomsetena ka neid tunnuseid, millel on rohkem vastusevariante, kuid uuritava ülesande seisukohalt lähtudes

võib osa variante ühendada: kodakondsus vastusevariantidega 'Eesti kodanik' ja 'pole Eesti kodanik'.

Enne tunnuse analüüsimist tuleb kindlaks teha selle tüüp, sest erinevat tüüpi tunnustega on lubatud erinevad statistikaprotseduurid.

1.2.3. Tunnuste kodeerimine

Mittearvuliste tunnuste puhul on sageli otstarbekas nad enne töötlemise algust *kodeerida*, st asendada sõnalised vastusevariandid arvude ehk *koodidega*. Kodeerimise puhul on kasulik silmas pidada alljärgnevaid otstarbekuse reegleid.

- Järjestustunnuste kodeerimisel tuleb jälgida, et koodid säilitaksid väärtuste sisulise järiestuse. Seega ei tohiks kasutada kodeeringut

  hea..... 1
  paha..... 2
  ei oska öelda... 3.

  Küll aga sobiksid kodeeringud

hea..... 1	hea..... 3	hea..... 1
ei oska öelda2	ei oska öelda ... 2	ei oska öelda ..0
paha.....3	paha..... 1	paha.....–1

Viimast kodeerimisreeglit ei tahaks siiski soovitada, tema puudus paistab silma kirjapildistki: miinuste kasutamine tekitab lisatööd ja suurendab eksimuste võimalust. Muidugi ei tähenda tunnuse kodeerimine naturaalarvudega seda, et saadud väärtusi ei tohiks edaspidi *teisendada* suvalisele kujule. Loomulikult võib töötlustulemuste interpreteerimisel kasutada uurijale kõige sobivamat *skaalat*, s.o tunnuse väärtuste või koodide hulka. Tuleb aga jälgida, et teisendamisel säiliksid tunnuse olulised omadused.

- Binaarse tunnuse kodeerimisel on samuti eelistatav lihtsaim võimalus, näiteks 1 ja 2 (või ka 0 ja 1, kui see on sisuliselt mõistetavam).
- Nominaaltunnuseid ei ole enamasti vaja arvuliseks kodeerida. Sageli on aga otstarbekas asendada pikad vastusevariandid kokkuleppeliste lühenditega.

1.2.4. Andmete õigsuse kontrollimine

Igasuguste andmete analüüsimise puhul on oluliseks etapiks andmete õigsuse *kontrollimine*. Andmetes võib esineda mitmetüübilisi vigu, loetleme neist olulisimad.

- *Süstemaatilised mõõtmisvead*, mida põhjustab ebatäpne instrument (kas taatlemata kaal või mõtlematult sõnastatud küsimus ankeedis).

Süstemaatilist mõõtmisviga on enamasti võimalik avastada vaid kordusuuringu abil, kus kasutatakse erinevat instrumenti.

- *Jämedad vead* (*erindid*) on enamasti tingitud mitmesugustest inimlikest eksimustest (mõõtmistulemus kirjutati sentimeetrites, mitte meetrites, nagu oli kokku lepitud; vahetati ära tunnuste järjekord, st kasvu asemele kirjutati kaal ja vastupidi; unustati üks tunnus mõõtmata ja selle tagajärjel kirjutati teiste tunnuste väärtused valedesse kohtadesse). Sellised vead on enamasti loogilise kontrolli teel leitavad.
- *Juhuslikud vead*, mida põhjustab mõõtmise ebatäpsus. Enamasti ei põhjusta sellised vead suuri eksimusi järeldustes.

Erindiks võib olla ka igati korrektse mõõtmise tulemus. Näiteks laste kaalude hulgas on ühe haiglaslikult tüseda lapse kaal. Selleks, et otsustada, kuidas erindiga käituda, tuleb lähtuda ülesande püstitusest. Kui oli tarvis leida normaalsete laste kaaalujaotust, tuleb erind kui *antud üldkogumisse mittekuuluv punkt* välja jätta. Kui aga on tarvis iseloomustada kõigi laste kaalusid, tuleb erind andmestikku alles jätta: tema esindab väga tüsedaid lapsi, keda tõenäoliselt on rohkemgi.

Enne igasuguste järelduste tegemist on oluline puhastada andmed jämedatest vigadest, sest need võivad põhjustada oluliselt ekslikke järeldusi.

1.2.5. Statistiline andmestik

Mingi kogumi mõõtmisel saadud mõõtmistulemused ehk *andmed* moodustavad *statistilise andmestiku*. Tüüpilise statistilise andmestiku korral mõõdetakse statistilise kogumi igas punktis (igal objektil) terve rida näiteks m tunnust. Sageli tähistatakse tunnuseid (millel on küll oma nimed, näiteks kasv, kaal, vanus) ka lühidalt $\underline{X}$, $\underline{Y}$, $\underline{Z}$ või $\underline{X}_1$, ..., $\underline{X}_m$. Enamasti esitatakse statistiline andmestik tabelina.

Selle tabeli kokkuleppeline kuju on alljärgnev.

- Iga tabeli veerg vastab tunnusele (st veerg indeksiga j sisaldab tunnuse $\underline{X}_j$ mõõtmistulemusi.
- Iga tabeli rida vastab mõõdetavale objektile, st i-s rida sisaldab i-nda objekti andmeid.
- Tabeli i-nda rea ja j-nda veeru lõikekohal olevas lahtris paikneb j-nda tunnuse väärtus, mis mõõdetud i-ndal objektil x_{ij}.
- Kui mingil objektil mingi tunnuse väärtus pole mõõdetud, siis on vastav koht tabelis tühi või seal paikneb puuduvat väärtust tähistav sümbol (näiteks –) ning vastav väärtus loetakse *puuduvaks*.

Puuduvate väärtuste tõttu on üldiselt iga *tunnuse valim* (st objektide arv, millel selle tunnuse väärtused on mõõdetud) erinev (väiksem vaiimi mahust n).

Näide 11

On esitatud ühe maakooli (nimetame seda Matsikooliks) 15-aastaste õpilaste antropomeetriliste andmete tabel. Sugu on kodeeritud: mees – 1; naine – 2. Kehaehituse tüüp on kodeeritud: leptosoom (sale kehaehitus) – 1; keskmine kehaehitus – 2; püknik (tüse kehaehitus) – 3.

Tabel 1

Jrk nr	Nimi	Kaal	Sugu	Kehaehituse tüüp	Õdede-vendade arv	Küla
1	Anne	54	2	1	0	Tammiku
2	Kalle	65	1	2	1	Järve
3	Teet	67	1	3	2	Järve
4	Eva	49	2	2	0	Järve
5	Mari	53	2	2	1	Lombi
6	Martin	60	1	1	1	Järve
7	Siim	72	1	2	0	Järve
8	Liis	65	2	3	0	Tammiku
9	Andres	49	1	1	1	Järve
10	Tanel	68	1	2	1	Järve
11	Priit	74	1	3	3	Raba
12	Marek	70	1	2	1	Järve
13	Katrin	51	2	1	1	Tammiku
14	Kristjan	55	1	2	0	Tammiku
15	Kristiina	58	2	2	2	Raba
16	Diana	52	2	2	2	Raba
17	Sander	65	1	2	1	Tammiku

Esimene tunnus nimi on nominaaltunnus ja kuna sellel on kõik väärtused erinevad, saab seda kasutada objektide identifitseerimiseks. Kui laste hulgas oleks olnud kaks ühenimelist ("Katrin 1" ja "Katrin 2"), siis oleks tulnud identifitseerimiseks kasutada mitut tunnust (lisades veel perekonnanime ja/või küla nime). Teine tunnus kaal on pidev arvtunnus (kuigi mõõtmistäpsuse piiratuse tõttu diskretiseeritud). Kolmas – sugu – on binaarne ja kodeeritud. Neljas – kehaehituse tüüp – on järjestustunnus, mis on kodeeritud nii, et koodi suuremale väärtusele vastab tüsedam kehaehitus. Viies tunnus – õdede-vendade arv – on diskreetne arvtunnus, seda ei ole kodeeritud. Kuues tunnus küla on samuti kodeerimata nominaaltunnus.

Märgime, et seda õpilaste hulka pole korrektne vaadelda valimina kõigi 15-aastaste eesti õpilaste seast – esiteks on kogumi maht liiga väike ja teiseks on vaatlused sõltuvad: tegemist on sama piirkonna elanikega. Niisugune uurimus on huvitav vaid ühe kooli seisukohast.

1.3. Tunnuse sagedus ja jaotus

1.3.1. Tunnuse sagedustabel

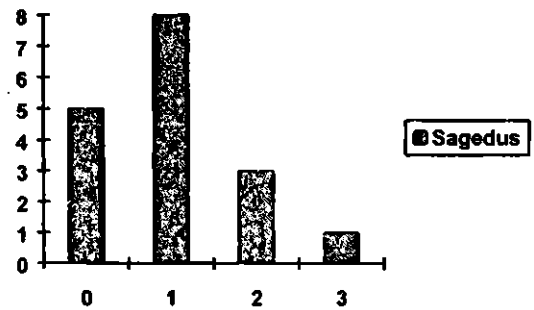
Tunnuse *sagedustabeli* moodustamiseks on tarvis lihtsalt kokku lugeda, mitu korda tunnuse iga väärtus esineb. Tunnuse väärtuse *sagedus* on tema esinemisarv vaadeldavas kogumis.

Näide 12

Moodustame sagedustabeli tabelis 1 esitatud tunnusest ödede-vendade arv. Saame alljärgneva sagedustabeli, mida illustreerib joonis 1.

Tabel 2

Ödede-vendade arv	0	1	2	3	Kokku
Sagedus	5	8	3	1	17



Joonis 1

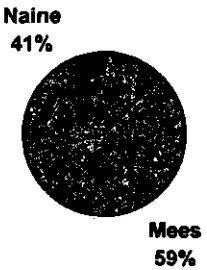
Sagedustabelis on tunnuse väärtused tavaliselt kasvavalt järjestatud. Samasuguse sagedustabeli saaksime teha ka tunnuse kehaehituse tüüp kohta. Tunnuse sagedustabelit illustreerib hästi tulpdiagramm ehk histogramm.

Näide 13

Olgu tarvis uurida ka tunnust sugu. Erilist mõtet pole küll teha sagedustabelit binaarse tunnuse kohta, sest sellel on ainult kaks sisukat lahtrit, kuid seda illustreerib päris kenasti nn sektordiagramm joonisel 2:

Tabel 3

Sugu	Mees	Naine	Kokku
Sagedus	10	7	17



Joonis 2

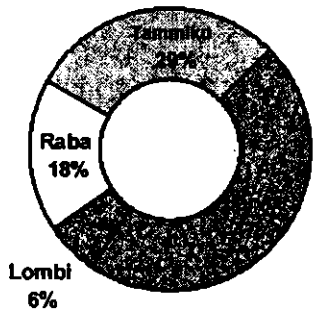
Sagedustabeli võime teha ka nominaaltunnuse küla kohta. Nominaaltunnuse puhul on aga väärtuste järjestus vabalt valitav.

Näide 14

Vaatleme tunnuse küla jaotust näitest 11. Seda iseloomustavad tabel 4 ja joonis 3.

Tabel 4

Küla	Järve	Lombi	Raba	Tammiku	Kokku
Sagedus	8	1	3	5	17



Joonis 3

1.3.2. Variatsioonrida

Mõnevõrra tülikam on teha sagedustabelit pideva arvtunnuse jaoks. Selle puhul on sobivam alustada tunnuse järjestamisest *variatsioonritta*. Variatsioonrea saame, kui järjestame tunnuse mõõdetud väärtused kasvavalt. Kui mõni väärtus esineb kogumis korduvalt, on ta ka variatsioonreas mitu korda. Sellisel juhul räägitakse *kordustega variatsioonreast*. Kui kordusi ei ole, öeldakse, et variatsioonrida on kordusteta. Variatsioonrea väikseimat elementi tähistatakse sümboliga *min*, suurimat elementi sümboliga *max*. Variatsioonrea *i*-ndat liiget tähistakse sümboliga x'_i ja seda nimetatakse *i-ndaks järkstatistikuks*. Vastava kogumi elemendi järjekorranumbrit variatsioonreas kutsutakse *astakuks*. Korduste esinemise korral variatsioonreas defineeritakse astak kui *sama väärtusega punktide keskmine järjekorranumber variatsioonreas.*

Näide 15

Tunnusele kaal saame alljärgneva variatsioonrea:

49, 49, 51, 52, 53, 54, 55, 58, 60, 65, 65, 65, 67, 68, 70, 72, 74.

Selle tunnnuse kohta näeme, et min = 49, max = 74 ja näiteks x'_3=51. Veendume, et tegemist on kordustega variatsioonreaga, kordub näiteks väärtus 65. Lihtne on määrata astakud, näiteks 52 kg kaaluva õpilase astak on 4, aga 49 kg kaaluva õpilase astak on 1,5.

1.3.3. Pideva arvtunnuse väärtuste klassifitseerimine

Et saada tabeli jaoks sobivat arvu lahtreid, tuleks tunnuse väärtused *klassifitseerida,* määrates selleks vajalikud *klassipiirid.* Praktikas sobivad niisugusteks klassipiiridcks tihti nn ümmargused, näit nulli või viiega lõppevad arvud. Niisuguste standardsete klassipiiride kasutamine muudab andmed hõlpsasti võrreldavateks. Näiteks rahvastikustatistikas kasutatakse tavaliselt 5-aastaseid vanusevahemikke, kusjuures klassipiirideks on 0 või 5-ga lõppev arv. Ühte klassi kuuluvad siis 5- kuni 9-aastased lapsed, kusjuures vanust loetakse täisaastates (ei ümardata).

Näide 16

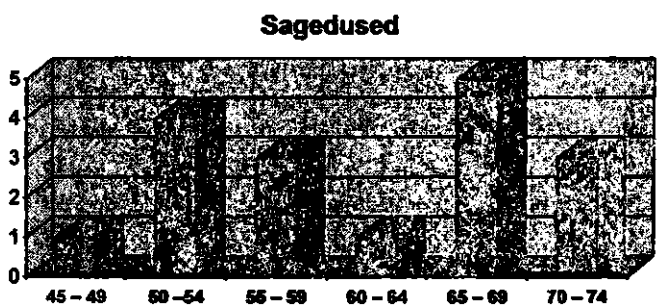
Tunnuse kaal jaoks valime klassipiirideks

$$45-49,\ 50-54,\ 55-59,\ 60-64,\ 65-69 \text{ ja } 70-74.$$

Siis saame alljärgneva tabeli, mida illustreerivad joonised 4 ja 5.

Tabel 5

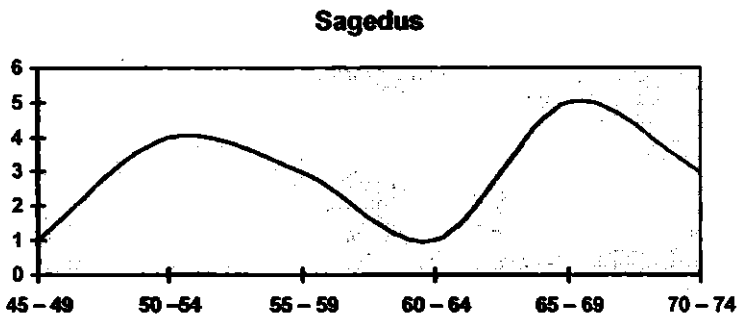
Kaaluklass	45 – 49	50 – 54	55 – 59	60 – 64	65 – 69	70 – 74	Kokku
Sagedus	1	4	3	1	5	3	17



Joonis 4

Pideva tunnuse sagedustabelit on sobiv illustreerida ka pideva *sagedus-kõvera* või *-murdjoonega.*

Käesoleval juhul ilmneb, et kogum koosneb kahest osakogumist – kergematest tütarlastest ja raskematest poistest, seetõttu on ta "kahe küüruga".



Joonis 5

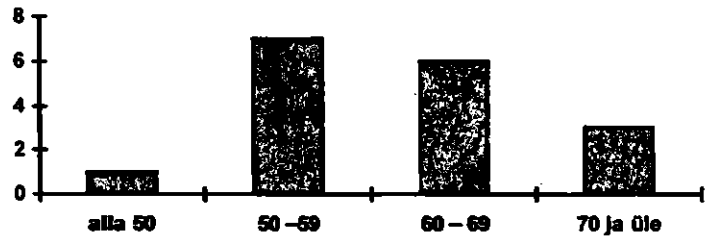
Klassipiirid võib valida ka teisiti (kasutades muuseas ka otstes nn *lahtisi klasse*).

Näide 16 (järg)

Tunnuse kaal jaoks valime uued klassipiirid, moodustame tabeli 6 ja seda illustreeriva joonise 6. Näeme, et suuremaid klasse kasutades läheb kaotsi eelmistel joonistel ilmnenud kahetipulisus.

Tabel 6

Kaaluklass	Alla 50	50 –59	60 – 69	70 ja üle	Kokku
Sagedus	1	7	6	3	17



Joonis 6

Kahe tabeli (*5* ja *6*) võrdlus veenab, et sagedustabeli ja seda illustreerivate graafikute kuju sõltub oluliselt klassifitseerimiseeskirjast.

1.3.4. Tunnuse jaotustabel

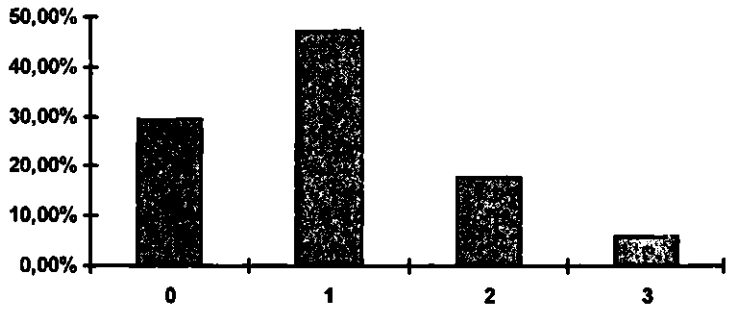
Tunnuse jaotustabelisse paigutatakse sageduste asemel *suhtelised sagedused* (mis tihti väljendatakse protsentides).

Näide 17

Asume nüüd uurima ödede-vendade arvu näites 11 esitatud andmestikust. See on diskreetne arvtunnus ning selle kirjeldamiseks moodustame jaotustabeli, vt tabel 7 ja joonis 7.

Tabel 7

Ödede-vendade arv	0	1	2	3	Kokku
Jaotus	29,41%	47,06%	17,65	5,88%	100%



Joonis 7

Jaotustabelis on alati suhteliste sageduste summa üks ehk 100%. Arvestades kogumi väikest mahtu, on käesolevas tabelis protsentide esitamistäpsusega liialdatud. Selle põhjuseks on asjaolu, et on jälgitud tava, mille kohaselt protsentide esitamise standardiks on kaks kümnendkohta. Sisuliselt oleksid need kümnendkohad õigustatud alles tuhandelise kogumi puhul.

Sagedus- ja jaotustabelid on ühesuguse kujuga, ka neid illustreerivad graafikud on üldiselt sarnased, seetõttu esitatakse andmed tihti ühise *sagedus-jaotustabelina*, vt tabel 8.

Tabel 8

Ödede-vendade arv	0	1	2	3	Kokku
Sagedus	5	8	3	1	17
Jaotus	29,41%	47,06%	17,65%	5,88%	100%

Arvtunnuse jaotust nimetatakse *sümmeetriliseks keskpunkti c suhtes*, kui igat tunnuse keskpunktist suuremat väärtust $c+t$ esineb sama sagedusega kui keskpunktist väiksemat väärtust $c-t$ ja vastupidi.

1.3.5. Sagedus- ja jaotustabelite tõlgendamine

Sagedus- ja jaotustabelid koostatakse statistilise kogumi, st kõigi tegelikult mõõdetud objektide mõõtmistulemuste põhjal. Tabeli koostamisel pole oluline see, kas on tegemist kõikse või valikuuringuga. Samuti toimuvad paljud kirjeldava statistika protseduurid formaalselt ühteviisi, sõltumata sellest, kas tegemist on valimi või üldkogumiga. Oluline on aga see, et nende tulemusi tuleb alati erinevalt tõlgendada.

Kõikse uuringu korral sisaldab sagedustabel kogu olemasoleva teabe üldkogumi vaadeldava tunnuse kohta. Sagedustabelist näeme, mitu last Tammiku külast käib Matsikoolis. Lisaks saame jaotustabelist teada, missuguses protsentuaalses vahekorras on naaberkülade lapsed nimetatud koolis.

Kui aga tegemist on valikuuringuga, soovitakse järeldusi teha üldkogumi, mitte valimi kohta. Sel juhul pakuvad valimi sagedused uurijale vähe huvi, küll aga püütakse nende põhjal hinnata tunnuse vastavate väärtuste esinemissagedusi üldkogumis.

Kui on mõõdetud üldkogum ning samast üldkogumist on võetud valim ja moodustatud jaotustabel nii üldkogumi kui ka valimi kohta, siis võivad küll valimi ja üldkogumi jaotustabelid olla erinevad, kuid esindava valimi korral iseloomustab tunnuse jaotus valimis siiski üldjoontes ka tunnuse jaotust üldkogumis. Seetõttu pakuvad valikuuringute korral sagedustabelitega võrreldes suuremat huvi jaotustabelid, mis on ühtlasi üldkogumi vastava jaotustabeli *hinnanguks*.

1.4. Arvkarakteristikud

1.4.1. Arvkarakteristiku mõiste

Arvkarakteristikud iseloomustavad uuritavat tunnust võimalikult kompaktselt ja annavad aluse tunnuste võrdlemiseks erinevates kogumites (nende osades). Kui on tehtud üldine eeldus tunnuse jaotuse kohta, võimaldab arvkarakteristikute väärtuste määramine ka tunnuse jaotust identifitseerida. Selletõttu nimetatakse arvkarakteristikuid matemaatilises statistikas ka *jaotusparameetriteks*. Põhilised arvkarakteristikute tüübid nende eesmärgi järgi on järgmised: *asendi-* (ehk paiknemise) ja *hajuvus-karakteristikud*. Lisaks neile on olemas ka jaotuse *kujukarakteristikud*, mida siiski harvem kasutatakse.

Vastavalt määramiseeskirjale on olemas *momenditüüpi* ja *järk-* (ehk *jaotusvabad*) *karakteristikud*. Viimaste kasutamine kuulub nn jaotusvaba ehk *mitteparameetrilise statistika* valdkonda.

1.4.2. Asendikarakteristikud. Keskmine

Keskmine on tuntuim asendikarakteristik, mis on defineeritud arvtunnuste korral. Keskmine iseloomustab tunnuse paiknemist. Keskmine arvutatakse ühesugusel viisil tunnuse kõigi üksikväärtuste aritmeetilise keskmisena nii üldkogumi kui valimi jaoks:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,. \qquad (1)$$

Kui tunnusel on k ($k<n$) erinevat väärtust $x'_1, \ldots, x'_k$ ja on olemas sagedustabel (väärtuse x'_i sagedus on n_i), siis saab selle valemi esitada ka kujul

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x'_i \,. \qquad (1')$$

Kui tegemist on pideva tunnusega ja väärtused on klassifitseeritud, siis tuleb keskmise arvutamiseks leida *klasside keskpunktid* $b_i = (a_{i-1} + a_i)/2$ ja kasutada valemit

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i b_i \,. \qquad (1'')$$

Saadud valem on aga eelmistest ebatäpsem, sest ei arvesta tunnuse väärtuste paiknemist klassi sees. Oluliseks muutub see ebatäpsus eelkõige siis, kui

väärtused ei paikne klassides sümmeetriliselt. Kui otsmised klassid on lahtised, st on määratud eeskirjaga $<a_1$ või $>a_k$, siis on ülalmärgitud valemi kasutamiseks tarvis leida või hinnata otsmistesse klassidesse kuuluvate väärtuste keskmine (klassi keskväärtus) ja kasutada neid b_1 ja b_k asemel.

Kui on tegemist kõikse uuringuga, siis annab valemi (1) või (1') rakendamine *üldkogumi keskmise*, mida nimetatakse ka *keskväärtuseks*. Kui vaadeldav statistiline kogum on aga valim mingist üldkogumist, siis nimetatkse valemiga (1) või (1') arvutatud keskmist *valimkeskmiseks*. Valimkeskmine on *üldkogumi keskmise hinnanguks*. Hinnangut kasutatakse sageli sellepärast, et enamikul praktikas esinevatel juhtudel pole üldkogumi keskmist vahetult võimalik arvutada.

1.4.3. Keskmise omadused

Keskmisel on alljärgnevad omadused.

- Keskmine asub kindlasti kogumi suurima ja väikseima väärtuse vahel. Piltlikult võib kujutleda, et keskmise asukohaks on vaatluspunktide raskuskese.

Sellest omadusest tuleneb kaks omadust, mis on sõnastatud eraldi nende tähtsuse tõttu:

- Mittenegatiivse tunnuse keskmine on mittenegatiivne.
- Konstantse tunnuse keskmine võrdub selle tunnuse väärtusega.

Järgmised kaks omadust on seotud tunnuste lineaarteisendustega:

- Kui tunnuse kõigi väärtustega tehakse lineaarteisendus
$$x_i \Rightarrow a+bx_i,$$
siis teiseneb vastavalt ka keskmine:
$$\bar{x} \Rightarrow a+b\bar{x}.$$
- Kahe tunnuse summa keskmine võrdub nende tunnuste keskmiste summaga.

Märgime, et täisarvulise tunnuse keskmine ei tarvitse olla täisarvuline. Mõnikord arvutatakse ka järjestustunnuse jaoks koodide keskmine. Teatavate reservatsioonidega võib seda kasutada mõningate järjestusega seotud hüpoteeside sõnastamiseks, kuid üldiselt ei ole see päris korrektne.

Näide 18

Leiame näites 11 vaadeldud maakooli õpilaste jaoks tunnuse kaal keskmise. See on 1/17 (54 + 65 + ... + 65) = 60,41. Kui aga kasutaksime tabelit 5, saaksime klasside keskmisteks 47,5, 52,5, 57,5 jne, ning kaalu keskmise hinnanguks on valemi (1') kohaselt 61,62.

1.4.4. Mediaan ja mediaanklass

Lõplikus kogumis on *mediaan* selline tunnuse väärtus, millest pooled kogumi punktide väärtused on väiksemad ja pooled suuremad. Selle definitsiooni kohaselt on mediaani leidmiseks tarvis kõigepealt kõik vaatlustulemused järjestada variatsioonritta, kus i-ndal kohal on väärtus x'_i.

- Paarituarvulise mahuga kogumi puhul on mediaaniks vaatlus järjekorranumbriga $(n+1)/2$.
- Paarisarvulise vaatluste arvu korral ei lange mediaan ühegi vaatlusega kokku, vaid see on kahe keskmise vaatluse poolsumma $0{,}5\,(x'_{n/2} + x'_{n/2+1})$.

Seega leitakse mediaan alljärgneva eeskirja abil:

$$med = \begin{cases} x_{(n+1)/2}, & \text{kui } n \text{ on paaritu,} \\ 0{,}5(x_{n/2} + x_{n/2+1}), & \text{kui } n \text{ on paaris.} \end{cases} \qquad (2)$$

Kui vaatlustest on moodustatud sagedustabel, siis saab kõigepealt leida *mediaanklassi*, st sellise klassi (järjekorranumbriga h), et sageduste summa $s(h-1) = n_1 + \ldots + n_{h-1}$ on väiksem kui $n/2$, kuid summa $s(h) = n_1 + \ldots + n_h$ on suurem suurusest $n/2$ või sellega võrdne. Mediaanklass on karakteristik, mida saab leida nii arv- kui ka järjestustunnuste jaoks.

Kui lisaks mediaanklassile soovitakse tabeli põhjal leida ka tunnuse mediaani, tuleb seda hinnata, pidades silmas, et mediaan paikneb alati mediaanklassis. Mediaani väärtust hinnatakse, eeldades, et punktid on klassis jaotatud võrdsete vahedega. Siis on h-ndas klassis punktide vahe $(a_h - a_{h-1})/n_h$ ja mediaani hinnangu saame lihtsa lineaarse interpolatsiooni valemi rakendamisel:

$$med = (n/2 - s(h-1))\,(a_h - a_{h-1})/\,n_i.$$

Kui on tegemist kõikse uuringuga, siis annab valemi (2) rakendamine *üldkogumi mediaani*. Kui vaadeldav statistiline kogum on aga valim mingist üldkogumist, siis nimetatkse valemiga (2) arvutatud suurust *valim-mediaaniks*. Valimmediaan on *üldkogumi mediaani hinnang*.

Näide 19

Matsikooli õpilaste kaalu mediaaniks on variatsioonreas 9. kohal asuva õpilase kaal, s.o Martini kaal, 60 kg (vaata variatsioonrida näites 15). Kasutades tabelit 6, saaksime mediaanklassiks neljanda klassi, ja kuna selles klassis on üksainus mõõtmistulemus, tuleks klasside järgi hinnatud mediaan klassi keskpunkti, s.o. 62,5 kg.

1.4.5. Mediaani ja keskmise vahekord

Sümmeetrilise jaotusega arvtunnuse puhul langevad mediaan ja keskväärtus kokku. Statistiline kogum ei ole aga peaaegu kunagi täiesti sümmeetriline, seetõttu on kogumi põhjal arvutatud mediaan ja keskmine tavaliselt rohkem või vähem erinevad. Erinevus on suur siis, kui jaotus on tugevasti ebasümmeetriline. Kui jaotusel on nn *raske saba* näiteks paremal, siis kallutab see saba ka keskväärtust paremale, kuid mõjutab suhteliselt vähe mediaani. Mediaan pole ka tundlik jämedate vigade suhtes: mediaani väärtust ei mõjuta see, kas variatsioonrea maksimaalne liige on üsna lähedane oma naaberliikmetele või erineb sellest sadu kordi. Keskväärtust aga mõjustab jäme viga või erind märgatavalt.

Näide 20

Ühel päeval registreeris perekonnaseisubüroos oma abielu 5 pruuti vanuses 19, 20, 20, 21 ja 60 aastat. Arvutame selle kogumi põhjal pruutide keskmise vanuse ja vanuse mediaani. Näeme, et keskmine vanus on 28 aastat, mediaan – 20 aastat. Suur erinevus nende kahe näitaja vahel on tingitud sellest, et üks mõõtmistulemus erineb ülejäänutest märgatavalt.

Näites 20 kirjeldatud kogumit ei saa käsitleda valimina üldkogumist, selleks on tema maht liialt väike. Seega ei saa siin toodud andmete põhjal teha järeldusi pruutide üldise abiellumisvanuse kohta.

1.4.6. Mood

Mood on see tunnuse väärtus, millele vastab suurim sagedus. Kui tunnusel on mitu sellist väärtust, millele vastab suurem sagedus kui naaberväärtustele, siis öeldakse, et tunnus on *mitme moodiga* ehk *multimodaalne* ja kõiki selliseid suurema sagedusega väärtusi nimetatakse moodideks. Kui tunnusel on kaks moodi, siis nimetatakse seda tunnust *bimodaalseks*. Moodi saab arvutada igat tüüpi tunnuste puhul, kuid suure väärtuste arvuga pideva arvtunnuse puhul pole moodil erilist mõtet, sest siis on tihti kõigi üksikväärtuste sagedused väiksed. Moodklass on see klass, mille sagedus on suurim.

Näide 21

Õpilaste andmetel saame tunnuse kaal moodiks väärtuse 65. Joonistelt 4 ja 5 näeme, et kaal on bimodaalne tunnus: üks mood vastab poiste, teine tüdrukute kaalule. Tunnuse küla moodiks on Järve, sest Järve külast on pärit rohkem lapsi kui teistest küladest.

1.4.7. Kaalutud keskmine

Kui on alust eeldada, et kogumi punktid ei ole otsustuste tegemise seisukohast samaväärsed, siis kasutatakse keskmise kõrval asendikarakteristikuna ka *kaalutud keskmist*. Kaalutud keskmist kasutatakse peamiselt kahel praktikas esineval juhul.

- On teada, et mõõtmised on erineva täpsusega, ning on olemas mingid mõõtmistäpsuse hinnangud. Selline olukord on eriti iseloomulik planeeritud katsete puhul. Siis võetakse suurema kaaluga arvesse täpsemad mõõtmised, et summaarset juhuslikku viga minimeerida.
- Valimi eeskirjast (disainist) järeldub, et valimi erinevad punktid esindavad erinevat hulka üldkogumi punkte. Sel juhul soovitatakse valikuteoorias valida punktide kaalud võrdeliselt nende laiendusteguritega.

Üldiselt on kaalutud keskmise arvutamise eeskiri alljärgnev: olgu igale valimi punktile x_i omistatud (ette teada olev) kaal w_i. Siis arvutatakse kaalutud keskmine alljärgneva eeskirja abil:

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

kus w tähistab kaalude summat,

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

1.4.8. Geomeetriline keskmine

Lisaks aritmeetilisele keskmisele kasutatakse tunnuse asendi iseloomustamiseks mõnikord ka *geomeetrilist keskmist*

$$\bar{x}_g = \sqrt[n]{x_1 x_2 \ldots x_n} .$$

Praktiliseks kasutamisjuhiseks on, et geomeetriline keskmine sobib positiivsete väärtustega tunnuste jaoks (nullväärtusi ei tohi olla), kui jaotusel on raske saba paremal (leidub üksikuid eriti suuri väärtusi, mis aga pole erindid). Sel juhul nihkub aritmeetiline keskmine tugevasti paremale võrreldes geomeetrilise keskmisega, mida üksikud hälbinud väärtused mõjustavad vähem.

Näide 22

Leiame näites 20 vaadeldud statistilise kogumi tunnuse vanus geomeetrilise keskmise: $(19 \times 20 \times 20 \times 21 \times 60)^{0,2} = 24,9$.

1.4.9. Tunnuste teisendamine keskmiste leidmiseks

Geomeetrilise keskmise leidmiseks kasutatakse tavaliselt logaritmilist teisendust:

$$\ln x_g = 1/n\,(\ln x_1 + \ldots + \ln x_n).$$

Põhimõtteliselt saab arvutada keskmisi, kasutades samal viisil ka teisi funktsioone:

- igale üksikvaatlusele rakendatakse teatavat funktsiooni $f(\cdot)$,
- leitakse saadud tulemuste keskmine,
- sellele keskmisele rakendatakse funktsiooni $f(\cdot)$ pöördfunktsiooni.

Niisugustest funktsioonkeskmistest on tuntuimad *ruutkeskmine* (mida kasutatakse eeskätt hajuvusega seotud ülesannetes), samuti ka *harmooniline keskmine*, mille puhul funktsiooniks $f(x)$ on $1/x$.

Näide 23

Leiame näites 20 vaadeldud kogumi vanuse ruutkeskmise:

$$\sqrt{1/5(19^2 + 20^2 + 20^2 + 21^2 + 60^2)} = 32{,}26$$

ja harmoonilise keskmise:

$$\frac{1}{\frac{1}{5}(1/19 + 1/20 + 1/20 + 1/21 + 1/60)} = 23{,}05.$$

Paneme tähele, et erindi osatähtsust võimendab ruutkeskmine kõige rohkem, seevastu aga harmooniline keskmine pigem tasandab erindi mõju.

Esitatust pole õige teha järeldust, et üks asendikarakteristik on *hea* ja teine *halb*, vaid tuleb mõista, et mitme asendikarakteristiku kasutamine annab valimi kohta rohkem teavet, eriti siis, kui nad üksteisest oluliselt erinevad.

1.4.10. Hajuvuse karakteristikud. Dispersioon

Tunnuse *hajuvus* iseloomustab seda, kui erinevad on selle tunnuse väärtused kogumi erinevatel objektidel. Kui kõigil kogumi objektidel on uuritava tunnuse väärtus sama, siis on tunnus konstantne ja temal puudub hajuvus. Niisugune tunnus ei ole juhuslik, ning tema uurimine ei paku statistika seisukohast huvi. Tunnuse hajuvust ja jaotuse kuju iseloomustavad karakteristikud leitakse sageli *hälbeid* kasutades. *Hälbeks* nimetatakse tunnuse üksikväärtuse erinevust tunnuse keskmisest $x_i - \bar{x}$.

Tunnuse hajuvust iseloomustavatest näitajatest on olulisim *dispersioon*.

Kõikse uuringu dispersioon, mida tähistatakse tähisega σ^2 (sigma-ruut), on vaatluste hälvete ruutude keskväärtus,

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 . \qquad (3)$$

Valimi põhjal arvutatakse *valimdispersioon* s^2, kasutades selleks eelnenust veidi erinevat arvutuseeskirja (seda on lähemalt selgitatud punktis 3.1.6):

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 . \qquad (4)$$

Valimdispersioon s^2 on üldkogumi dispersiooni σ^2 hinnanguks. Valimdispersiooni on sobiv kasutada siis, kui on tegemist valikuuringuga, mille põhjal soovitakse teha järeldusi üldkogumi kohta. Valemite (3) ja (4) võrdlusest järeldub, et väga suurte valimite puhul, mille maht ulatub tuhandeni ja üle, muutub erinevus nende rakendamise tulemusel saadud dispersioonihinnangute vahel tühiselt väikseks.

Dispersiooni definitsioonist järelduvad tema omadused.

- Mida rohkem on tunnusel keskväärtusest tugevasti hälbivaid väärtusi ja mida suuremad on need hälbed, seda suurem on dispersioon.
- Konstantse tunnuse dispersioon on 0.
- Dispersioon on alati mittenegatiivne.
- Juhusliku suuruse X nihe $X \Rightarrow X+a$ ei mõjuta tema dispersiooni.
- Juhusliku suuruse korrutamisel konstandiga c suureneb tema dispersioon c^2 korda.

Näide 24

Arvutame näites 20 esitatud tunnuse X (pruudi abiellumisvanus vaadeldaval päeval) dispersiooni $\sigma^2 = 1/5\ (\ (-9)^2 + (-8)^2 + (-8)^2 + (-7)^2 + (32)^2\ ) = 256{,}4$. *Kuna me uurime selles näites vaid ühel päeval ühes perekonnaseisubüroos abiellunud pruutide vanust, mitte pruutide abiellumisvanust üldiselt, siis moodustavad antud viis pruuti üldkogumi ja me kasutame üldkogumi dispersiooni* σ^2.

1.4.11. Standardhälve

Dispersioonil on oluline tähtsus mitmesugustes teoreetilistes arutlustes, kuid tema puuduseks on, et ta on väljendatud nõ ruutühikutes (võrreldes tunnuse enese väärtusega). Selletõttu kasutatakse praktikas rohkem *standardhälvet*. Standardhälve on defineeritud kui ruutjuur dispersioonist. Kõikse uuringu standardhälve σ on ruutjuur kõikse uuringu dispersioonist σ^2. Valimi standardhälve s on ruutjuur valimdispersioonist: $s = \sqrt{s^2}$. See on ühtlasi hinnanguks (tundmatule) *üldkogumi standardhälbele*.

Näide 25

Kasutades näites 24 leitud pruutide abiellumisvanuste dispersiooni σ^2=256,4, leiame selle standardhälbe: σ = 16,0.

1.4.12. Hajuvuse hindamine järkstatistikute abil

Järkstatistikud on variatsioonrea liikmed (vt punkt 1.3.2). Lisaks mediaanile arvutatakse variatsioonrea põhjal tavaliselt ka *kvartiilid*, s.o. sellised variatsioonrea punktid, mis jagavad koos mediaaniga variatsioonrea neljaks võrdse elementide arvuga osaks. Praktiliselt soovitatakse selleks leida mediaaniga poolitatud variatsioonrea kummastki osast omakorda mediaan. Variatsioonrea vasakpoolse osa mediaani nimetatakse sel juhul alumiseks, parempoolse osa mediaani aga ülemiseks kvartiiliks. Saab kõnelda kolmest kvartiilist: esimene on *alumine kvartiil*, teine on mediaan ja kolmas on *ülemine kvartiil*.

Kõik kvartiilid iseloomustavad tunnuse paiknemist, kuid ülemise ja alumise kvartiili vahe, nn *kvartiilhaare*, on ühtlasi ka tunnuse hajuvuse karakteristikuks – mida suurem see on, seda suurem on üldiselt tunnuse hajuvus.

Võrreldes standardhälvet ja kvartiilhaaret tunnuse hajuvuse karakteristikuna, tuleb märkida järgmist.

- Standardhälbe eeliseks on see, et teda kasutatakse paljudes standardsetes statistikaprotseduurides (*t*-test, usalduspiiride arvutamine jne).
- Kvartiilhaare on stabiilsem erindite suhtes: tema suurust ei mõjuta üksik keskmisest väga kaugel paiknev väärtus (erind).

Enamasti on tunnuse kvartiilhaare mõnevõrra väiksem kui selle tunnuse kahekordne standardhälve, kuid see reegel ei ole absoluutne. Mida väiksem on kvartiilhaare kahekordse standardhälbega võrreldes, seda tihedamini on punktid koondunud mediaani lähedusse.

Näide 26

Leiame näites 19 vaadeldud tunnuse kaal kvartiilid, need on vastavalt 53 (alumine kvartiil) ja 67 (ülemine kvartiil). Kvartiilhaare on seega 14 kg.

Lisaks kvartiilidele kasutatakse väga sageli ka *detsiile*, st punkte, mis jagavad variatsioonrea kümneks osaks ja *protsentiile*, mis jagavad variatsioonrea sajaks osaks. Hajuvusnäitajana on käibel ka *variatsiooniulatus* ehk *haare*, mis avaldub variatsioonrea äärmiste punktide vahena $max - min$. Tuleb arvesse võtta, et see viimane näitaja sõltub valimi mahust: valimi suurenedes üldiselt minimaalne valimi element väheneb ja maksimaalne suureneb. Siiski on teatavatel eeldustel võimalik ka variatsiooniulatust kasutada hajuvuse hindamiseks, kuid see on mõeldav üksnes väikeste valimite korral.

1.4.13. Jaotuse kujukarakteristikud

Lisaks asendi- ja hajuvuskarakteristikutele tuntakse veel karakteristikuid, mis iseloomustavad tunnuse jaotusgraafikute (näit sageduskõvera või -murdjoone) kuju. Nendest nn *kujukarakteristikutest* olulisim on *asümmeetriakordaja a*, mis arvutatakse tunnuse hälvete kuupide summa järgi:

$$a = \frac{1}{n\sigma^3} \sum_{i=1}^{n} (x_i - x)^3. \qquad (5)$$

Asümmeetriakordaja väärtus võib olla positiivne või negatiivne vastavalt sellele, kas rohkem suuri hälbeid on tunnuse suurte või väikeste väärtuste poolel.

- Positiivne on asümmeetriakordaja siis, kui jaotusel on *raske saba* suurte väärtuste poolel, so paremal.
- Negatiivne on asümmeetriakordaja siis, kui jaotusel on *raske saba* vasakul, st leidub keskväärtusest eriti kaugeid väikesi väärtusi.
- Sümmeetrilise jaotuse korral on asümmeetriakordaja 0. Nullilisest asümmeetriakordajast ei järeldu, et jaotus oleks täiesti sümmeetriline, kuid enamasti järeldub sellest asjaolu, et jaotus on sümmeetrilisele küllalt lähedane.

Teine oluline jaotuse kujukarakteristik on *ekstsess e* (ka järsakus), mis arvutatakse hälvete neljandate astmete summa kaudu:

$$e = \frac{1}{n\sigma^4} \sum_{i=1}^{n} (x_i - \bar{x})^4 - 3. \qquad (6)$$

Positiivne ekstsess tunnistab, et jaotusel on terav tipp ja rasked sabad (kas ühele või mõlemale poole). Negatiivne ekstsess aga iseloomustab jaotusi, mille väärtused paiknevad keskväärtusele suhteliselt lähedases piirkonnas ja sabad puuduvad või on väga kerged.

Valemitega (5) ja (6) on esitatud kõikse uuringu asümmeetriakordaja ja ekstsess. Kui on tegemist valikuuringuga, mille põhjal soovitakse üldkogumi asümmeetriakordajat ja ekstsessi hinnata, siis arvutatakse vastavad valimkarakteristikud, milleks valemeid (5) ja (6) muudetakse alljärgnevalt:

- σ asendatakse selle hinnanguga s;
- n asendatakse vahega $n - 1$.

Näide 27

Leiame näites 11 vaadeldud tunnuse kaal asümmeetriakordaja ja ekstsessi, kasutades üldkogumi asümmeetriakordaja ja ekstsessi arvutusvalemeid (5) ja (6). Nende väärtusteks saame vastavalt 0,078 (asümmeetriakordaja) ja –1,40 (ekstsess). Järelikult ei paista jaotus silma erilise ebasümmeetria poolest. Ekstsessi väärtus näitab, et sabad puuduvad hoopis ja ka teravat tippu ei ole, tunnuse jaotus on pigem ühtlaselt lauge kujuga.

Näide 28

Leiame näites 20 vaadeldud tunnuse abiellumisvanus asümmeetriakordaja ja ekstsessi. Saame tulemuseks a=1,49 ja e=0,24. Näeme, et abielluvate naiste jaotus on ebasümmeetriline (asümmeetriakordaja on positiivne – järelikult leidub abiellujate hulgas keskmiselt märksa vanemaid naisi).

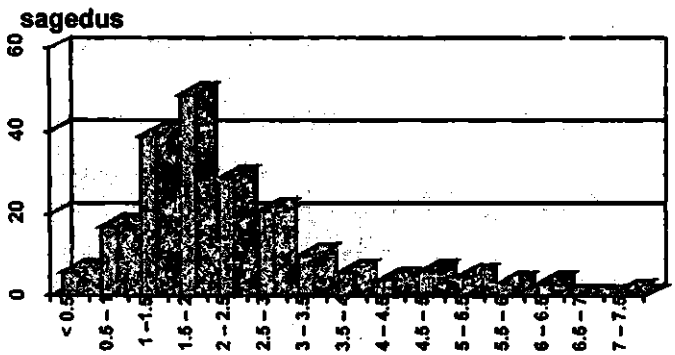
Näide 29

Vaatleme tunnusena elanike arvu Eesti valdades. Tabelis 9 on Eesti valdade jaotus elanike arvu järgi (tuhandetes). Tabelit illustreerib joonis 8.

Tabel 9

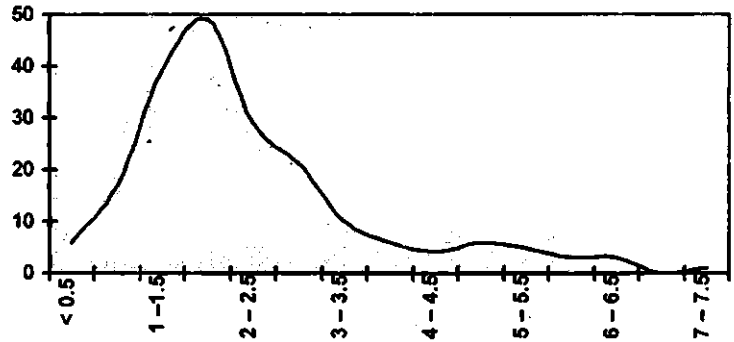
Elanike arv	<0,5	0,5– 1	1 -1,5	1,5 – 2	2 – 2,5	2,5 – 3	3 – 3,5	3,5 – 4
Sagedus	6	17	39	49	29	21	10	6

Elanike arv	4 – 4,5	4,5 – 5	5 – 5,5	5,5 – 6	6 – 6,5	6,5 – 7	7 – 7,5
Sagedus	4	6	5	3	3	0	1



Joonis 8

Kuna tegemist on kõikse uuringuga, vaadeldud on kõiki Eesti valdu, siis saame arvutada selle tabeli põhjal üldkogumi tegelikud arvkarakteristikud. Vallaelanike arvu keskväärtus on 2214,2; mediaan 1893,5; dispersioon 1754503 ja standardhälve 1324,6. Asümmeetriakordaja on 1,399 ja ekstsess 1,816. Alumine ja ülemine kvartiil on 1327 ja 2665, miinimum ja maksimum on 62 ja 7121. Lisame ka silutud sageduskõvera (hulknurga) graafiku (vt joonis 9). Tõlgendame jaotuse karakteristikuid.



Joonis 9

- *Näeme joonistelt, et jaotusel on märgatav raske saba suurte väärtuste pool.*
- *Maksimum erineb mediaanist ligi kolm (täpsemalt: 2,85) korda rohkem kui miinimum.*
- *Parempoolset asümmeetriat näitab ka positiivne asümmeetriakordaja.*
- *Jaotuskõvera kujule on iseloomulik suhteliselt terav tipp ja pikaks veninud saba – seda näitab positiivne ekstsess.*
- *Ka see, et mediaan on keskväärtusest märksa väiksem, näitab just parempoolse saba olemasolu.*
- *Kvartiilhaare on märksa väiksem kahekordsest standardhälbest, seegi näitab, et mediaani ümbrusse on koondunud palju punkte (jaotuskõveral on suhteliselt kõrgele ulatuv tipp).*

2. Tõenäosusteooria põhimõisted

2.1. *Sündmus ja tõenäosus*

2.1.1. Tõenäosusteooria alused teoreetilises (matemaatilises) statistikas

Statistiliste andmete põhjal otsustusi tehes seisab uurija silmitsi tõsiasjaga, et tal on pidevalt oht teha eksijäreldusi selletõttu, et *andmestik on juhuslik* ja selle põhjal arvutatud jaotusparameetrid (arvkarakteristikud) erinevad vastavatest parameetritest üldkogumis. Matemaatiline statistika on välja töötanud 'mängureeglid', mis võimaldavad niisugust ohtu hoida teatud mõttes kontrolli all. Nende mängureeglite mõistmiseks on tarvis teatavaid minimaalseid üldteadmisi tõenäosusteooriast, mida me käesolevas peatükis esitamegi.

2.1.2. Sündmus

Tõenäosusteooria põhimõisteks on *sündmus* ja see defineeritakse *katse* abil. *Katse* all mõistetakse teatavat juhuslikku valikut etteantud *katsetulemuste* hulgast. Eeldatakse, et katse on suvaline arv kordi korratav. Katsetulemuste kohta eeldame:

- katsel on lõplik arv võimalikke tulemusi;
- katse teostamisel esineb alati täpselt üks katsetulemus;
- katsetulemused on võrdvõimalikud, st et neil kõigil on niisama suur võimalus katse tulemusena esineda.

Sündmuseks nimetatakse iga katsetulemuste hulka. Sündmuses kui hulgas sisalduvad katsetulemused on vaadeldavale sündmusele *soodsad.*

Näide 30

Tavalised näited katse kohta on:

- *Täringuvise, kus katsetulemusteks on 1, 2, 3, 4, 5 ja 6 silma pealejäämine. Sündmuseks on näiteks paarisarvulise tulemuse saamine, see koosneb katsetulemustest 2, 4 ja 6.*
- *Mündivise, kus katsetulemusteks on kirja- ja vapipoole pealelangemine, samad on ühtlasi sündmused.*
- *Kaardi tõmbamine kaardipakist. Standardpaki korral on 52 erinevat katsetulemust, erinevaid sündmusi saab nende abil defineerida väga palju, nimelt 2^{52}, so ligikaudu $4{,}5 \times 10^{15}$. Sündmuste näiteiks on – saadakse potimastist kaart; saadakse piltkaart jne. Ärtukuninga saamine katsetulemusena on soodus sündmusele 'saadi piltkaart', kuid ebasoodus sündmusele 'saadi potikaart'.*

Sündmusi tähistatakse tavaliselt suurte tähtedega: *A, B, C, ...*

Statistikaülesannete lahendamisel võime katsena kirjeldada ühe punkti (objekti) väljavalimist kas üldkogumist või ka valimist. Esimesel juhul on võimalike katsetulemuste arv n, teisel juhul N, ning vastavalt sellele on ka erinevate sündmuste arv kas 2^n või 2^N.

Näide 31

Vaatleme näites 11 kirjeldatud andmestikku (maakoolide lapsed). Olgu katseks juhusliku õpilase valik. Siis on katsel 17 võimalikku tulemust. Nende katsetulemuste abil saab defineerida näiteks alljärgnevad sündmused: A – valitud õpilane on tütarlaps; B – valitud õpilane on pärit Tammiku külast; C – valitud õpilane on tütarlaps Järve külast. Tabelist 1 näeme, et sündmuse A jaoks on 7 soodsat katsetulemust, sündmuse B jaoks – 3 soodsat katsetulemust (Anne, Liis ja Kristjan) ning sündmuse C jaoks ainult üks soodus katsetulemus (valitakse Eva).

2.1.3. Tehted sündmustega

Kuna sündmus on defineeritud kui katsetulemuste hulk, siis saab sündmuste abil defineerida ka tehteid, mille tulemusena saadakse uued sündmused. Need tehted langevad sisuliselt ühte tehetega katsetulemuste hulkadega.

- *Sündmuste A* ja *B summa* $A \cup B$, mis toimub siis, kui toimub kas sündmus A, sündmus B või mõlemad.
- *Sündmuste A* ja *B korrutis* $A \cap B$, mis toimub siis, kui toimuvad mõlemad sündmused A ja B.
- *Sündmuste vahe* $A \setminus B$, mis toimub siis, kui toimub A, aga ei toimu B.
- *Sündmuse A vastandsündmus* A^c, mis toimub alati siis, kui A ei toimu.

Näide 32

Vaatleme näites 31 defineeritud sündmusi. Sündmus $B \cup C$ *toimub siis, kui valituks osutub kas laps Tammiku külast või tütarlaps Järve külast, seega soodsaid tulemusi on 4. Sündmus* $A \cap B$ *toimub siis, kui valituks osutub üksik tütarlaps Tammiku külast, selle sündmuse jaoks on soodsaid katsetulemusi 2 (Anne ja Liis). Sündmus* $B \setminus A$ *toimub siis, kui valituks osutub laps Tammikult, kes ei ole tüdruk, seega on* $B \setminus A$ *jaoks ainult üks soodus katsetulemus (Kristjan). Sündmuse A vastandsündmus on, et valitud lapseks osutub poiss, ja selle jaoks on kümme soodsat katsetulemust.*

2.1.4. Sündmustevahelised seosed

Kasutades sündmuse määratlust katsetulemuste hulgana, saame kindlaks teha ka sündmuste vahekorrad.

- Kui mingi sündmuse jaoks ei leidu ühtki soodsat katsetulemust (sündmusele vastav katsetulemuste hulk on tühi), siis nimetatakse seda sündmust *võimatuks sündmuseks*. Võimatu sündmuse tähis on $\varnothing$.
- Kui mingi sündmuse jaoks on kõik katsetulemused soodsad, siis on see sündmus *kindel sündmus*. Kindla sündmuse tähis on Ω.
- Sündmust, mis pole ei kindel ega võimatu, nimetatakse *juhuslikuks sündmuseks*. Juhusliku sündmuse toimumine või mittetoimumine sõltub juhusest, st sellest, missuguse tulemuseni katse sooritamisel jõuti.
- Kui sündmuse A jaoks soodsate katsetulemuste hulk sisaldub sündmuse B jaoks soodsate katsetulemuste hulgas, siis *järeldub* sündmuse A toimumisest sündmuse B toimumine. Seda sisaldussuhet märgitakse $A \subset B$.

Näide 33

Vaatleme näiteis 31 ja 32 kirjeldatud sündmusi ja veendume, et

- *kõik seni kirjeldatud sündmused on juhuslikud;*
- *sündmus C sisaldub sündmuses A, sest alati, kui toimub sündmus C (valituks osutus tütarlaps Järve külast), toimub ka A (valiti tütarlaps);*
- *sündmuste A ja A^c summa on kindel sündmus, valitud laps on kas poiss või tüdruk;*
- *sündmuste B ja C korrutis on võimatu sündmus: valitu ei saa olla üheaegselt laps Tammiku külast ja Järve küla tütarlaps.*

Sündmusi, mis ei saa üheaegselt toimuda, nimetatakse (üksteist) *välistavateks sündmusteks*.

- Katsetulemused on alati üksteist välistavad.
- Sündmus ja tema vastandsündmus on üksteist välistavad.
- Kui sündmused A ja B on üksteist välistavad, siis on nende korrutis võimatu sündmus,

$$A \cap B = \varnothing.$$

2.1.5. Sündmuse tõenäosus

Kui vaadelda sama katse tulemuste kaudu määratletud sündmusi, on selge, et nende toimumise võimalikkus on erinev. Sellel sündmusel, mille jaoks on rohkem soodsaid katsetulemusi, on rohkem võimalusi ka toimuda. Seda tõsiasja väljendabki sündmuse *tõenäosus*.

Sündmuse A *tõenäosus* $P(A)$ on sündmuse jaoks soodsate katsetulemuste arvu $k(A)$ ja kõigi katsetulemuste arvu n suhe:

$$P(A) = k(A)/n.$$

- Tõenäosus on arv 0 ja 1 vahel.
- Mida suurem on sündmuse tõenäosus, seda rohkem on alust loota selle sündmuse toimumist.
- Kui sündmuste A ja B vahel on (range) sisaldusseos $A \subset B$, siis kehtib võrratus $P(A) < P(B)$.
- Võimatu sündmuse tõenäosus on 0.
- Kindla sündmuse tõenäosus on 1.

Näide 34

Vaatleme näiteis 31–32 kirjeldatud sündmusi ja arvutame nende tõenäosused, arvestades, et katsetulemuste koguarv n=17.

$P(A) = 7/17 = 0{,}412$; $P(B) = 3/17 = 0{,}176$; $P(C) = 1/17 = 0{,}059$; $P(A^c) = 0{,}588$; $P(B \cup C) = 0{,}235$.

2.1.6. Tehted tõenäosustega

Kui me teeme sündmustega mingi tehte ja teame lähtesündmuste tõenäosusi, siis saame arvutada ka tehete tulemusena saadud sündmuse tõenäosuse.

- Kahe teineteist välistava sündmuse A ja B summa tõenäosus võrdub liidetavate tõenäosuste summaga

$$P(A \cup B) = P(A) + P(B), \text{ kui } A \cap B = \varnothing.$$

- Sündmuse A vastandsündmuse A^c tõenäosus avaldub ühe ja sündmuse A tõenäosuse vahena

$$P(A^c) = 1 - P(A).$$

2.1.7. Sõltumatud sündmused

Kui ühe sündmuse toimumine ei mõjuta teise sündmuse toimumise tõenäosust, siis on need sündmused *sõltumatud*. Sõltumatute sündmuste korral kehtib võrdus:

$$P(A \cap B) = P(A)\,P(B).$$

Viimast võrdust saab kasutada ka sõltumatuse defineerimiseks – *kui sündmuste korrutise tõenäosus võrdub tegursündmuste tõenäosuste korrutisega, siis on need sündmused sõltumatud.*

Näide 35

Olgu katseks kaardipakist kaardi tõmbamine, sündmuseks A potimastist kaardi saamine ja sündmuseks B – piltkaardi saamine. Need sündmused on sõltumatud, sest: P(A) = 13/52 = 1/4; P(B)= 12/52 = 3/13 ja P(A ∩ B)= 3/52=1/4×12/52.

2.1.8. Tinglik tõenäosus

Kui on teada, et mingi sündmus A kindlasti toimub või toimus, siis on teada ka, et katse tulemusena toimus üks selle sündmuse jaoks soodsatest katsetulemustest. Sellisel juhul jäävad ülejäänud katsetulemused arvestusest välja, ja me saame arvutada nn *tinglikud tõenäosused.* Mingi teise sündmuse B jaoks saame siis leida selle tingliku tõenäosuse tingimusel, et *A toimub (lühidalt: B tinglik tõenäosus tingimusel A)*, mida tähistatakse sümboliga $P(B|A)$. Tinglik tõenäosus arvutatakse alljärgnevast valemist:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

- Kui sündmused A ja B on sõltumatud, st kui sündmuse A toimumine ei muuda sündmuse B toimumise tõenäosust, siis on sündmuse B tinglik tõenäosus tingimusel A võrdne selle sündmuse tõenäosusega,

$$P(B|A) = P(B).$$

Näide 36

Leiame tõenäosuse, et juhuslikult valitud õpilane pärineb Raba või Järve külast, kui on teada, et ta on tütarlaps. Leiame järgmiste sündmuste tõenäosused: A – valitud õpilane on tütarlaps, P(A)=0,412 ja P(A ∩ D) – õpilane on tütarlaps Raba või Järve külast 3/17=0,176. Otsitav tinglik tõenäosus on

$$P(D|A) = 0,176/0,412 = 0,429.$$

Leiame ka sündmuse D (õpilane pärineb Järve või Raba külast) tõenäosuse, see on 11/17=0,647. Tuleb välja, et sündmused A ja D ei ole sõltumatud, st tütarlaste jaotus külade vahel on erinev. Tõepoolest, Järve külast on koolis kaheksa õpilast, neist ainult üks tütarlaps.

2.1.9. Sõltumatud katsed ja katseseeria

Mingil viisil korraldatud katset võib tavaliselt nii korrata, et katsetingimused ja katsekorraldus jääb muutumatuks. Kui see nii on, siis on katsed sõltumatud. *Sõltumatuid katseid* võib tavaliselt korrata ükskõik kui palju kordi, saades sel viisil sõltumatute katsete jada ehk *katseseeria.*

Näide 37

Katseseeriad on näiteks:

- *Mündivisete seeria.*
- *Kuuli võtmine urnist, mis sisaldab mitmevärvilisi kuule. Peale kuuli värvi registreerimist pannakse kuul urni tagasi (katsekorralduse jäävuse tagamiseks). Seejärel katset korratakse.*

2.1.10. Sündmuse suhteline sagedus katseseerias. Suurte arvude seadus

Oletame, et oleme määranud mingi katse, ja selle tulemuste kaudu defineerinud sündmuse A, mille tõenäosus $p(A)$ on meil teada. Kui me selle katse sooritame, siis tekib kaks võimalust: sündmus A toimus või ei toimunud. Kordame nüüd katset n korda (katsed on sõltumatud) ja loendame, mitu korda katse A selles katseseerias esines – olgu see arv $k(A)$. Siis sündmuse A esinemise suhteline sagedus katseseerias on suhe $k(A)/n$. Sündmuse suhtelist sagedust ja tõenäosust seob alljärgnev, statistika seisukohalt erakordselt tähtis suurte arvude seadus:

- Katseseeria lõpmatul pikenemisel läheneb sündmuse suhteline sagedus üldiselt tema tõenäosusele,

$$k(A)/n \Rightarrow P(A). \qquad (7)$$

Valemis (7) toimuv *lähenemine* (piirprotsess) ei tähenda sugugi seda, et iga uue katse tegemisel suhteline sagedus aina läheneb tõenäosusele. Igaüks, kes proovib suurte arvude seadust katseliselt kontrollida (kasvõi näiteks münzi viskamise teel) märkab, et lähenemine toimub nõ sakiliselt, mõned katsetulemused nagu "rikuvad" lähenemisprotsessi. Mida rohkem on aga katseid tehtud, seda väiksemaks jääb selliste häirivate katsetulemuste mõju.

See piirprotsess, mis suurte arvude seaduse puhul toimub, kannab *tõenäosuse järgi koondumise* nime.

2.1.11. Statistiline tõenäosus. Tõenäosuse hinnag

Tõenäosust, mis defineeriti punktis 2.1.5, nimetatakse klassikaliseks tõenäosuseks. Tema kasutusala on piiratud nõudega, et katsetulemuste hulk, mille kaudu sündmused defineeritakse, peab olema lõplik ja katsetulemused võrdvõimalikud. Laiemalt saab kasutada *suhtelise sageduse* mõistet:

- Kui sündmus A on defineeritud mingi katse abil (ilma et me katsetulemuste võrdvõimalikkust ja lõplikkust eeldaksime) ning on tehtud n sõltumatust katsest koosnev katseseeria, mille tulemusena sündmus A toimus $k(A)$ korda, siis nimetatakse suhet $k(A)/n$ sündmuse A statistiliseks tõenäosuseks.
- Statistiline tõenäosus muutub üldjuhul kui katseseeria pikkus n muutub.

- Kui katseseeria pikkus on fikseeritud, siis säilitab statistiline tõenäosus kõik klassikalise tõenäosuse omadused (vt p. 2.1.5 ja 2.1.6), kuid sündmust, mille statistiline tõenäosus on 0, ei nimetata võimatuks (sest see võib edaspidi esineda), ja sündmust, mille statistiline tõenäosus on 1, kindlaks, sest see võib edaspidi mitte esineda.
- Kui mingi sündmuse jaoks ei ole võimalik tõenäosust klassikalise eeskirja järgi arvutada, siis nimetatakse statistilist tõenäosust selle sündmuse *tõenäosuse hinnanguks*. Hinnang on juhuslik, see sõltub konkreetsest katseseeriast ning üldiselt saadakse teise katseseeria korral samale sündmusele mõnevõrra erinev tõenäosuse hinnang. Mida pikem on katseseeria, seda täpsem on tavaliselt hinnang ja seda vähem erinevad üksteisest erinevate katseseeriate abil saadud hinnangud.

Toodud üldistus on meile oluline, kuna võime edaspidi kõneleda sündmustest alati, kui need on defineeritavad katseseeriate kaudu, ning omistada neile sündmustele ka tõenäosused, ilma et tarvitseksime kontrollida, kas need on arvutatavad lõpliku arvu võrdtõenäoste katsetulemuste kaudu.

2.1.12. Tõenäosuse kasutamine suhtelise sageduse ennustamisel

Väga tähtis järeldus suurte arvude seadusest on see, et teades mingi sündmuse tõenäosust (või ka selle tõenäosuse hinnangut), võime ennustada selle esinemise suhtelist sagedust.

- Näiteks, kui teame, et mingi sündmuse tõenäosus on 0,1, siis esineb see sündmus keskmiselt kümne katse korral üks kord. Muidugi ei välista see võimalust, et see sündmus esineb esimesel korral.
- Sündmus, mille tõenäosus on 0,01, esineb keskmiselt vaid ühel korral saja katse kohta,
- ja sündmus, mille tõenäosus on 0,001, on juba üpris vähetõenäone: ta esineb keskmiselt üks kord tuhande katse kohta. Kuigi niisuguse tõenäosusega sündmuse esinemine üksikkatsel on äärmiselt vähe tõenäone, oleks viga sellise sündmuse toimumise võimalust eirata pikkade katseseeriate korral.

Kui mingisse haigusesse haigestumise tõenäosus on 0,001, siis on seda üsna lohutav teada igal inimesel eraldi võetuna. Ometigi haigestub 1 000 000 inimesega populatsioonis sellesse haigusesse keskmiselt 1000 inimest, mis pole sugugi väike arv.

Samuti võib korrastada liikluse nii, et igas üksikus kohas iga üksiku auto õnnetusse sattumise tõenäosus kaduvväike. Et aga liiklusest haaratud kohti on palju ja autosid, mis neid läbivad, samuti palju, siis on katsete arv nii suur, et õnnetusi siiski juhtub.

2.2. *Juhuslik suurus*

2.2.1. Diskreetne juhuslik suurus

Oletame, et meil on defineeritud katse, millel on n võrdtõenäost katsetulemust. Niisugust suurust, mille väärtus sõltub selle katse tulemusest (matemaatiliselt öeldes – on *katsetulemuse funktsioon*), nimetatakse *diskreetseks juhuslikuks suuruseks*. Edaspidi näeme, et katsetulemuste võrdtõenäosus ei ole selle definitsiooni juures tarvilik, ning me nimetame *diskreetseks juhuslikuks suuruseks niisugust katsetulemuse funktsiooni, millel on lõplik arv võimalikke väärtusi.*

Ka seda määratlust on võimalik üldistada, lugedes diskreetsete juhuslike suuruste hulka ka sellised, mille väärtuste hulk on loenduvalt lõpmatu (st väärtused on nummerdatavad täisarvuliste järjekorranumbritega, kusjuures suurimat järjekorranumbrit ei ole ette määratud). Käesolevas paragrahvis käsitleme eeskätt diskreetseid juhuslikke suurusi.

Näide 38

Vaatleme katsena täringuviset ja katsetulemusena seda, missugune täringu tahk viskel peale jääb. Loeme juhusliku suuruse X väärtuseks täringuviske tulemusena saadud silmade arvu. Näeme, et X on juhuslik suurus, sest tema väärtuseks on 1, 2, 3, 4, 5 või 6 sõltuvalt sellest, missugune tahk jääb täringuviskel peale.

2.2.2. Diskreetse juhusliku suuruse jaotus

Juhuslikku suurust iseloomustab

- tema väärtuste hulk, mida tavaliselt tähistatatakse $x_1, \ldots, x_k$.
- iga väärtuse esinemise tõenäosus, $P(X = x_i) = p_i$, mille leiame, lugedes kokku nende katsetulemuste arvu, mille korral vastav väärtus esineb.

Juhusliku suuruse üksikväärtuste tõenäosusi saab esitada jaotustabeli, graafiku või valemina, kusjuures alati kehtib võrdus

$$\sum_{i=1}^{k} p_i = 1 .$$

Tabel, graafik ja valem (tõenäosusfunktsioon) on kõik *juhusliku suuruse jaotuse esitused.* Juhusliku suuruse jaotus näitab tema väärtuste esinemise tõenäosusi.

Juhusliku suuruse abil saab defineerida sündmusi ja leida nende tõenäosusi. Tüüpilised juhusliku suuruse abil defineeritud sündmused on järgmised:

- $X = a$, st sündmus, et X väärtus võrdub arvuga a.
- $X < a$ (või $X > a$), st sündmus, et X väärtus on väiksem kui a (või suurem kui a).
- $a < X < b$, st sündmus, et X väärtus on a ja b vahel.

Kui juhusliku suuruse jaotus on teada, on nende sündmuste tõenäosused lihtsalt arvutatavad.

Näide 39

Olgu juhuslikuks suuruseks X juhuslikult valitud lapse kaal näite 11 andmetel. Siis kehtivad järgmised võrdused:

$P(X=65) = 0,176$; $P(X=66)=0$; $P(X> 50) = 0,882$; $P(X<75) =1$; $F(50<X<59)= 0,353$.

2.2.3. Arvtunnus ja diskreetne juhuslik suurus

Iga arvtunnus määrab juhusliku suuruse. Selle juhusliku suuruse väärtusteks on arvtunnuse väärtused, iga väärtuse tõenäosuseks aga selle suhteline sagedus üldkogumis. Kui üldkogum on lõplik, siis on tunnuse kaudu defineeritav juhuslik suurus kindlasti diskreetne.

- Kui uuring on kõikne, on ka uuritava juhusliku suuruse jaotus täielikult teada.
- Kui on tegemist valikuuringuga, siis ei ole tunnuse jaotus üldkogumil (nn *tegelik* ehk *teoreetiline jaotus*) teada, ning teada on vaid valimi põhjal moodustatud nn *valimjaotus* (näiteks peatükis 1.3.4 kirjeldatud jaotustabeli kujul). Valimjaotusele on aga valimi juhuslikkuse tõttu rakendatav suurte arvude seadus, ja me võime öelda, et valimjaotus on *teoreetilise jaotuse hinnanguks*.
- Mida suurem on valimi maht, seda lähedasem on üldiselt valimjaotus teoreetilisele jaotusele.

Väga suur osa statistika ülesannetest seisnebki selles, et püüda kirjeldada ja hinnata mitte teadaolevat üldkogumi jaotust valimjaotuse põhjal.

2.2.4. Jaotusseadus

Igas ülesandes kasutatavad tunnused on üldiselt erinevad, ning ka sama tunnuse jaotus võib erinevate kogumite puhul olla vägagi erinev. Sellest näivast eripärade rohkusest hoolimata on tunnuste jaotustel ka palju ühiseid jooni. Nendele tuginedes on defineeritud teatavad standardsed jaotuseeskirjad, mis kirjeldavad (ligikaudselt) paljudes tüüpilistes olukordades tekkivaid juhuslikke suurusi. Neid nimetatakse *jaotusseadusteks* (ehk

parameetrilisteks jaotuste peredeks). Parameetriks, mis jaotust iseloomustab, on tavaliselt jaotuse mingi arvkarakteristik. Näiteks tunnused, mille jaotused on oma kuju poolest sarnased, kuid erinevad keskväärtuse poolest, kuuluvad ühte jaotuste peresse. Samuti võivad ühte peresse kuuluda ka jaotused, mille dispersioonid on erinevad.

Kui on kindlaks tehtud *tunnuse jaotusseadus*, siis tuleb tunnuse *jaotuse* määramiseks arvutada vaid jaotusparameetriteks olevate arvkarakteristikute väärtused.

Jaotusseadus on uuritava tunnuse *mudel*.

2.2.5. Diskreetse juhusliku suuruse arvkarakteristikud

Juhusliku suuruse ja tunnuse arvkarakteristikud on üldiselt samad, kuid sageli kasutatakse veidi erinevat tähistust. Ka arvutuseeskiri tõenäosusfunktsiooni kaudu on sarnane.

- Keskväärtus EX arvutatakse alljärgneva summana:

$$EX = \sum_{i=1}^{k} x_i p_i \,. \qquad (8)$$

  Paneme tähele, et tähistades $p_i = n_i/n$, saame valemist (8) valemi (1') tunnuse keskmise arvutamiseks. Tulemus on ootuspärane, sest kõikse uuringu puhul ühtib ju tunnuse keskmine vastava juhusliku suuruse keskväärtusega. Sama olukord kehtib ka teiste arvkarakteristikute korral.
- Dispersioon DX on defineeritud keskväärtuse kaudu, $DX = E(X-EX)^2$, kusjuures tema arvutusvalem on

$$DX = \sum_{i=1}^{k} p_i (x_i - EX)^2 \,.$$

- Standardhälve σX arvutatakse ruutjuurena dispersioonist, $\sigma X = \sqrt{DX}$.

Diskreetse juhusliku suuruse defineerimiseks sobib kõige paremini tõenäosusfunktsioon, mis tavaliselt sõltub mõnest jaotusparameetrist. Alljärgnevas esitamegi kaks diskreetset jaotusseadust.

2.2.6. Bernoulli ehk kahe väärtusega jaotus

Kõige lihtsam jaotusseadus määrab nn *Bernoulli jaotuse*, mis sõltub ühest-ainsast parameetrist. Bernoulli jaotusega juhuslik suurus X on defineeritud teatava sündmuse A kaudu:

- kui sündmus A toimub, siis $X=1$;
- kui sündmus A ei toimu, siis $X=0$.

Selle sündmuse tõenäosus $P(A)$ ongi Bernoulli jaotuse parameeter p (mis võib omandada väärtuse nulli ja ühe vahel). Bernoulli jaotusega on lähendatavad kõigi binaarsete tunnuste jaotused. Bernoulli jaotusega juhusliku suuruse arvkarakteristikud on väga lihtsalt arvutatavad:

$$EX = p,\ DX = p(1-p),\ \sigma X = \sqrt{p(1-p)}.$$

2.2.7. Binoomjaotus

Üks kõige sagedamini kasutatavaid diskreetsete jaotuste peresid määrab nn *binoomjaotuse*. Tegemist on kaheparameetrilise jaotuste perega, kus üheks parameetriks on sündmuse A tõenäosus p ja teiseks parameetriks sooritatud katsete arv m. Juhusliku suuruse X väärtuseks loetakse sündmuse A esinemiste arvu katseseerias.

- Juhusliku suuruse X väärtused on täisarvud 0 ja m vahel (otspunktid kaasa arvatud).
- Tõenäosus, et juhuslik suurus X omandab väärtuse k $(0 \le k \le m)$, on arvutatav järgmise valemiga:

$$P(X=k) = \frac{m!}{k!(m-k)!} p^k (1-p)^{m-k}.$$

- Binoomjaotusega juhusliku suuruse keskväärtus on mp.
- Binoomjaotusega juhusliku suuruse dispersioon on $mp(1-p)$.
- Binoomjaotusega juhusliku suuruse standardhälve on $\sqrt{mp(1-p)}$.

Binoomjaotus sobib mudeliks järgmistes ülesannetes:

- Sünnitusmajas sündinud laste hulgast poisslaste (tütarlaste) arv mingil fikseeritud ajaperioodil või mingi fikseeritud arvu sünnituste seast; poisslapse sünni statistiline tõenäosus on teada.
- Mängija C poolt võidetud partiide arv C ja D vahelises fikseeritud pikkusega matšis, kui C võidu tõenäosuse hinnang on teada ja see ei muutu matši vältel.

Märgime, et Bernoulli jaotus on käsitletav ka binoomjaotuse erijuhuna, kui katsete arv $m=1$.

Näide 40

Sünnitusmajas sündis 8 last. Juhuslik suurus X on poiste arv nende hulgas. Lihtsuse mõttes on siin eeldatud, et poisi ja tütarlapse sündimise tõenäosus on võrdne. Juhusliku suuruse X tõenäosusfunktsioon on esitatud tabelis 10. Siit näeme, et tõenäosus selleks, et poiste ja tüdrukute arv on võrdne, on veidi üle veerandi, nimelt 0,2734. Tõenäosus selleks, et ei poiste ega tüdrukute arv pole väiksem kui 3, on üle 70%, vt ka joonist 10.

Tabel 10

Väärtus	*0*	*1*	*2*	*3*	*4*
Tõenäosus	*0,0039*	*0,0312*	*0,1094*	*0,2188*	*0,2734*

Väärtus	*5*	*6*	*7*	*8*
Tõenäosus	*0,2188*	*0,1094*	*0,0312*	*0,0039*

Joonis 10

2.2.8. Pidev juhuslik suurus

Lisaks diskreetsetele juhuslikele suurustele eksisteerib teisigi juhuslikke suurusi, olulisimad neist on *pidevad juhuslikud suurused.* Ka pideva juhusliku suuruse väärtus sõltub katsetulemusest, s.o juhusest. Pidev juhuslik suurus võib omandada väärtusi fikseeritud (lõplikust või lõpmatust) intervallist või ka kogu arvsirgelt. Pideva juhusliku suuruse defineerimiseks kasutatakse *tihedusfunktsiooni* $f(x)$. Tihedusfunktsioonil on järgmised omadused:

- $f(x)$ on mittenegatiivne;
- tihedusfunktsiooni integraal üle kogu juhusliku suuruse määramispiirkonna on 1:

$$\int_{-\infty}^{\infty} f(x)dx = 1;$$

- juhusliku suuruse abil defineeritud sündmuste tõenäosused on leitavad tihedusfunktsiooni integraalina:

$$P(a<X<b)=\int_a^b f(x)dx\,.$$

Ka pidevate juhuslike suuruste jaoks on kirjeldatud mudelina kasutatavaid jaotusseadusi, mis omakorda sõltuvad parameetritest. Tuntuim neist jaotusseadustest on *normaaljaotus.*

2.2.9. Normaaljaotus

Juhuslik suurus X, mille tihedusfunktsioon $f(x)$ on alljärgneva kujuga

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

on normaaljaotusega. Seda tähistatakse järgmiselt: $X \sim N(\mu, \sigma)$. Tihedusfunktsiooni valemist näeme, et normaaljaotusel on kaks parameetrit. Need on

- μ, mis tähistab juhusliku suuruse X keskväärtust;
- σ, mis tähistab juhusliku suuruse standardhälvet.

Konstantidel π ja e on nende tavaline tähendus ja x on juhusliku suuruse väärtus.

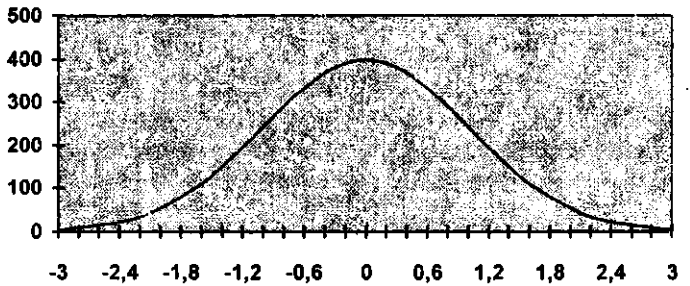
2.2.10. Normaaljaotuse omadused

Normaaljaotusel on järgmised olulised omadused.

- Normaaljaotus on pidev.
- Normaaljaotus on oma keskväärtuse suhtes sümmeetriline.
- Normaaljaotuse tihedusfunktsioon on ühe tipuga (unimodaalne jaotus).
- Normaaljaotuse keskväärtus, mood ja mediaan ühtivad.
- Normaaljaotuse keskväärtuse μ muutmisel nihkub normaaljaotuse tihedusfunktsiooni graafik x-teljel, kusjuures graafiku kuju ei muutu.
- Normaaljaotuse standardhälbe suurenemisel muutub graafik madalamaks ja tema ulatus laieneb, standardhälbe vähenemisel aga ulatus väheneb ja tipp tõuseb kõrgemale (meenutagem, et kõvera alune pindala on konstantselt võrdne ühega).
- Normaaljaotusega juhuslike suuruste summa ja lineaarkombinatsioon on samuti normaaljaotusega.

- Normaaljaotuse väärtuste hulk ei ole tõkestatud, st et põhimõtteliselt iga arv võib olla iga normaaljaotuse väärtuseks. See fakt on aga üksnes teoreetilise tähtsusega, sest tegelikkuses on keskväärtusest kaugel paiknevatesse piirkondadesse sattumise tõenäosus kaduvväike.

Normaaljaotuse tihedusfunktsioon on kujutatud joonisel 11.



Joonis 11

2.2.11. Standardiseeritud normaaljaotus

Erijuhul, kui $\mu = 0$ ja $\sigma = 1$, on tegemist standardiseeritud normaaljaotusega. Kui meil on suvalise normaaljaotusega juhuslik suurus X, kus $X \sim N(\mu, \sigma)$, siis saame sellest standardiseeritud normaaljaotuse järgmiselt:

- tsentreerime juhusliku suuruse X, st lahutame X väärtustest keskväärtuse μ;
- normeerime saadud juhusliku suuruse, st. jagame ta läbi standardhälbega.

Seega uus juhuslik suurus $X_0 = \dfrac{x - \mu}{\sigma}$ on standardiseeritud normaaljaotusega, $X_0 \sim N(0,1)$.

- Standardiseeritud normaaljaotuse tihedusfunktsioon on tabuleeritud.
- Standardiseeritud normaaljaotuse *jaotusfunktsiooni* $P(X < x)$ tähistatakse sümboliga $\Phi(x)$ ja selle väärtused on tabuleeritud (vt tabel 36 Lisas). Normaaljaotuse jaotusfunktsiooni tabeleid on sobiv kasutada normaaljaotusega juhusliku suuruse abil defineeritud sündmuste tõenäosuste arvutamiseks.

- Standardiseeritud normaaljaotusega juhusliku suuruse puhul on kasulik meelde jätta järgmiste sündmuste tõenäosused:

  $P(X < 0) = 0{,}5$; $P(X > 0) = 0{,}5$;

  $P(-1 < X < 1) = 0{,}6827$ ($\approx 68\%$); $P(X < -1) = 0{,}159$ ($\approx 16\%$); $P(X > 1) = 0{,}159$ ($\approx 16\%$);

  $P(-2 < X < 2) = 0{,}9545$ ($\approx 95{,}5\%$); $P(X < -2) = 0{,}0228$ ($\approx 2{,}3\%$); $P(X > 2) = 0{,}0228$ ($\approx 2{,}3\%$);

  $P(-3 < X < 3) = 0{,}9974$ (=99,74%); $P(X < -3) = 0{,}0013$ (=0,13%); $P(X > 3) = 0{,}0013$ (=0,13%);

Standardiseeritud normaaljaotusega juhuslik suurus võib põhimõtteliselt omandada kõikvõimalikke väärtusi, kuid reaalsuses on juba piirkonnast (–3;3) väljaspool asuva väärtuse esinemine väga väikese tõenäosusega. Veelgi kaugemate väärtuste esinemise tõenäosus on seega praktiliselt võimatu sündmus.

Statistikaülesannete lahendamiseks on kasulik pidada meeles ka mõned normaaljaotuse abil defineeritud sündmused, millel on ette antud "ümmargune" tõenäosus.

- $P(X < 1{,}65) = 0{,}95$;
- $P(-1{,}96 < X < 1{,}96) = 0{,}95$;
- $P(-2{,}58 < X < 2{,}58) = 0{,}99$.

Väärtusi 1,65; 1,96 ja 2,58 nimetatakse ka vastavalt 0,95-; 0,975- ja 0,995-kvantiilideks. Rakenduslikus kirjanduses öeldakse nende kohta vahel ka 95%, 97,5% ja 99,5% *protsentpunktid*.

2.2.12. Normaaljaotuse tabelite kasutamine sündmuste tõenäosuste leidmiseks

Normaaljaotuse tabelid on küll koostatud standardiseeritud normaaljaotuse jaoks, kuid lihtsa teisendusega saab neid kasutada ka suvalise normaaljaotusega juhusliku suuruse kaudu defineeritud sündmuste tõenäosuste arvutamiseks. Selleks arvestame järgmisi seoseid.

- Kui $X \sim N(\mu, \sigma)$, siis $X_0 = (X-\mu)/\sigma$ on standardiseeritud normaaljaotusega, $X_0 \sim N(0,1)$.
- Kui $X_0 \sim N(0,1)$, siis $X = \sigma X_0 + \mu$ on normaaljaotusega $X \sim N(\mu, \sigma)$.
- Sündmused $\{X < a\}$ ja $\{(X-\mu)/\sigma < (a-\mu)/\sigma\}$ ühtivad, seega on nende tõenäosused võrdsed.

Näide 41

Olgu eesti mehe keskmine pikkus 180 cm standardhälbega 8 cm. Kui suur on tõenäosus, et juhuslikult valitud mees osutuks enam kui 196 sentimeetri pikkuseks? Enam kui 2 m pikkuseks?

Leiame suurused (196-180)/8 = 2 ja (200 – 180)/8 = 2,5 ning seejärel kasutame normaaljaotuse tabelit. Sümmeetria tõttu kehtib võrdus

$$P(X>2)=P(X<-2)=\Phi(-2)=0{,}0282 \text{ ja } P(X>2{,}5)=\Phi(-2{,}5)=0{,}0062.$$

Seega näeme, et tõlgendades leitud tõenäosusi sagedustena, peaks iga saja mehe kohta tulema ligi kolm üle 196 sentimeetri pikkust, kuid üle 2-meetrise kasvuga on vaid üks mees saja kuuekümne hulgast.

2.2.13. Piirteoreem. Normaaljaotus kui mudel

Kui me vaatleme binoomjaotuse tulpdiagrammi ja normaaljaotuse tihedusfunktsiooni graafikut (vt jooniseid 10 ja 11), siis märkame nende kuju sarnasust. See sarnasus suureneb siis, kui katsete arv m kasvab ja sobivalt muudetakse skaalasid telgedel (vaadeldakse tunnuse hälbeid keskmisest, kasutades horisontaalteljel standardhälbe ühikuid ja vertikaalteljel suhtelisi sagedusi). Üks tõenäosusteooria tähtsamaid tulemusi, nn *klassikaline piirteoreem*, väidab, et katsete arvu m suurenemisel läheneb binoomjaotusega juhusliku suuruse jaotus normaaljaotusele.

Kehtib veelgi üldisem teoreetiline tulemus: kui liidetakse ühesuguse, kuid suvalise jaotusega sõltumatuid juhuslikke suurusi, ka siis läheneb summa jaotus (peale selle standardiseerimist) standardiseeritud normaaljaotusele. See tõsiasi võimaldab normaaljaotusega juhuslikke suurusi laialdaselt kasutada mudelina mitmesuguste praktiliste ülesannete lahendamisel.

2.2.14. Normaaljaotusega juhusliku suuruse identifitseerimine

Kui meil on alust oletada, et mingi tunnuse jaotus üldkogumis on normaaljaotusega, siis tuleb jaotuse määramiseks kindlaks teha selle tunnuse keskväärtus μ ja standardhälve σ, ning sellega ongi üldkogumi jaotus täielikult teada. Praktikas aga ei ole tunnuse parameetrite õiged väärtused teada, ning nende asemel kasutatakse hinnanguid $\bar{x}$ ja s.

3. Parameetrite hindamine ning hüpoteeside kontrollimine

3.1. Matemaatilise statistika põhiülesanne – valimi põhjal üldkogumi kohta järelduste tegemine

Meenutame, et uurimisobjekti, mille kohta järeldusi soovitakse teha, nimetatakse matemaatilises statistikas *üldkogumiks*. Statistiline uuring võib aga olla kas *kõikne* (uuritakse läbi kogu üldkogum) või *valikuline* (uuritakse läbi üldkogumit esindav osa, nn valim). Kõikse uuringu tulemused on lõplikud, neid ei saa kasutada edasisteks üldistusteks. Valikuuringu eesmärgiks on aga valimi põhjal järelduste tegemine üldkogumi kohta, valikuuringute tulemuste üldistamine. Selleks, et tehtavad üldistused oleksid korrektsed ja veenvad, on matemaatilises statistikas välja töötatud teatavad mängureeglid üldistuste tegemiseks. Nende reeglite järgimine ei taga küll 100%-liselt eksimatuid tulemusi (see pole valikuuringute korral tavaliselt põhimõtteliselt võimalik), kuid hoiab ära jämedad vead ja garanteerib tulemuste õigsuse näiteks 99% juhtudest.

Käesolevas peatükis käsitletakse eranditult ainult valikuuringuid. Kõik need ülesanded, mida siin lahendatakse, on mõttekad üksnes siis, kui valimi abil on tarvis teha järeldusi üldkogumi kohta. Kõikse uuringu puhul pole mõtet ega vajadust konstrueerida usalduspiire ega kontrollida statistilisi hüpoteese, sest kõikse uuringu puhul on üldkogumi jaotus ja selle parameetrid ju täpselt teada.

3.1.1. Üldised eeldused

Selleks, et töötada välja eeskirjad valikuuringute tulemuste üldistamises, tuleb kõigepealt fikseerida teatavad eeldused. Käesolevas peatükis me kasutame järgmisi eeldusi:

- Me eeldame, et üldkogumis on mõõdetud arvtunnus X. Selle tunnuse jaotust (nn *teoreetilist jaotust*) me ei tea.
- Meil on olemas esindav valim selle tunnuse väärtustest üldkogumis $(x_1, ..., x_n)$. Nagu varem kokku lepitud, on see valim *lihtne juhuslik valim*, kusjuures n tähistab *valimi mahtu*.
- Meil võib olla tunnuse X jaotusseadus teada või mitte teada.
- Üldkogum võib olla kas lõplik (mahuga N, $N>n$) või lõpmatu, kusjuures see, kumma juhuga on tegemist, on meil teada.

Näeme, et tehtud eeldused on väga loomulikud ega esita mingeid erilisi kitsendusi. Kõige tähtsam on nende eelduste puhul see, et *valim peab olema üldkogumi jaoks tõepoolest esindav*, st et igal üldkogumi objektil on võrdne tõenäosus valimisse sattuda.

3.1.2. Milliseid järeldusi üldkogumi kohta tehakse?

Statistikas ollakse seisukohal, et kui teatakse tunnuse jaotust, siis on tunnuse kohta kõik teada. Kui tunnuse kohta on eeldatud, et ta kuulub mingisse jaotuste peresse, siis on tema jaotuse kindlakstegemiseks tarvis määrata vaid ühe või paari arvkarakteristiku väärtused. Märgime, et matemaatilises statistikas nimetatakse neid arvkarakteristikuid tavaliselt parameetriteks. Näiteks, kui on teada, et tunnus on normaaljaotusega, siis on tema jaotuse täpseks kindlaksmääramiseks vaja leida selle tunnuse keskväärtuse ja standardhälbe väärtused, sest need karakteristikud on normaaljaotuse parameetriteks.

- Siit tuleneb esimene matemaatilise statistika ülesanne – *parameetrite hindamine valimi põhjal.*

Kuivõrd valimi põhjal leitud hinnangud üldiselt erinevad parameetri *õigest* väärtusest, siis on sageli otstarbekas lisaks arvulisele hinnangule ehk nn *punkthinnangule* leida parameetri jaoks ka *vahemikhinnang,* so vahemik, millesse hinnatava parameetri õige väärtus tõenäoliselt kuulub.

- Teine matemaatilise statistika ülesanne on *parameetrite vahemik-hindamine.*

Tihti on tarvis jaotuste või ka jaotusparameetrite kohta teha järeldusi, mis kas kinnitavad või kummutavad sisuliselt olulisi hüpoteese üldkogumi kohta. Niisugusteks hüpoteesideks võivad olla niihästi praktilised järeldused kui ka teaduslike teooriate tuletised.

- Kolmas matemaatilise statistika ülesanne on *statistiliste hüpoteeside kontrollimine.*

3.1.3. Valimi juhuslikkus

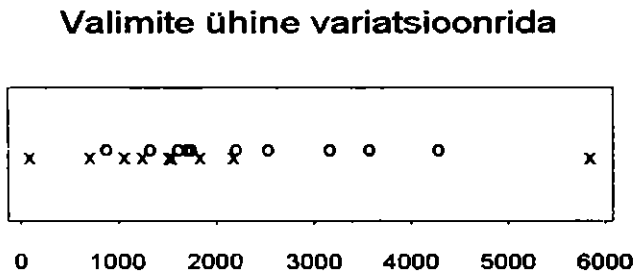
Valimi põhjal järelduste tegemisel on peamiseks probleemiks *valimi juhuslikkus.* Kuigi esindava valimi jaotus on tavaliselt küllaltki lähedane üldkogumi jaotusele, on üksikud valimid (eriti siis, kui valimi maht on suhteliselt väike) üksteisest üsna erinevad.

Näide 42

Näites 29 vaatlesime Eesti valdade jaotust elanike arvu järgi. Teeme nüüd kaks juhuslikku 10-objektilist valimit kõigi valdade hulgast. Need valimid on alljärgnevad (sulgudes on näidatud tunnuse X väärtus – elanike arv 1. jaanuaril 1996).

- *Emmaste (1524), Jõgeva (5846), Laekvere (2178), Misso (1068), Muhu (2180), Oisu (1548), Piirissaare (95), Valgjärve (1840), Varbla (1254), Öru (703).*
- *Kohtla (1614), Koigi (1321), Käina (2532), Lohusuu (886), Mäksa (1708), Märjamaa (3578), Põlva (4287), Ridala (3164), Sangaste(1750), Sõmerpalu (2209).*

Valimite võrdlemiseks jälgime nende ühist variatsioonrida, kus esimese valimi punktid on tähistatud sümboliga x ja teise valimi punktid sümboliga o (vt joonis 12).



Joonis 12

- *Näeme, et esimesse valimisse kuulub nii minimaalne väärtus (Piirissaare) kui ka maksimaalne väärtus (Jõgeva), variatsiooniulatus (maksimumi ja miinimumi vahe) on esimeses valimis 5751, teises valimis aga 1,7 korda väiksem – ainult 3401.*
- *Arvutame kummagi valimi mediaani. Esimese valimi mediaani saame Emmaste ja Oisu keskmisena, s.o 1536. Teise valimi mediaaniks on Sangaste ja Sõmerpalu keskmine 1979,5.*
- *Vaatame nüüd ka seda, kuidas paikneb ühe valimi mediaan teise valimi variatsioonreas. Näeme, et esimese valimi mediaanist on teises valimis väiksemaid punkte 2 ja suuremaid 8, teise valimi mediaanist aga esimeses valimis väiksemaid 7 ja suuremaid 3.*
- *Kvartiilhälve (Laekvere ja Misso valdade elanike arvu erinevus) on esimeses valimis 1110, teises valimis aga on kvartiilhälve 1550, järelikult on see vahekord vastupidine variatsiooniulatuse suurusvahekorrale.*

Need tähelepanekud näitavad, et valimid on oma iseloomult üsna erinevad, kuigi pärinevad samast üldkogumist ja on moodustatud lihtsa juhuvalikuna.

3.2. *Arvkarakteristikute hindamine*

3.2.1. Miks on arvkarakteristikuid vaja hinnata?

Nagu nägime, on igasuguse statistilise uurimise eesmärk – teha järeldusi üldkogumi kohta. Üks lihtsamaid taolisi järeldusi on näiteks alljärgnev väide – *tunnuse X keskmine uuritavas kogumis on* $\bar{x}$. See *tunnus X* võib olla perekonna summaarne sissetulek, laste arv peres, vastsündinu kaal, kuu aja jooksul tarbitud liha hulk jpm. Uuritavaks kogumiks võib olla kogu Eesti elanikkond mingil fikseeritud ajal, kuid samuti ka mingi linna või valla elanikkond. Teistsuguse probleemiseade korral võib uuritava üldkogumi ja tunnuse iseloom olla hoopiski erinev, näiteks võib selleks olla Peipsi järve kalade kaal, taevakehade mass teatavas maailmaruumi piirkonnas jpm.

Kui kõikset uuringut pole võimalik või otstarbekas korraldada (nii on see aga valdaval enamikul juhtudest), siis tuleb huvipakkuvat arvkarakteristikut *hinnata*, kasutades selleks valimi andmeid. Selleks, et hindamine oleks võimalikult erapooletu ja korrektne, on hindamiseks välja töötatud üldtunnustatud, kuid ühtlasi piisavalt paindlikud reeglid.

Hindamisreeglite väljatöötamine ja saadud hinnangute iseloomustamine ongi matemaatilise statistika ühe osa – *hinnangute teooria* – ülesanne. Järgmises punktis tutvustamegi neid matemaatilisi eeldusi ja mõisteid, millele hinnangute teoorias tuginetakse, ja veendume, et sisuliselt on kõik need mõisted meile juba tuttavad.

3.2.2. Punkthinnang

Esimene ülesanne on tavaliselt uuritava tunnuse arvkarakteristikute ehk parameetrite *hindamine*, kusjuures silmas peetakse nende nn *teoreetilisi* ehk *tegelikke väärtusi*, st väärtusi üldkogumis. Enamasti on kõige olulisemad parameetrid *keskväärtus* ja *dispersioon*. Iga parameetri hindamise tulemusena saadakse arv, mida nimetatakse *punkthinnanguks* (lihtsuse mõttes sageli ka lihtsalt *hinnanguks*). Et hinnang arvutatakse valimi põhjal ja valim on võetud juhuslikult, siis sõltub ka hinnangu väärtus juhusest, lühidalt – *hinnang on juhuslik suurus*. Hinnangu juhuslikkus väljendub selles, et samast üldkogumist võetud erinevate ühesuuruste valimite põhjal saadud hinnangute väärtused on tavaliselt erinevad.

Punkthinnanguid tähistatakse enamasti sama tähega nagu hinnatava parameetri *õiget väärtust* üldkogumis (mida me ei tea), lisades juurde mingi täiendava sümboli – tärnikese, katuse või laine. Teeme seda meiegi, kusjuures defineerime iga sümboli just siis, kui meil seda tarvis on.

Hinnangute arvutamisel saame kasutada vaid meile teadaolevat informatsiooni ehk tunnuse väärtusi valimis. Uuritavat tunnust tähistame edaspidi tähega X, valimi mahtu tähistame tähega n ning tunnuse väärtused valimis tähistame $x_1, ..., x_n$.

3.2.3. Punkthinnangu nihe. Nihketa hinnang

Et hinnang üldiselt hinnatava parameetri *õigest väärtusest* erineb, siis tekib küsimus, kas hinnangus sisaldub mingi *süstemaatiline viga*, st kas hinnang enamasti üle- või alahindab hinnatava parameetri õiget väärtust. Et seda kindlaks teha, on kasutusele võetud hinnangu *nihke* mõiste.

Olgu m hinnatava parameeteri õige väärtus ja m^* mingi eeskirja järgi arvutatav hinnang sellele parameetrile. Et m^* on juhuslik suurus, võime arvutada tema keskväärtuse Em^*. Hinnangu keskväärtuse ja hinnatava parameetri vahet b,

$$b = Em^* - m, \qquad (9)$$

nimetatakse hinnangu *nihkeks*. Kui nihe on positiivne, siis ülehindab hindamiseeskiri m^* hinnatavat parameetrit, kui nihe on negatiivne, siis alahindab. Ütleme ka, et neil juhtudel sisaldab hinnang m^* *süstemaatilist viga*. Kui aga kehtib võrdus

$$Em^* = m,$$

siis öeldakse, et hinnang m^* on *nihketa*, ta on keskmiselt õige. Loomulikult võib iga konkreetse valimi korral saadav hinnang erineda parameetri õigest väärtusest, olles kas suurem või väiksem. Kuid nihketa hinnangu puhul on tavaliselt need juhud, kus hinnang on õigest parameetri väärtusest suurem, ja need juhud, kus hinnang on õigest parameetrist väiksem, omavahel tasakaalus.

Nihketus on üks hinnangute olulisemaid omadusi. Kui võimalik, kasutatakse alati nihketa hinnanguid.

3.2.4. Hinnangu bajuvus ja täpsus

Kuna parameetri hinnang on juhuslik suurus, nagu räägitud eelmistes punktides, siis on võimalik lisaks tema keskväärtusele (mida me kasutasime valemis (9) hinnangu nihke määratlemisel) leida ka *hinnangu hajuvust*. Hinnangu hajuvuse kindlakstegemisel on oluline praktiline tähtsus, sest *mida vähem hinnang parameetri õige väärtuse ümber hajub, seda täpsem see hinnang on.*

Hinnangu hajuvus sõltub oluliselt valimi mahust. Enamiku praktikas kasutatavate hinnangute puhul on hinnangu dispersioon pöördvõrdeline valimi mahuga ja hinnangu standardhälve – ruutjuurega valimi mahust. See tähendab, et näiteks valimi neljakordsel suurendamisel väheneb hinnangu dispersioon neli korda ja standardhälve kaks korda. Et just standardhälve on aluseks hinnangu täpsuse määramisel, tulenebki siit reegel: valimi mahu K-kordne suurendamine suurendab hinnangute täpsust $\sqrt{K}$ korda. Oluline on siinjuures see, et valimi mahtu suurendades võib hinnangu teha nii täpseks kui vaja.

3.2.5. Üldkogumi keskväärtuse hinnang

Olgu X arvtunnus, mille jaotus üldkogumil on tundamatu, ja olgu antud esindav valim x_1 ,..., x_n selle tunnuse väärtustest. Kõige tavalisem viis arvtunnuse X keskväärtuse EX hindamiseks on valimkeskmise $\bar{x}$ kasutamine, vt ka valem (1):

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

Tuleb meeles pidada, et erinevad valimid annavad ühele ja samale üldkogumi keskväärtusele enamasti erineva väärtusega hinnangu.

Samuti saab ühele ja samale parameetrile ühe ja sama valimi põhjal leida ka mitmesuguseid hinnanguid sõltuvalt *erinevatest hindamiseeskirjadest*. Näiteks, kui üldkogumi jaotus on sümmeetriline, siis ühtivad juhusliku suuruse mediaan ja keskväärtus ning keskväärtuse hinnanguna saab kasutada ka üldkogumi mediaani hinnangut – valimmediaani.

3.2.6. Normaaljaotusega juhusliku suuruse valimkeskmise jaotus

Järgmiseks ülesandeks, mida me proovime lahendada, on *hinnangu kui juhusliku suuruse jaotuse leidmine*. See jaotus on teoreetiline ning sellele vastavat üldkogumit polegi nii lihtne ette kujutada (see on kõikvõimalike antud mahuga valimite põhjal arvutatud hinnangute kogum). Õnneks on aga mõnel praktikas olulisel erijuhul võimalik hinnangu jaotust teoreetiliselt arvutada. Nii on see näiteks normaaljaotusega tunnuse valimkeskmise puhul.

Olgu X juhuslik suurus, mille jaotus üldkogumis (tegelik jaotus) on normaaljaotus, $X \sim N(\mu, \sigma)$. Valimi juhuslikkust ja esindavust arvestades on üsna lihtne tõestada, et siis on ka valimkeskmine normaaljaotusega,

$$\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}}).$$

Ülal kirja pandud järeldusest tulenevad mõningad olulised tõsiasjad.

- Valimkeskmine $\bar{x}$ on üldkogumi keskmisele nihketa hinnanguks, st et valimi keskmise kasutamisel keskväärtuse hinnanguna ei teki süstemaatilist viga, $\bar{x}$ ei ala- ega ülehinda keskväärtust.
- Valimkeskmise standardhälve on lihtsalt arvutatav üldkogumi standardhälbe järgi, millest ta on $\sqrt{n}$ *korda väiksem.*

Siit järeldub näiteks, et hindamistäpsuse suurendamiseks kolm korda tuleb valimi mahtu suurendada üheksa korda. Oluline on aga see, et valimi mahu suurendamisel võib hindamise teha kuitahes täpseks.

3.2.7. Juhusliku suuruse valimkeskmine üldkogumi suvalise jaotuse korral

Eelmises punktis normaaljaotusega juhusliku suuruse kohta saadud tulemus kehtib ligilähedaselt ka üldkogumi suvalise jaotuse korral. Selle aluseks on piirteoreem: valimkeskmise arvutamisel liidetakse kokku valimi üksikelementide väärtused, mis on ju kõik sama jaotusega. Järelikult on (peale sobivat normeerimist) saadud summa ja järelikult ka aritmeetiline keskmine ligikaudselt normaaljaotusega. See tulemus on statistika rakenduste seisukohalt väga oluline:

- praktiliste ülesannete lahendamisel kasutatakse valimkeskmise jaotuse mudelina normaaljaotust, mille keskväärtuseks on üldkogumi keskväärtus μ ja standardhälbeks $\sigma/\sqrt{n}$;
- mida suurem on valimi maht n, seda parem see mudel üldiselt on.

Näide 43

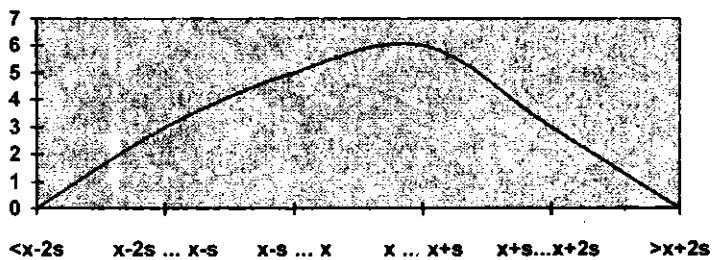
Näites 29 kirjeldatud andmestikust – Eesti valdade elanike arvud – tehti seitseteist valimit. Seitsmeteistkümne 20-objektilise näitevalimi põhjal saime keskväärtuste hinnangutest alljärgneva variatsioonrea:

1822, 1849, 1843, 1958, 2012, 2054, 2141, 2153, 2216, 2252, 2317, 2376, 2413, 2433, 2565, 2567, 2679

Valimkeskmiste keskmine on 2214,706; mediaan 2216 ja standardhälve 269,5. Valimkeskmiste jaotustabel (kasutades nn standardhälbe ühikuid) on esitatud tabelina 11 ja vastav jaotuskõver joonisel 13.

Tabel 11

vahemik s-ühikutes	*<x-2s*	*x-2s...x-s*	*x-s...x*	*x...x+s*	*x+s...x+2s*	*>x+2s*
sagedus	*0*	*3*	*5*	*6*	*3*	*0*



Joonis 13

Näeme, et valimkeskmiste keskmine (2214,7) on tõepoolest väga lähedane tegelikule keskmisele.

Valimkeskmiste standardhälve peaks teoreetiliselt olema $\sqrt{17}$ = 4,123 korda üldkogumi standardhälbest väiksem, seega peaks olema 321,3; on aga 269,5. Mediaanide keskmine on 1875 (teoreetiliselt 1893,5). Üldiselt on kooskõla teooriaga hea. Paneme tähele ka seda, et valimkeskmise jaotus on visuaalse hinnangu järgi märksa lähedasem normaaljaotusele, kui oli vallaelanike jaotus (vt. joonist 9).

3.2.8. Dispersiooni hinnang. Nihke parandamine

Punktis 1.4.10 tutvusime kahe erineva valemiga dispersiooni arvutamiseks – üks neist oli *kõikse uuringu* andmetel leitav *üldkogumi dispersioon* σ^2 (vt valemit (3)) ja teine *valimi dispersioon* s^2, mille eeskiri on antud valemiga (4). Selgitust erinevatele dispersiooniavaldistele me selles punktis ei andnud.

Kui eeldada esindava juhusliku valimi olemasolu, siis näib igati loogiline olevat kasutada σ^2 hinnangu arvutamiseks samuti valemit (3). Paraku selgub, et sel viisil arvutatud hinnang alahindaks üldkogumi dispersiooni. *Nihe* tuleneb sellest, et me ei tea tegelikku keskväärtust ja kasutame selle asemel hinnangut $\bar{x}$. Valimi keskmine aga sõltub valimi iseärasustest, ja selletõttu on hälbed valimi keskmisest keskmiselt mõnevõrra väiksemad, kui oleksid hälbed tegelikust keskmisest. Siit järeldub ka see, et valemiga (3) arvutatud dispersioonihinnang on liiga väike. Valem (4), mis annab dispersioonihinnangu

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

võimaldab leida üldkogumi dispersioonile *nihketa hinnangu* ning selletõttu nimetataksegi sellega määratud statistikut s^2 *valimdispersiooniks.*

On selge, et suurte ($n > 100$) valimite korral on hinnangute σ^2 ja s^2 erinevus tühine, kuid väikeste valimite puhul on see oluline.

Valikuuringu korral arvutatakse standardhälve s kui ruutjuur valimdispersioonist s^2. See on tundmatu üldkogumi standardhälbe jaoks väga väikse nihkega hinnang, mis kõiki praktilise kasutamise eesmärke rahuldab.

Näide 44

Leiame näites 20 vaadeldud kogumi jaoks dispersiooni hinnangu, kasutades juba leitud keskväärtust $\bar{x}=28$*:*

$$s^2 = 1/4\{(-9)^2 + (-8)^2 + (-8)^2 + (-7)^2 + 32^2\} = 320,5.$$

Veendume, et võrreldes näites 24 leitud dispersiooniga $\sigma^2 = 256,4$ *on erinevus üsna suur. Selle põhjuseks on muidugi väga väike valimi maht.*

Märgime, et kuna valimdispersioon hindab üldkogumi dispersiooni, siis valimi suurenedes valimdispersioon küll mõnevõrra muutub, kuid ta ei suurene ega vähene süstemaatiliselt.

3.2.9. Valimi keskväärtuse hinnangu dispersioon. Standardviga

Valimi dispersiooni abil saame leida ka *valimi keskväärtuse* $\bar{x}$ *dispersiooni hinnangu,*

$$s^2_{\bar{x}} = \frac{s^2}{n}, \qquad (10)$$

kus valimdispersioon s^2 on arvutatud valemist (4).

Ruutjuurt valimi keskväärtuse e valimikeskmise dispersioonist nimetatakse *standardveaks*. Standardviga kasutatakse väga tihti mitmesugustes statistilistes otsustustes, ning sageli tähistatakse seda tähega *m*.

Mõnevõrra teistsuguse hinnangu saame aga valimi keskmise dispersioonile siis, kui meil on tegemist lõpliku üldkogumiga, mille mahtu N me teame. Sel juhul sõltub valimkeskmise dispersioon ka üldkogumi mahust ning me saame valemi (10) asemele alljärgneva valemi

$$s_{\bar{x}}^2 = \frac{1}{n}(1-\frac{n}{N})s^2. \qquad (11)$$

Seda valemit on mõtet kasutada siis, kui valimi maht on üldkogumiga võrreldes suhteliselt suur. Näiteks, kui on tegemist 10%-lise valimiga, siis väheneb valemi (11) kasutamisel valimkeskmise dispersiooni hinnang 10% ja standardvea hinnag 5% võrra.

3.2.10. Hinnangu efektiivsus.

Hindamisel on võimalik kasutada mitmesuguseid hindamiseeskirju. Näiteks keskväärtuse hindamiseks võib kasutada lisaks valimkeskmisele ka suurust $0{,}5(max+min)$. Sümmeetrilise tunnuse korral kasutatakse tihti keskväärtuse hindamiseks valimi mediaani. Tekib küsimus, missugust hinnangufunktsiooni kasutada siis, kui nende hulgast on võimalik valida.

Üks oluline nõue, millest juba juttu oli, on nihketus, st süstemaatilise vea puudumine. Teine nõue on, et ka hinnangu hajuvus (st dispersioon) peab olema võimalikult väike. Mida väiksem on hinnangu dispersioon, seda *efektiivsem* ta on. Hinnangut, mis on nihketa ja millel on antud valimi mahu korral väikseim võimalik dispersioon, nimetatakse *efektiivseks hinnanguks*.

Efektiivse hinnangu kasutamine on otstarbekas, kuna võimaldab saavutada vajalikku täpsust kõige väiksema vaatluste arvuga, st kõige odavamalt. Kui mingi hinnang ei ole efektiivne ja tema dispersioon on 20% võrra suurem kui efektiivse hinnangu dispersioon, siis kulub selle hinnangu kasutamisel sama täpsuse saavutamiseks 20% võrra rohkem vaatlusi.

Keskväärtuse hinnang $\bar{x}$ on efektiivne.

3.3. *Jaotusparameetrite vahemikhinnangud*

3.3.1. Vahemikhinnangu mõiste

Et valimi põhjal leitud parameetri hinnang on juhuslik ega ühti tavaliselt parameetri õige väärtusega, on otstarbekaks osutunud *vahemikhinnangu* kasutusele võtmine. Vahemikhinnang on valimi põhjal määratud *vahemik, mis katab õige parameetri väärtuse antud (küllalt suure) tõenäosusega.* Kõige sagedamini kasutatakse vahemikhindamisel *usaldusvahemikku*, mille otspunktideks on vastavalt *alumine* ja *ülemine usalduspiir*. Kui b on hinnatav parameeter, siis on tema usaldusvahemikuks vahemik $(\underline{b}, \bar{b})$, kus usalduspiirid on määratud seosega

$$P(\underline{b} < b < \bar{b}) = 1 - \alpha \qquad (12)$$

Hinnatava parameetri õige väärtus on b. Tõenäosust $1-\alpha$ nimetatakse *usaldusnivooks*. Usaldusnivoo määratluses esinev väike positiivne arv α kannab nimetust *olulisuse nivoo*, ning tema väärtuseks on tavaliselt kas 0,01 või 0,05.

Harvem kasutatakse praktilistes ülesannetes *ühepoolseid* usalduspiire. *Ühepoolne alumine* $(1-\alpha)$-usalduspiir $\underline{b}'$ on defineeritud seosega

$$P(\underline{b}' < b) = 1 - \alpha \qquad (12')$$

ja *ühepoolne ülemine* $(1 - \alpha)$-usalduspiir seosega

$$P(b < \bar{b}') = 1 - \alpha. \qquad (12'')$$

Näiteks võib pakkuda huvi supelranna vee bakterite sisalduse ülempiir, mida annab siis võrrelda etteantud standardiga.

Lihtne on näha, et ühepoolne ülemine $(1-\alpha)$-usalduspiir ja ühepoolne alumine $(1-\alpha)$-usalduspiir moodustavad koos $(1-2\alpha)$-usaldusvahemiku.

3.3.2. Normaaljaotuse keskväärtuse usalduspiirid (suure valimi korral)

Normaaljaotuse keskväärtuse μ usalduspiiride konstrueerimisel kasutatakse tõsiasja, et normaaljaotusega tunnuse korral on valimi keskmine $\bar{x}$ samuti normaaljaotusega, kuid tema dispersioon on n korda väiksem üldkogumi dispersioonist σ^2. Kasutades normaaljaotuse tabeleid, saamegi lihtsalt konstrueerida alljärgneva vahemikhinnangu:

$$P(\bar{x} - q\,\sigma/\sqrt{n} < \mu < \bar{x} + q\,\sigma/\sqrt{n}) = 1 - \alpha,$$

- kus $\bar{x} - q\,\sigma/\sqrt{n}$ on alumine ja $\bar{x} + q\,\sigma/\sqrt{n}$ ülemine usalduspiir,
- $\sigma/\sqrt{n}$ on valimkeskmise standardhälve ehk standardviga,
- q on konstant, mis sõltuvalt α väärtusest leitakse normaaljaotuse tabelist (vt. tabel 36 ja punkt 2.2.11). On kasulik meelde jätta, et

  kui α =0,05 ja $1-\alpha = 0{,}95$, siis $q = 1{,}96$;

  kui $1 - \alpha = 0{,}99$, siis $q = 2{,}58$.

Esitatud valemit saab kasutada siis, kui üldkogumi dispersioon σ^2 on teada (mis on ebareaalne eeldus) või kui valimi maht on küllalt suur, näiteks $n>100$. Suure valimi korral võime nimelt kasutada σ asemel tema hinnangut s.

3.3.3. Tõenäosuse p usaldusvahemik

Väga sageli esinevaks praktiliseks ülesandeks on mingi sündmuse A tõenäosuse $p=P(A)$ hindamine katseseeria põhjal, milles sooritatakse n katset. Kui sündmus A toimus k katse korral, siis on tõenäosuse p hinnanguks selle suhteline sagedus k/n, vt punkt 2.1.11. Et leida eeskirja p usaldusvahemiku leidmiseks, arutleme järgnevalt.

Vaatleme juhuslikku suurust X, mis on defineeritud kui sündmuse A esinemiste arv katseseeria jooksul. Iga katse defineerib ühe Bernoulli jaotusega juhusliku suuruse ning katsed on sõltumatud. Bernoulli jaotusega juhusliku suuruse keskväärtus on p ja standardhälve $\sqrt{np(1-p)}$, vt punkt 2.2.6. Kui sündmus A toimus k korda, siis on juhuslike suuruste summa k ja valimi keskmine k/n ning standardviga

$$s_p = \sqrt{\frac{p(1-p)}{n}} \approx \frac{1}{n}\sqrt{\frac{k(n-k)}{n}}. \qquad (13)$$

Viimane avaldis on saadud, asendades tõenäosuse p tema hinnanguga. Kasutades punktis 2.2.13 märgitud tõsiasja, et ühesuguse jaotusega juhuslike suuruste summa läheneb normaaljaotusele, saame usalduspiiride arvutamiseks kasutada normaaljaotuse tabelit.

Seega saame kokkuvõttes:

$$P\left(k/n - \frac{q}{n}\sqrt{\frac{k(n-k)}{n}} < p < k/n + \frac{q}{n}\sqrt{\frac{k(n-k)}{n}}\right) = 1-\alpha. \qquad (14)$$

3.3.4. Studenti *t*-jaotus

Eelmistes punktides kasutasime me usalduspiiride arvutamisel tõsiasja, et kuna valimi keskmine $\bar{x}$ on normaaljaotusega $N(\mu, \frac{\sigma}{\sqrt{n}})$, siis peale tema tsentreerimist ja normeerimist (vt punkt 2.1.11) saame standardiseeritud normaaljaotusega juhusliku suuruse

$$(\bar{x}-\mu)\frac{\sqrt{n}}{\sigma} \sim N(0,1) \qquad (15)$$

Probleemiks on siin aga see, et valemis sisalduv õige standardhälve σ pole üldiselt teada ja selle asemel kasutatakse praktikas hinnangut. Eelmistes punktides toodud valemid on kasutatavad siis, kui valimi maht on küllalt suur, ning me võime "unustada", et me oleme üldkogumi dispersiooni σ^2 asendanud selle hinnanguga valimi põhjal s^2. Väikeste valimite puhul aga nii teha ei tohi, sest siis on standardhälbe hinnang s väga hajuv, ja juhusliku suuruse

$$t = (\bar{x}-\mu)\frac{\sqrt{n}}{s} \qquad (16)$$

jaotus erineb märgatavalt standardiseeritud normaaljaotusest. Erinevus on seda suurem, mida väiksem on valimi maht n.

Valemiga (16) defineeritud juhusliku suuruse jaotust nimetatakse Studenti t-jaotuseks. Studenti t-jaotus on tegelikult *jaotuste pere*, mis sõltub ühest täisarvulisest parameetrist f, mida nimetatakse *vabadusastmete arvuks*. Ülesande niisuguse püstituse puhul, nagu me käesolevas punktis vaatleme, on vabadusastmete arv tavaliselt ühe võrra väiksem valimi mahust, st $f = n - 1$. Üldiselt võib f omandada suvalisi täisarvulisi väärtusi. Studenti

t-jaotus on sümmeetriline ja läheneb vabadusastmete arvu suurenedes normaaljaotusele. Tema tihedusfunktsioon on kujult üsna sarnane normaaljaotusega ka väikeste vabadusastmete arvude korral (visuaalselt on vahet teha raske harjunudki silmal), kuid sabad on pisut raskemad, ning sellest tulenevalt on nn protsentpunktid erinevad. Kuna t-jaotust kasutatakse praktiliste ülesannete lahendamisel väga sageli, siis on tema kohta koostatud tabelid, (vt Lisa, tabel 37). Märgime, et tabelis on käsitletud mitte üht jaotust, vaid kahtekümmet erinevat jaotust, mis kõik kuuluvad t-jaotuste perre. Need jaotused vastavad erinevatele vabadusastmete arvudele, mis on märgitud tabeli esimesse veergu. Usalduspiiride arvutamisel on vabadusastmete arv ühe võrra väiksem valimi mahust n. Et tabelis paiknevad arvud muutuvad suuremate vabadusastmete arvude korral suhteliselt vähe, on tabelist nö "ridu vahelt ära jäetud". Puuduvate vabadusastmete arvude asemel võib alati kasutada lähimat väiksemat vabadusastmete arvu, näiteks 22 asemele võtta 20.

Väärtused, mis on antud tabeli teises ja kolmandas reas, on nn α-täiendkvantiilid, st niisugused arvu q väärtused, mille korral kehtib võrdus

$$P(X>q)=\alpha, \qquad (17)$$

st, et arvust q suuremaid väärtusi omandab juhuslik suurus X väga harva, üksnes tõenäosusega α. Siin on α olulisuse nivoo, seega üsna pisike arv. Tabelis 37 on olulisuse nivooks vastavalt 5% või 1%.

Tabeli neljandas ja viiendas veerus on arv q' defineeritud pisut teisiti, nimelt võrdusega

$$P(|X|>q')=\alpha, \qquad (18)$$

seega kui arv, millest absoluutväärtuse poolest suuremaid väärtusi omandab juhuslik suurus küllalt harva, nimelt tõenäosusega α. Kui juhuslik suurus X on sümmeetriline, siis on valemiga (18) määratud arv q' sama, mis oleks valemiga (17) määratud q siis, kui tema valemis võtta α asemele $\alpha/2$.

Valemiga (17) määratud arvu q kasutatakse ühepoolsete, valemiga (18) määratud arvu q' aga kahepoolsete $(1-\alpha)$-usalduspiiride arvutamiseks. Rakenduslikes käsiraamatutes nimetatakse neid arve mõnikord ka lihtsalt tabeliväärtusteks.

3.3.5. Studenti *t*-jaotuse kasutamine keskväärtuse usalduspiiride arvutamisel

Kui valimi maht pole väga suur (rusikareeglina öeldes: on alla saja vaatluse), siis tuleb normaaljaotuse keskväärtuse usalduspiiride arvutamisel kasutada *t*-jaotust. Tundmatu keskväärtuse μ jaoks $(1-\alpha)$-usaldusvahemikku kirjeldavad usalduspiirid $\underline{\mu}$ ja $\overline{\mu}$ arvutatakse alljärgnevalt:

$$\underline{\mu} = \bar{x} - q'(\alpha, n-1)\frac{s}{\sqrt{n}}; \quad \overline{\mu} = \bar{x} + q'(\alpha, n-1)\frac{s}{\sqrt{n}}, \qquad (19)$$

kus $q'(\alpha,n-1)$ tähistab *t*-jaotuse tabeliväärtust kahepoolsete $(1-\alpha)$-usalduspiiride arvutamiseks (veerud 4 ja 5 tabelis 37) vabadusastmete arvu $n-1$ korral. Keskväärtuse ja standardhälbe hinnangud on $\bar{x}$ ja s.

Näide 45

Vaatleme valimit mahuga 20 vaatlust ülalkirjeldatud andmestikust (Eesti valdade elanike arv). Olgu selle valimi põhjal keskväärtuse ja standardhälbe hinnangud vastavalt 2544,7 ja 1289. Olgu ülesandeks leida üldkogumi keskväärtusele 99%-lised usalduspiirid.

- *Et pole eraldi räägitud ühepoolsetest usalduspiiridest, moodustame kahepoolsed usalduspiirid.*
- *Leiame t-jaotuse tabelist sobiva veeru, see on parempoolseim. Puudub sobiv rida juhuks n – 1 = 19, arvutame selle lineaarse interpolatsiooni teel: q = 0,5 (2,88 + 2,85) = 2,865.*
- *Leiame m = $s/\sqrt{20}$ = 1289/4,472 = 288,23.*
- *Arvutame usalduspiirid: 1718,92 ja 3370,48.*

Näeme, et meie teadaolev üldkeskmine paikneb ootuspäraselt arvutatud usalduspiirkonnas.

Et tsentraalse piirteoreemi kohaselt ühesuguse jaotusega sõltumatute liidetavate normeeritud summa läheneb normaaljaotusega juhuslikule suurusele, siis on ülalkirjeldatud keskväärtuse usalduspiiride leidmise eeskiri lähendina rakendatav ka sellisele juhule, kus üldkogumi jaotus erineb normaaljaotusest. Loomulikult sõltub selle lähendi headus sellest, kui tugevasti lähtejaotus erineb normaaljaotusest ja kui suur on valimi maht.

3.4. Statistiliste hüpoteeside kontrollimine

3.4.1. Statistilise hüpoteesi mõiste

Üks kõige olulisemaid matemaatilise statistika ülesandeid on kontrollida mitmesuguseid hüpoteese. Inimesed võivad tavakogemuse tasemel teha mitmesuguseid tähelepanekuid, kuid tihti on raske otsustada, kas niisuguste tähelepanekute korral on tegemist juhuslike nähtuste või süstemaatiliste tendentsidega. Toome mõningaid näiteid.

- Talved on muutunud lumevaesemaks ja soojemaks.
- Pojad on keskmiselt pikemad kui isad.
- Inimesed sõidavad praegu vähem bussidega kui mõni aasta tagasi.

Arusaadavalt saaks kõikse uuringu tulemusena öelda, kas nimetatud väited on õiged, ning kui on, siis näidata ka täpne muudatuse suurus. Käesolevas peatükis me aga põhimõtteliselt ei tegele kõikse uuringuga, vaid eeldame, et meie käsutuses on üksnes valimi andmed. Et valim on alati juhuslik, tuleb meil lahendada alljärgnev probleem:

Kas meie poolt valimi andmete baasil tehtud tähelepanek kehtib ka üldkogumis? Kas me saame kinnitada, et tegemist on üldkehtiva seaduspärasusega?

Et kasutada matemaatilise statistika aparatuuri püstitatud ülesande lahendamiseks, tuleb toimida alljärgnevalt:

- sõnastada tõestamist vajav sisuline hüpotees;
- sõnastada sellele hüpoteesile vastav statistiline hüpotees ja selle vastandhüpotees;
- kontrollida püstitatud statistilise hüpoteesi eeldusi ja leida selle kontrollimiseks sobiv metoodika (test);
- kontrollida hüpoteesi, kasutades selleks valimi andmeid;
- vastavalt statistilise hüpoteesi kontrollimise tulemusele teha sisuline järeldus.

Märgime, et statistiliste hüpoteeside kontrollimise juures on peaaegu alati tarvis teha üldkogumi jaotuse kohta mõningaid eeldusi. Sageli eeldatakse, et üldkogumi jaotus on normaaljaotusega. Seda eeldust kasutame meiegi käesolevas paragrahvis. Õnneks kehtivad saadavad tulemused ka siis, kui üldkogumi jaotus normaaljaotusest mõnevõrra erineb.

3.4.2. Statistiliste hüpoteeside liigid

Statistilisi hüpoteese sõnastatakse *tavaliselt* kas

- jaotuste või
- jaotusparameetrite kohta.

Statistilised hüpoteesid sõnastatakse enamasti alternatiivsete hüpoteeside paarina H_0 ja H_1, kusjuures nendest hüpoteesidest saab tõene olla üks ja ainult üks.

- Hüpoteesi H_0 nimetatakse *nullhüpoteesiks*. Tavaliselt väljendab null-hüpotees ebahuvitavat, väheinformatiivset juhtu ("nii nagu alati on arvatud"). *Nullhüpoteesi ei ole võimalik tõestada.* Selle vastuvõtmine võib tähendada uuringu jätkumist.
- Hüpoteesi H_1 nimetatakse *sisukaks hüpoteesiks.* Enamasti on uurija sooviks tõestada sisukas hüpotees.

Vaatleme esialgu hüpoteese parameetri (tunnuse arvkarakteristiku) kohta. Kõik need hüpoteesid väidavad midagi parameetri *tegeliku väärtuse* ν kohta, mida me paraku ei tea. Kõige lihtsam hüpotees (nn *lihthüpotees*) parameetri ν kohta on väide, et see parameeter võrdub mingi antud konstandiga ν_0. Selle antud konstandi väärtuseks on praktilistes ülesanetes sageli null. Üks väga sageli esinev statistiliste hüpoteeside paar ongi alljärgnev:

- H_0: $\nu = \nu_0$, H_1: $\nu \neq \nu_0$.

Teine võimalus on väita, et parameeter pole mingist teadaolevast arvust väiksem. Seda väidet esitab ühepoolsete hüpoteeside paar:

- H_0: $\nu \geq \nu_0$, H_1: $\nu < \nu_0$.

Kolmas võimalus on väita, et parameeter on väiksem või niisamasuur kui mingi teadaolev arv. Ka seda väidet esitab ühepoolsete hüpoteeside paar:

- H_0: $\nu \leq \nu_0$, H_1: $\nu > \nu_0$.

Siintoodud hüpoteeside paarid on küll laialt levinud, kuid moodustavad siiski vaid väikese osa kõikvõimalikest statistiliste hüpoteeside paaridest.

Hüpoteesid jaotuse kohta sõnastatakse tavaliselt alljärgnevalt.

- H_0: juhuslik suurus X on mingi teadaoleva (teoreetilise) jaotusega P_0.
- H_1: juhusliku suuruse X jaotus erineb jaotusest P_0.

Loomulikult mõeldakse siin juhusliku suuruse X *teoreetilist ehk tegelikku jaotust*, st jaotust üldkogumis.

Teine sageli kontrollitav hüpoteeside paar on selline:

- H_0: juhuslikud suurused X ja Y on sama jaotusega;
- H_1: juhuslike suuruste X ja Y jaotused on erinevad.

3.4.3. Vead hüpoteeside kontrollimisel

Et statistiliste hüpoteeside kontrollimisel tehakse valimi põhjal järeldusi üldkogumi kohta, tuleb paratamatult arvestada võimalusega, et otsustamisel tekib viga ja tehakse vale järeldus. Statistiliste hüpoteeside kontrollimise teooria põhiliseks iseärasuseks on, et *otsustamise juures reguleeritakse vigade esinemise tõenäosusi.* Et hüpoteesid H_0 ja H_1 ei ole samaväärsed, pole seda ka vead. Kokkuvõttes on otsustamise käigus võimalikud järgmised tulemused.

Tabel 12

Tegelikkus\ Otsustus	Võetakse vastu H_0	Võetakse vastu H_1
Kehtib H_0	õige	1. liiki viga
Kehtib H_1	2. liiki viga	õige

Statistilised hüpoteesid sõnastatakse reeglina nii, et eriti ebasoovitav on esimest liiki viga (loetakse tõestatuks sisukas hüpotees, kuigi tegelikult on õige nullhüpotees). Seda arvestades konstrueeritakse otsustuse vastuvõtmise eeskiri (nn *kriteerium*) nii, et esimest liiki vea tõenäosus oleks küllalt väike.

Tähtis on siinjuures see, et otsustuse tegijal on enesel võimalik valida, kui suure tõenäosusega peab ta võimalikuks lubada esimest liiki veal esineda.

- Esimest liiki vea tegemise suurimat lubatavat tõenäosust nimetatakse *olulisuse nivooks* ja tema tavaliseks tähiseks on α.
- Olulisuse nivoo on alati väike positiivne arv, tema väärtuseks valitakse tavaliselt 0,05, 0,01 või 0,001.
- Mida väiksem on olulisuse nivoo, seda *rangem* on kriteerium, ning võib öelda, et seda *raskem* on sisukat hüpoteesi tõestada.

Viimane väide tähendab seda, et võib esineda olukord, kus sisuka hüpoteesi saab küll tõestada siis, kui olulisuse nivoo väärtuseks on 0,05, kuid seda ei õnnestu teha olulisuse nivoo väärtusel 0,01.

Üldiselt toimub esimest liiki vea tõenäosuse vähendamine teatavas mõttes teist liiki vea tõenäosuse arvel. Teist liiki vea tõenäosus võib põhimõtteliselt ulatuda kuni väärtuseni $1 - \alpha$. Enamasti püütakse hüpoteeside kontrollimise kriteeriumid konstrueerida nii, et esimest liiki vea tõenäosus piiratakse olulisuse nivooga ja lisaks sellele minimeeritakse teist liiki vea tõenäosus. Selliseid tingimusi rahuldavat kriteeriumi nimetatakse *võimsaimaks.*

3.4.4. Hüpoteeside kontrollimine normaaljaotuse keskväärtuse kohta

Üks kõige sagedamini praktikas esinevaid ülesandeid on normaaljaotusega juhusliku suuruse keskväärtuse võrdlemine antud konstandiga. Paneme tähele, et seda ülesannet lahendades me ei tõesta, et keskväärtus võrduks antud konstandiga, vaid vastupidi – üritame tõestada, et keskväärtus ei võrdu antud konstandiga. See, missuguse konstandiga tundmatut keskväärtust võrreldakse, sõltub ülesande sisulisest püstitusest. Niisiis on üks kõige lihtsamaid hüpoteeside paare normaaljaotuse keskväärtuse kohta alljärgnev:

$$H_0: \mu = \mu_0,$$
$$H_1: \mu \neq \mu_0.$$

Siin μ tähistab normaaljaotusega juhusliku suuruse keskväärtust üldkogumis ja μ_0 on antud konstant. Sageli on ülesanne sõnastatud nii, et μ_0 väärtus on null.

Näide 46

Mõõtku X ühe majandusnäitaja muutust esimesest teise kvartalini Eesti valdades. Hüpoteesiks, mida soovitakse tõestada, on see, et kogu riigis majanduskasv erineb nullist. Olulisuse nivooks on 0,05. Valimisse on võetud andmed 15 vallast, n=15, ning need moodustavad alljärgneva variatsioonrea:

-256, -130, -55, 12, 15, 29, 201, 278, 347, 377, 451, 589, 621, 699, 850.

Seda hüpoteesipaari on võimalik tõestada järgmiselt.

- *Leiame majandusmuutuste variatsioonrea põhjal valimkeskmise* $\bar{x} = 268{,}5$ *ja valimi standardhälbe 330,83.*
- *Arvutame valemi (19) põhjal majandusmuutuse keskväärtuse jaoks 0,95-usaldusvahemiku (85,3; 451,7), kasutades usaldusnivood* $1-\alpha$*, kus* α *on antud olulisuse nivoo.*
- *Kontrollime, kas punkt* μ_0 *sisaldub usalduspiirkonnas.*
- *Kui jah, siis loeme nullhüpoteesi vastuvõetuks, ent mitte tõestatuks. Vajaduse korral jätkatakse uurimist.*
- *Kui ei, siis loeme sisuka hüpoteesi tõestatuks. Uurimine on lõpetatud, eesmärk saavutatud.*

Käesoleva näite puhul me saime sisuka hüpoteesi tõestada, andmed kinnitavad nullist erineva majanduskasvu olemasolu.

Toodud näite põhjal veendusime, et statistilise hüpoteesi kontrollimiseks keskväärtuse kohta saab kasutada keskväärtuse usalduspiire. Tegelikult ei ole aga usalduspiiride väljaarvutamine üldse vajalik, näites 46 esitatud arutluse saab esitada ka alljärgneva üldise skeemi järgi.

- Arvutame andmestiku põhjal valimi keskmise $\bar{x}$ ja standardhälbe s.
- Kasutades leitud suurusi ja teadaolevat valimi mahtu n leiame t-statistiku väärtuse

$$t = \frac{|\bar{x} - \mu_0|}{s}\sqrt{n}. \qquad (20)$$

- Võrdleme t-statistiku väärtust t-jaotuse tabelis antud kriitilise väärtusega, kasutades
  - vabadusastmete arvu $n - 1$,
  - seda tabeli osa, mis käib kahepoolse hüpoteesi kohta,
  - kokku lepitud olulisuse nivoole α vastavat veergu.
- Kui kehtib võrratus $t > t(\alpha,n-1)$, siis loetakse sisukas hüpotees tõestatuks, vastasel korral võetakse vastu nullhüpotees, järeldades seega, et olemasoleva materjali korral pole võimalik sisukat hüpoteesi tõestada.

Seda tulemust on ka üsna lihtne tõlgendada: üldkogumi keskväärtus ei ühti antud arvuga μ_0 siis, kui valimkeskmine erineb arvust μ_0 palju. Kui aga erinevus on väike, on täiesti võimalik, et tegelik keskväärtus võrdub arvuga μ_0.

Näide 47

Vaatame eelmises näites käsitletud valdade majandusnäitajate andmestikku. Oletame, et mainekas majandusteadlane väitis Eesti maa-alade kolme kuu majanduskasvu olevat 200 ühikut uuritava majandusnäitaja järgi. Soovides kontrollida tema väite paikapidavust, püstitame hüpoteeside paari H_0: Eesti valdades kasvas vaadeldud majandusnäitaja kolme kuu jooksul 200 ühiku võrra ($\mu = 200$); H_1: $\mu \neq 200$. t-statistiku väärtuseks saame:

$$t = \frac{|268{,}5 - 200|}{330{,}38 \cdot 3{,}87} = 0{,}80$$

Kuna vastav kriitiline t-väärtus t(0,05; 14)=2,14 on suurem kui leitud t-statistik, siis ei saa me nullhüpoteesi ümber lükata – st antud valimi põhjal pole meil võimalik tõestada, et majandusnäitaja muutus Eesti valdades erines 200-ühikulisest kasvust. Sama järeldust kinnitab ka eelmine näide – majandusnäitaja kasv 200 ühiku võrra jääb 0,95-usaldusvahemikku.

3.4.5. Ühepoolsete hüpoteeside kontrollimine normaaljaotuse keskväärtuse kohta

Alati pole sisuline ülesanne niisugune, et sellele vastaks kahepoolne hüpotees, millega sisuliselt tõestatakse, et keskväärtus ei võrdu teatava etteantud arvuga. Arvatavasti on hoopis sagedamini tarvis kontrollida ühepoolsete hüpoteeside paari, näiteks

$$H_1: \mu > \mu_0,$$
$$H_0: \mu \le \mu_0.$$

Niisuguse hüpoteesipaari kontrollimiseks saame kasutada *t*-testi üsna sarnaselt eelmises punktis kirjeldatud juhule, kusjuures erinevusi on vaid kaks: *t*-statistiku defineerimisel ei kasutata absoluutväärtuse märke ja tabeliväärtus valitakse sellest tabeli osast, mis vastab ühepoolsele hüpoteesile. Seega saame järgmise eeskirja.

- Arvutame andmestiku põhjal valimi keskmise $\bar{x}$ ja standardhälbe *s*.
- Kasutades leitud suurusi ja teadaolevat valimi mahtu *n*, leiame *t-statistiku*

$$t = \frac{\bar{x} - \mu_0}{s}\sqrt{n}. \qquad (21)$$

- Võrdleme *t*-statistiku väärtust *t*-jaotuse tabelis antud kriitilise väärtusega, kasutades
  - vabadusastmete arvu $n - 1$,
  - seda tabeli osa, mis käib ühepoolse hüpoteesi kohta,
  - kokku lepitud olulisuse nivoole α vastavat veergu.

Kui kehtib võrratus $t > t(\alpha, n - 1)$, siis loetakse sisukas hüpotees tõestatuks, vastasel korral võetakse vastu nullhüpotees, järeldades seega, et olemasoleva materjali korral pole võimalik sisukat hüpoteesi tõestada.

Ühepoolse hüpoteesi tõestamise korral tuleb hoolikalt jälgida, et *t*-statistiku lugejas olev avaldis kirjutataks õigesti. Kui me tahame tõestada, et *keskväärtus on arvust μ_0 suurem*, siis tuleb kirjutada avaldisse *valimi keskmise ja μ_0 vahe*, ja see vahe peab kindlasti olema *positiivne*. Kui arvutamise tulemusena saadakse valimi keskväärtus $\bar{x}$, *mis on antud arvust μ_0 väiksem*, siis võib *ilma edasiste arvutusteta võtta vastu nullhüpoteesi*: keskväärtus ei saa olla arvust μ_0 suurem.

Täpselt samuti võib esineda ülesandeid, mille korral pakub huvi tõestada ühepoolne hüpotees:

$$H_1: \mu < \mu_0,$$

millele vastab nullhüpotees

$$H_0: \mu \ge \mu_0.$$

Selle ülesande lahendamiseks tuleb t-statistik määratleda alljärgnevalt:

$$t = \frac{\mu_0 - \bar{x}_0}{s}\sqrt{n}. \qquad (22)$$

Muu osa ülesande lahenduskäigust langeb täpselt kokku vastupidise ühepoolsete hüpoteeside paari kontrollimise ülesandega, millega tutvusime käesoleva punkti esimeses pooles.

Näide 48

Tegelikult pakub eelmises näites sisulist huvi tõestada, et majanduskasv on positiivne. Seega oleks meil hoopiski arukam sõnastada ühepoolse hüpoteesi tõestamise ülesanne:

$$H_0: \mu_x \le 0, \qquad H_1: \mu_x > 0.$$

Arvutame t-statistiku: $t = \frac{\bar{x} - \mu_0}{s}\sqrt{n} = \frac{268,5}{330,83}\sqrt{15} = 3,14.$

Näeme, et arvutatud t-statistiku väärtus on suurem kui kriitiline väärtus t(0,05;14)=1,76 (vt tabel 37 Lisas, ühepoolsete hüpoteeside veerg). Seega saame tõestada alternatiivse hüpoteesi – majanduskasv on olnud positiivne.

Märkame, et ühepoolse hüpoteesi puhul on t-statistiku kriitiline väärtus märksa väiksem kui kahepoolse hüpoteesi korral. Mõlemal juhul sarnaselt arvutatav t-statistik ületab seega ühepoolse hüpoteesi testimisel kriitilise väärtuse märksa "kergemini". Seetõttu saab ühepoolset hüpoteesi tõestada keskmiselt väiksema valimi põhjal, st odavamalt. Järelikult, kui uurijat tegelikult huvitab ühepoolse hüpoteesi tõestamine, siis on otstarbekam tõestamiseks validagi ühepoolsete hüpoteeside paar.

Mõnikord peetakse ühepoolsete hüpoteeside kontrollimist keerukamaks. Tegelikult pole see nii, kui vaid järgida allpool esitatud näpunäiteid.

- Alati tuleb hüpoteesid sõnastada nii, et sisukas hüpotees ühtib sellega, mida sisuliselt soovitakse tõestada.
- t-statistik määratakse nii, et kui panna selle avaldisse $\bar{x}$ asemele tõeline keskmine μ, siis on murru lugejas olev vahe positiivne sisuka hüpoteesi kehtimise korral.

Kui soovitakse tõestada ühepoolset hüpoteesi, *tuleb see sõnastada enne andmetöötluse algust.* Kahepoolse hüpoteesi asendamine ühepoolsega töötlustulemuste põhjal ei ole uurijaeetika seisukohalt korrektne.

3.5. *Kahe üldkogumi keskmiste võrdlemine*

3.5.1. Kahe üldkogumi keskväärtuste võrdlemine (nn sõltumatute vaatluste ja ühise dispersiooni juhtum)

Üks tihti püstitatavaid statistikaülesandeid on kahe üldkogumi keskväärtuste võrdlemine. Sellised ülesanded on näiteks: kas Eesti elanike keskmine vanus erineb Läti elanike keskmisest vanusest? kas Põhja-Eesti ja Lõuna-Eesti vallad on keskmiselt sama elanike arvukusega? jne. Standardseks vahendiks sellise ülesande lahendamisel on t-jaotusel baseeruv t-test.

Eeldame, et meil on kaks üldkogumit ja mõlemas on meid huvitav tunnus X normaaljaotusega. Tähistame selle tunnuse keskväärtuse esimeses üldkogumis tähega μ_1 ja teises üldkogumis tähega μ_2. Eeldame veel lihtsuse mõttes, et tunnuse dispersioon on mõlemas üldkogumis sama, olgu see σ^2.

Sõnastame tõestamist vajava hüpoteesi (tunnusel on erinevates üldkogumites erinev keskväärtus) ja seda välistava nullhüpoteesi ning valime välja meie ülesande jaoks sobiva olulisuse nivoo α:

$$H_0: \mu_1 = \mu_2,$$
$$H_1: \mu_1 \neq \mu_2.$$

Kui nullhüpotees oleks õige, siis oleks mõlemas üldkogumis sama jaotus. Oletame, et meil on olemas järgmised andmed: valim mahuga n_1 esimesest üldkogumist, $x_1, \ldots, x_{n_1}$, ja valim mahuga n_2 teisest üldkogumist, $y_1, \ldots, y_{n_2}$.

Kuivõrd tegemist on erinevate üldkogumitega, siis on ka kõik mõõdetavad objektid erinevad, seda tähendabki *sõltumatute vaatluste nõue*. Oletame, et kummagi valimi põhjal on arvutatud keskväärtuse ja standardhälbe hinnangud $\bar{x}$ ja $\bar{y}$ ning s_1^2 ja s_2^2. Et eelduse kohaselt on mõlemas üldkogumis on ühine dispersioon, siis me võime dispersioonihinnangud ühendada. Tulemuseks saame nn *ühise dispersioonihinnangu* s^2:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}. \qquad (23)$$

Kui nullhüpotees on õige, siis on avaldis

$$t = \frac{\bar{x} - \bar{y}}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \qquad (24)$$

t-jaotusega vabadusastmete arvuga $n_1 + n_2 - 2$.

Püstitatud hüpoteesi kontrollimine on t-tabelite abil lihtne.

- Kui arvutatud *t-statistiku* absoluutväärtus $|t|$ on väiksem kui tabelist leitud kriitiline väärtus valitud olulisuse nivoo ja vabadusastmete arvu $n_1 + n_2 - 2$ korral, siis meil ei õnnestu sisukat hüpoteesi tõestada ja me jääme nullhüpoteesi juurde, mida aga ei saa lugeda tõestatuks.
- Kui $|t|$ on vastavast kriitilisest väärtusest suurem, siis me loeme, et sisukas hüpotees on antud olulisuse nivoo korral tõestatud.

Ka kahe üldkogumi võrdlemisel on võimalik, et uurijal on kasutada teatav eelinformatsioon, mille põhjal ta soovib tõestada mitte kahepoolset väidet kahe üldkogumi keskväärtuste erinevuse kohta, vaid näidata, et ühes konkreetses üldkogumis on keskväärtus suurem. Kasutades sama tähistust mis varemgi, saaksime siis näiteks alljärgneva statistiliste hüpoteeside paari:

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

Et tõestada ühepoolset hüpoteesi, tuleb vastavalt määrata ka t-statistik. Selle lugejas on niisugune vahe, mis on siis positiivne, kui sisukas hüpotees on õige ja valimkeskmiste asemel üldkogumi keskmised μ_1 ja μ_2. Käesoleva hüpoteesipaari kontrollimiseks sobib valemiga (24) avaldatud t-statistik.

Sisuka hüpoteesi kontrollimiseks tuleb selle statistiku väärtust võrrelda ühepoolse hüpoteesi kontrollimiseks ette nähtud t-jaotuse kriitiliste väärtustega.

Näide 49

Olgu meil eesmärgiks võrrelda tööealiste inimeste (16 – 60-aastaste meeste ja 16 – 55-aastaste naiste) osatähtsust Põhja- ja Lääne-Eesti valdades ning Lõuna- ja Kagu-Eesti valdades. Võtame aluseks hüpoteesi, et Põhja- ja Loode-Eestis on tööealiste inimeste osatähtsus keskmiselt suurem kui Lõuna- ja Kagu-Eestis.

Tabel 13

Põhja- ja Loode-Eesti		Lõuna- ja Kagu-Eesti	
Vald	Tööealiste protsent	Vald	Tööealiste protsent
Aseri	56,7	Ahja	52,1
Haljala	57,8	Antsla	54,5
Harku	58,9	Haaslava	51,5
Kernu	54,1	Hummuli	52,8
Kohtla	57,7	Nõo	53,7
Kullamaa	55,5	Olustvere	51,1
Nissi	54,4	Pajusi	52,0
Saku	57,3	Puhja	54,0
Sonda	55,1	Põltsamaa	53,7
Sõmeru	57,5	Pühajärve	53,6
		Urvaste	50,1
		Valgjärve	51,8

Tähistame tööealiste osatähtsuse tähega X ja võtame esimeseks üldkogumiks Põhja- ja Loode-Eesti ning teiseks Lõuna- ja Kagu-Eesti. Tähistame tunnuse X keskväärtused nendes üldkogumites vastavalt μ_1 ja μ_2. Nüüd saame oma tõestamist vajava sisuka hüpoteesi sõnastada alljärgnevalt:

$$H_1 : \mu_1 > \mu_2 .$$

Nullhüpotees on sel juhul

$$H_0 : \mu_1 \leq \mu_2 .$$

Olulisuse nivooks valime 0,01.

Soovitava hüpoteesi tõestamiseks teeme järgmised sammud:

- *Arvutame mõlema valimi jaoks keskmised; saame $\bar{x}_1 = 56,5$ ja $\bar{x}_2 = 52,575$.*
- *Arvutame mõlema valimi jaoks dispersiooni hinnangud; saame $s_1^2 = 2,633$ ja $s_2^2 = 1,817$.*
- *Näeme, et dispersiooni hinnangud on küllaltki lähedased (seda eeldust võib kasutada, kui suurem dispersiooni hinnang ületab väiksemat vähem kui kaks korda). Seejärel arvutame valemi (23) abil ühise dispersioonihinnangu $s^2 = 2,184$.*
- *Leiame ühise standardhälbe hinnangu $s = 1,478$.*
- *Arvutame nüüd t-statistiku, mis vastab meie poolt sõnastatud nullhüpoteesile ja valimi mahtudele $n_1 = 10$ ning $n_2 = 12$:*
  $t = (3,975 \times 2,335)/1,478 = 6,28$.
- *Saadud t-statistiku väärtus ületab suurelt t-jaotuse 0,01-täiendkvantiili väärtuse vabadusastmete arvu 20 korral, seega võime oma hüpoteesi lugeda väga veenvalt tõestatuks: tööealiste elanike suhtarv on Põhja- ja Loode-Eestis kõrgem kui Lõuna- ja Kagu-Eestis.*

Näide 50

Püstitame uue hüpoteesi (H_1): Põhja- ja Lõuna-Eestis on vallad keskmise elanike arvu poolest erineva suurusega. Püüame seda hüpoteesi tõestada olulisuse nivool α=0,05. Nullhüpoteesiks (H_0) on siis väide, et nii Põhja- kui Lõuna-Eestis on elanike arv vallas keskmiselt sama. Loeme Põhja-Eestiks Harjumaa, Ida- ja Lääne-Virumaa ning Raplamaa valdu, Lõuna-Eestiks aga Valga-, Võru-, Põlva-, Viljandi- ja Tartumaa valdu. Moodustame kummastki üldkogumist valimi: Põhja-Eesti valimisse kuuluvad Harku, Kose, Raasiku, Saku, Alajõe, Iisaku, Lüganuse, Haljala, Rakvere, Tamsalu ja Rapla vald, seega kokku on neid 11. Elanike arvu keskväärtuse hinnang on 3139,2 ja dispersiooni hinnang 3 580 407. Lõuna-Eesti valdade valimisse sattusid Mikitamäe, Veriora, Haaslava, Nõo, Halliste, Karula, Kolga-Jaani, Lasva ja Rõuge vald, kokku 9. Keskmine elanike arv neis valdades on 2061,9 ja valimdispersioon 472819. Näeme, et käesolevas näites on aga valimite dispersiooni hinnangud väga erinevad (nende suhe on 7,57). Seega ei saa me kasutada eeldust, et mõlemas üldkogumis on sama dispersioon, ning me ei saa ülesannet ülalesitatud meetodi abil lahendada.

3.5.2. Kahe üldkogumi keskväärtuste võrdlemine (nn sõltumatute vaatluste ja erinevate dispersioonide juhtum)

Näites 50 esitatud ülesande lahendamisel tekib kahtlus, kas on õige kasutada eeldust, et mõlemas üldkogumis on dispersioonid võrdsed, kui ühe valimi põhjal saadud dispersiooni hinnang on üle seitsme korra suurem kui teises valimis. Järelikult pole õige arvutada ka ühist dispersioonihinnangut s^2. Peatumata siinkohal dispersioonide võrdlemise metoodikal, toome alternatiivse t-statistiku valemi, mida võib kasutada siis, kui üldkogumite dispersioonid on erinevad.

- Arvutame kummagi üldkogumi jaoks dispersiooni s_i^2 ja keskväärtuse hinnangu dispersiooni s_i^2/n_i.
- Leiame keskväärtuse hinnangute vahe $\bar{x} - \bar{y}$ dispersiooni hinnangu $s_v^2 = s_1^2/n_1 + s_2^2/n_2$.
- Arvutame t-statistiku absoluutväärtuse $| \bar{x} - \bar{y} | / s_v$.
- Võrdleme saadud statistikut t-jaotuse kriitilise väärtusega olulisuse nivoo α ja vabadusastmete arvu $\min(n_1, n_2)$ korral.

Olgu märgitud, et see test on ligikaudne. Selle jaoks sobivaima vabadusastmete arvu määramiseks on olemas täpsem, kuid keerukam valem. Et aga t-jaotuse täiendkvantiilid suuremate vabadusastmete arvude korral muutuvad üsna vähe, siis võimaldab ka vaadeldav ligikaudne test väga suure tõenäosusega teha õigeid järeldusi üldkogumi kohta.

Näide 51

Lahendame eelmises näites sõnastatud ülesande, arvestades dispersioonide erinevust üldkogumites.

- *Arvutame valimikeskmiste dispersioonihinnangud 3 580 407/11 ja 472819/9.*
- *Leiame vahe standardhälbe 614,8.*
- *Arvutame t-statistiku 1077,3/614,8=1,75.*

Kuna t-statistiku arvutatud väärtus on väiksem t-jaotuse kriitilisest väärtusest 2,23, mis vastab kahepoolsele hüpoteesile olulisuse nivoo 0,05 ja vabadusastmete arvu 10 puhul, siis tuleb meil vastu võtta nullhüpotees: me ei saa oma andmete põhjal väita, et Lõuna-Eesti ja Põhja-Eesti valdade vahel oleks süstemaatiline erinevus keskmises elanike arvus.

3.5.3. Keskväärtuste võrdlemine sõltuvate vaatluste korral

Sageli esineb ülesandeid, mille puhul võrreldavad valimid *ei ole sõltumatud*. Niisugune on olukord näiteks siis, kui soovitakse mõõta *muutusi*, mis on toimunud uuritavas üldkogumis. Sellisel juhul mõõdetakse samu objekte kahel korral. Mõlemal korral on valimi maht sama (olgu see n), ning vaatlused on sõltuvad. Tähistame esimese mõõtmise tulemused vastavalt sümbolitega $x_1, \ldots, x_n$ ja teise mõõtmise tulemused vastavalt $y_1, \ldots, y_n$. Olgu nende mõõtmiste keskväärtused vastavalt μ_1 ja μ_2. Kontrollimist vajab sel juhul sama hüpoteesipaar, mis punkti 3.5.1 alguses:

$$H_0: \mu_1 = \mu_2,$$
$$H_1: \mu_1 \neq \mu_2.$$

Samuti on võimalik sõnastada ühepoolsed hüpoteesid, mis määratlevad juba enne andmetega tutvumist võimalike muutuste suuna.

Ekslik oleks sõltuvate valimite korral kasutada punktis 3.5.1 tutvustatud metoodikat, kus vaatlused olid eeldatavalt sõltumatud. Selle asemel on otstarbekas vaadelda üksikobjektide muutusi ning teha selgeks, *kas keskmine muutus erineb nullist*. Seega on lihtsaim võimalus käesoleva ülesande lahendamiseks alljärgnev.

- Arvutame iga vaatlustepaari jaoks muutuse $d_i = y_i - x_i$.
- Leiame muutuste keskväärtuse $\bar{d}$ ja standardhälbe s_d.
- Kontrollime hüpoteesi: keskmine muutus erineb nullist, kasutades selleks punktides 3.4.4 ja 3.4.5 esitatud metoodikat.

Näide 52

Olgu meil valim Eesti linnalistest asulatest, kusjuures on kindlaks tehtud nende elanike arvud 1. jaanuaril 1994 ning 1. jaanuaril 1996. Ülesandeks on tõestada, et kahe aasta jooksul on keskmine asulate elanike arv muutunud (olulisuse nivoo on 0,05).

Valimi andmed on esitatud tabelis 14, kusjuures tabeli viimastes ridades on esitatud valimkeskmised ja -standardhälbed, samuti ka usalduspiiride arvutamiseks vajalik konstant $t \cdot m$, kus t on t-jaotuse kriitiline väärtus vastavalt vabadusastmete arvule ja olulisuse nivoole ning m on nn standardviga, $m = s/\sqrt{n}$.

Näeme, et muutuse 95%- usalduspiirid on –143,3 ja –22,5, seega nullpunkti usalduspiirkonnas ei sisaldu ja me võime kinnitada, et asulate keskmine elanikkond on möödunud kahe aasta jooksul vähenenud. Keskmise vähenemise hinnang on 82,9 ja vähenemise 95%-usalduspiirkond 82,9 ±60,4. Meie poolt soovitud hüpotees on sellega tõestatud.

Tabel 14

Asula	1.01.94	1.01.96	muutus
Aegviidu	1054	1094	40
Kehra	3787	3735	-52
Saue	4581	4725	144
Kärdla	4405	4341	-64
Kiviõli	9392	9017	-375
Püssi	2406	2277	-126
Jõgeva	7053	6669	-384
Paide	10543	10314	-229
Lihula	1810	1775	-35
Rakvere	18769	18543	-226
Võsu	720	755	35
Räpina	3425	3299	-126
Kilingi-Nõmme	2585	2661	76
Vändra	3005	3052	47
Järvakandi	1930	1940	10
Rapla	6483	6235	-248
Elva	6389	6357	-32
Otepää	2491	2500	9
Abja-Paluoja	1724	1693	-31
Suure-Jaani	1493	1433	-60
Antsla	1608	1494	-114
Keskmine	4554,905	4471,857	-82,9048
Sthälve	4237,5	4153,4	141,2179
t×m			60,39874

Kui me aga "unustaksime ära" vaatluste sõltuvuse ja kasutaksime valemeid (23) ja (24), saaksime hoopis erineva tulemuse. Sel juhul oleks ühine standardhälbe hinnang s=4195,6 ja t-statistiku väärtus kõigest 0,064, mis on palju väiksem kriitilisest väärtusest. Sellisel viisil poleks sisukat hüpoteesi võimalik tõestada.

Vaatluste sõltuvus ei tarvitse alati tuleneda sama tunnuse korduvast mõõtmisest. Selle põhjuseks võib olla ka näiteks see, et mõõtmisi toimetati samas piirkonnas, kuigi mõõdetavad objektid ise on erinevad.

Näide 53

Vaatleme näites 49 tutvustatud Põhja-Eesti valdu, ning uurime mees- ja naissoost töötealiste elanike arvukuse vahekorda nendes. Arvestades naiste suuremat osatähtsust elanikkonnas üldse, püstitame kaks tööhüpoteesi. Need on:

- *Põhja-Eesti töötealise maaelanikkonna hulgas on naiste ülekaal.*
- *Põhja-Eesti valdades erineb töötealiste meeste keskmine arvukus töötealiste naiste keskmisest arvukusest.*

Tähistame meeste arvu tähega X, naiste arvu tähega Y ja meeste ning naiste arvu vahe tähega D. Töötealiste maameeste arvu keskväärtust Põhja-Eesti valdades tähistagu μ_1 ja töötealiste maanaiste arvu keskväärtust Põhja-Eesti valdades μ_2. Esimene sisukas hüpotees, mida me tahaksime tõestada, on koos vastava nullhüpoteesiga alljärgnev:

$$H_1 : \mu_1 < \mu_2,$$
$$H_0 : \mu_1 \geq \mu_2.$$

Tähistades vahe keskväärtuse sümboliga μ_d, saame hüpoteesid ümber kirjutada alljärgnevalt:

$$H_1 : \mu_d < 0,$$
$$H_0 : \mu_d \geq 0.$$

Valime olulisuse nivooks $\alpha = 0{,}01$.

Andmed on esitatud alljärgnevas tabelis. Esimene pilk tabelile näitab, et esimene püstitatud hüpotees on väga vähe tõepärane: kõigis valdades on töötealisi mehi rohkem kui naisi.

Tabel 15

Vald	*Tööealisi mehi X*	*Tööealisi naisi Y*	*Erinevus D = X--Y*
Aseri	*828*	*725*	*103*
Haljala	*967*	*879*	*88*
Harku	*1499*	*1409*	*90*
Kernu	*392*	*346*	*46*
Kohtla	*500*	*432*	*68*
Kullamaa	*523*	*412*	*111*
Nissi	*1015*	*868*	*147*
Saku	*1798*	*1763*	*35*
Sonda	*434*	*337*	*97*
Sõmeru	*1202*	*1088*	*114*

Seega võime isegi ilma arvutusi tegemata võtta esimese nullhüpoteesi vastu: pole võimalik tõestada, et tööealisi naisi oleks rohkem kui mehi.

Asume nüüd teise hüpoteesi tõestamise juurde. See on kahepoolne hüpotees, mille kontrollimine taandub alljärgneva hüpoteesipaari kontrollimisele:

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

Teeme vajalikud arvutused:

- *vahede keskväärtus on 89,9 ja standardhälve 33,22;*
- *t-statistiku väärtus, mille arvutame valemi (20) kohaselt, on 8,57;*
- *et arvutatud t-statistiku väärtus on märksa suurem vastavast kriitilisest väärtusest, saame sisuka hüpoteesi lugeda veenvalt tõestatuks.*

Kaht järeldust kokku võttes saame kinnitada, et Põhja-Eesti valdades on tööealiste meeste ja tööealiste naiste keskmised arvukused erinevad, kusjuures tööealisi mehi on keskmiselt rohkem.

Ka käesolevas näites oli oluline kasutada sõltuvate vaatluste metoodikat: on ju üsna loogiline, et asulates, kus elab rohkem mehi, elab ka rohkem naisi, st et need näitajad on omavahel seotud.

Kui me oleksime unustanud, et vaatlused on tehtud samades punktides, ja kasutanud sõltumatute vaatluste metoodikat, oleksime saanud ühiseks standardhälbe hinnanguks s = 480,25 ja t-statistiku väärtuseks 0,418, ning sisukat hüpoteesi poleks tõestada saanud. Niisuguse, olemasolevat lisainformatsiooni põhjendamatult eirava (me ju teame, et valimid on tegelikult sõltuvad) metoodika valik oleks põhjustanud eksijäreldusi või pannud uurija olukorda, kus ta hüpoteesi tõestamiseks peaks tegema täiendavat tööd (valimeid suurendama).

3.6. *Hüpoteeside kontrollimine tõenäosuste kohta*

3.6.1. Tõenäosuste võrdlemine

Oletame, et meil on kaks üldkogumit ning mõlema puhul on võimalik sündmuse A toimumine. Meid huvitab, kas selle sündmuse toimumise tõenäosus esimeses üldkogumis p_1 erineb selle sündmuse toimumise tõenäosusest teises üldkogumis p_2. Hüpoteeside kontrollimise eeskirja konstrueerimisel kasutame seda, et katsete arvu suurenemisel läheneb sündmuse suhtelise sageduse jaotus normaaljaotusele. Seega on saadav eeskiri asümptootiline, st annab seda usaldusväärsema tulemuse, mida suurem on valimi maht.

Tõstatatud küsimusele vastamiseks tuleb fikseerida olulisuse nivoo α ja valida välja üks hüpoteesipaar alljärgnevaist:

$$\begin{aligned} &H_0: p_1 = p_2, \\ &H_1: p_1 \neq p_2. \end{aligned} \qquad (25)$$

$$\begin{aligned} &H_0: p_1 \leq p_2, \\ &H_1: p_1 > p_2. \end{aligned} \qquad (26)$$

$$\begin{aligned} &H_0: p_1 \geq p_2, \\ &H_1: p_1 < p_2. \end{aligned} \qquad (27)$$

Milline neist hüpoteesipaaridest valitakse kontrollimiseks, sõltub ülesande sisulisest püstitusest. Kui eelinformatsioon puudub, lähtutakse tavaliselt lihtsast nullhüpoteesist (25).

Olgu meil olemas järgmine vaatlusandmestik: esimesest üldkogumist on tehtud n_1 vaatlust, millest tulemusega A lõppes k_1, ja teisest üldkogumist on tehtud n_2 vaatlust, millest tulemusega A lõppes k_2. Kasutades sarnast mõttekäiku nagu punktis 3.5.1, saame järgmise protseduuri.

- Leiame kummagi katseseeria jaoks suhtelise sageduse k_i/n_i.
- Kui nullhüpotees oleks õige, siis oleks ühine tõenäosuse hinnang $\bar{p} = (k_1 + k_2)/(n_1 + n_2)$.
- Leiame nullhüpoteesi kehtivust arvestava ühise dispersiooni hinnangu $\bar{p}(1 - \bar{p})$.
- Leiame ka valimi mahtusid arvestava teguri $\sqrt{1/n_1 + 1/n_2}$.
- Nüüd arvutame z-statistiku väärtuse

$$z = \left(\frac{k_1}{n_1} - \frac{k_2}{n_2}\right) \Big/ \sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Selle statistiku jaotus on meie eeldustel ligikaudselt standardiseeritud normaaljaotusega, ja seetõttu kasutataksegi püstitatud hüpoteesi kontrollimiseks normaaljaotuse tabelit (vt tabel 36). Sellest tabelist saame täiendkvantiilid leida esimesest veerust, kusjuures teises veerus on tõenäosused.

- Mugavaim on täiendkvantiile otsida tabeli parempoolsest osast, pidades silmas, et kui $P(X > q) = \alpha$, siis $P(X < q) = 1 - \alpha$, ning seetõttu paikneb 0,05-täiendkvantiil q arvude 1,6 ja 1,7 vahel. Meie arvutustäpsuse juures sobib kasutamiseks ligikaudne väärtus $q = 1{,}65$.
- Kahepoolse hüpoteesi tõestamiseks vajaliku tabeliväärtuse saame aga arvestades, et

$$P(|X| > q') = P(X > q') + P(X < -q').$$

Siit tuleneb X jaotuse sümmeetrilisuse tõttu vajadus määrata q' kui $(\alpha/2)$-täiendkvantiil. Meie juhul on tarvis leida 0,025-täiendkvantiil, mis asub 1,9 ja 2,0 vahel; täpsemalt on selle väärtus 1,96.

Edasi tuleb meil jälgida, missuguse väärtusega tuleb statistikut z võrrelda.

- Kahepoolse hüpoteesi tõestamiseks (juhul kui nullhüpoteesiks on lihthüpotees, vt valem (25)) leiame z absoluutväärtuse ja võrdleme väärtusega q'. Kui $|z| > q'$, siis võetakse vastu sisukas hüpotees H_1, vastasel korral jäädakse H_0 juurde: *tõenäosuste erinevust ei saa tõestada*. Muidugi ei järeldu siit, et *oleks tõestatud* tõenäosuste võrdsus üldkogumites.

Ühepoolse hüpoteesi puhul tuleb jälgida tõenäosuste vahe märki.

- Kui me tahame tõestada, et esimeses üldkogumis on sündmuse esinemise tõenäosus suurem (hüpoteesipaar (26)), siis peab z-statistik olema kindlasti positiivne ja küllalt suur, kusjuures z-statistiku väärtust võrdleme täiendkvantiiliga q. Kui $z > q$, siis võetakse vastu sisukas hüpotees H_1, vastasel korral jäädakse H_0 juurde: ei saa tõestada, et p_1 oleks suurem kui p_2. Muidugi ei järeldu siit veel nullhüpoteesis märgitud vastupidine võrratus tõenäosuste vahel. *Märgime, et kui arvutamise tulemuseks on negatiivne z väärtus, siis tuleb kindlasti vastu võtta nullhüpotees* (ilma, et oleks tarvis tabeleid kasutada).
- Kui me tahame tõestada, et esimeses üldkogumis on sündmuse esinemise tõenäosus väiksem (hüpoteesipaar (27)), siis peab z-statistik olema kindlasti negatiivne ja küllalt väike (absoluutväärtuselt suur!), kusjuures z-statistiku väärtust võrdleme täiendkvantiiliga q. Kui $z < -q$, siis võetakse vastu sisukas hüpotees H_1, vastasel korral jäädakse H_0 juurde: ei saa tõestada, et p_1 oleks väiksem kui p_2. Muidugi ei järeldu siit veel nullhüpoteesis märgitud vastupidine võrratus tõenäosuste

vahel. *Märgime, et kui arvutamise tulemuseks on positiivne z väärtus, siis tuleb kindlasti vastu võtta nullhüpotees* (ilma, et oleks tarvis tabeleid kasutada).

Näide 54

Olgu meil tarvis teada, kas elanikkonna vähenemise tõenäosus Põhja- ja Lõuna-Eesti valdades on erinev. Et meil pole eelteavet, sõnastame kahepoolse sisuka hüpoteesi, milles väidame, et vähenemise tõenäosus on erinev. Lepime kokku, et olulisuse nivoo on 0,05.

Vaatlusandmed paiknevad alljärgnevas tabelis, kus arvuga 1 on märgitud meid huvitava sündmuse (elanikkonna vähenemine) toimumine ja arvuga 0 mittetoimumine.

Tabel 16

Põhja-Eesti		*Lõuna-Eesti*	
Vald	***Vähenemine***	***Vald***	***Vähenemine***
Anija	*0*	*Kanepi*	*0*
Kiili	*0*	*Mooste*	*1*
Loksa	*1*	*Värska*	*0*
Vasalemma	*1*	*Konguta*	*1*
Avinurme	*1*	*Meeksi*	*0*
Lohusuu	*0*	*Tähtvere*	*0*
Mäetagus	*0*	*Võnnu*	*0*
Toila	*0*	*Rõngu*	*1*
Lihula	*1*	*Helme*	*1*
Risti	*0*	*Põdrala*	*1*
Avanduse	*1*	*Pärsti*	*0*
Rakke	*1*	*Saarepeedi*	*1*
Kehtna	*0*	*Meremäe*	*1*
		Mõniste	*1*

Läbime nüüd samm-sammult hüpoteeside kontrollimise protseduurid.

- *Leiame kummagi üldkogumi jaoks tõenäosuse hinnangu: Põhja-Eestis on vähenemise suhteline sagedus 6/13 = 0,462, Lõuna-Eestis 8/14 = 0,571.*
- *Ühine tõenäosuse hinnang, mis kehtiks siis, kui nullhüpotees oleks õige, on 14/27 = 0,5185.*
- *Seda hinnangut kasutades arvutame ruutjuure väärtuse – standardvea hinnangu – ja saame 0,1925.*
- *Tõenäosuse hinnangute vahe on –0,110*
- *z-statistiku väärtuseks saame –0,57.*

Et z-statistik on absoluutväärtuselt palju väiksem kui 1,96, siis tuleb vastu võtta nullhüpotees: tõenäosuste erinevust meie andmetel ei õnnestu tõestada.

3.6.2. Valimi mahu planeerimine esmase uuringu põhjal

Sageli juhtub, et hüpoteeside kontrollimise tulemus uurijat ei rahulda. Nii on see siis, kui andmed näitavad sisuka hüpoteesi suunalise tendentsi olemasolu, kuid see on liiga nõrk, et sisukat hüpoteesi tõestatuks lugeda. On aga teada, et valimi mahu suurenedes statistikute hajuvus üldiselt väheneb võrdeliselt ruutjuurega valimi mahust. See annab alust loota, et kui sisukas hüpotees üldkogumis kehtib, siis võib valimit küllalt palju suurendades jõuda olukorrani, kus sisukas hüpotees õnnestub tõestada.

Illustreerimaks öeldut vaatleme viimast näidet. Ka selle puhul olid tõenäosuse hinnangud üldkogumites erinevad, kuid erinevus oli sisuka väite tõestamiseks liiga väike. Oletame nüüd, et me suurendame mõlemat valimit c korda ja selle juures jäävad tõenäosuse hinnangud täpselt samaks (tavaliselt see küll nii ei ole, kuid selle oletusega me teeme enesele elu kergemaks). Siis suureneb z-statistiku väärtus $\sqrt{c}$ korda. Selleks, et sisukat hüpoteesi vastu võtta, peaks z väärtus olema vähemalt 1,96. Seega saame c määramiseks avaldise: $c = (1{,}96/0{,}57)^2 = 12$.

Näeme, et vajaliku hüpoteesi tõestamiseks peaks valimi maht olema Põhja-Eestis 156 ja Lõuna-Eestis 168. Jõudsime absurdse tulemuseni, sest saadud numbrid ületavad vastavate üldkogumite mahud. Niisugusel puhul tuleb väga tõsiselt kahelda, kas sisukas hüpotees üldkogumis üldse kehtib. Igatahes pole erilist lootust seda tõestada valikuuringu tulemusena.

4. Tunnustevahelised seosed. Jaotuste võrdlemine

4.1. Kahe tunnuse ühisjaotus

4.1.1. Kahe juhusliku suuruse ühisjaotus

Vaatleme kaht tunnust X ja Y ning eeldame, et meil on statistiline kogum, mis sisaldab mõlema tunnuse väärtusi n punkti jaoks. Seni oleme selgitanud, kuidas saame uurida nende tunnuste jaotusi ükshaaval. Tekib küsimus, kas me saame täiendavat teavet, kui me uurime kaht tunnust üheskoos.

Üldiselt on vastus jaatav. Mitut tunnust koos uurides on võimalik avastada ka nende tunnuste omavahelisi seoseid. Kahe tunnuse kohta annab kogu olemasoleva teabe edasi nende *ühisjaotus*.

Kahe tunnuse ühisjaotus esitatakse enamasti tabelina, kus ühe tunnuse väärtused/väärtusklassid määravad rea ja teise tunnuse väärtused/väärtusklassid veeru. Kirjeldame alljärgnevas kahe tunnuse ühisjaotuse tabeli konstrueerimist.

- Tähistame esimese tunnuse tähega X ja tema väärtused sümboliga x_i, $i = 1, ..., k$. Kui juhuslik suurus X on pidev, siis klassifitseerime väärtused. Tähistame i-nda klassi piirid a_{i-1}, a_i ja lepime kokku, et klasside arv on k. Klassipiiride suhtes tuleb teha erikokkulepe, näiteks, et igasse klassi kuulub selle alumine, kuid mitte ülemine piir. Erandiks võib olla viimane klass.
- Tähistame teise tunnuse tähega Y ja tema väärtused sümboliga y_j, $j = 1, ..., h$. Kui juhuslik suurus Y on pidev, siis klassifitseerime väärtused. Tähistame j-nda klassi piirid b_{j-1}, b_j ja lepime kokku, et klasside arv on h.
- Tabeli esimesse veergu paigutatakse tunnuse X väärtused või väärtusklassid ja esimesse ritta tunnuse Y väärtused või väärtusklassid. Klassifitseerimiseks kasutatakse põhimõtteliselt samu võtteid nagu üksiktunnuse jaotus- või sagedustabeli puhulgi.
- Alustuseks on mõttekas konstrueerida *kahemõõtmeline* ehk *ühissagedustabel*, mille abil on hiljem lihtne leida *kahemõõtmelist* ehk *ühisjaotustabelit*. Ühissagedustabelis esitatakse i-nda rea ja j-nda veeru ristumiskohas paiknevas lahtris nende objektide arv, mille korral tunnus X omandab väärtuse x_i ja tunnus Y väärtuse y_j. Seda arvu (sagedust) tähistatakse tähega n_{ij}.

- Sagedustabelist saadakse jaotustabel, jagades kõigi lahtrite sisu valimi mahuga n. Seega saame i-nda rea ja j-nda veeru ristumiskohta vastava väärtuspaari esinemise *suhteline sageduse* (mis kõikse uuringu korral võrdub tõenäosusega, valikuuringu korral aga on tõenäosuse hinnanguks). Suhtelise sageduse tähiseks on p_{ij}.
- Kahemõõtmelise sagedustabeli viimane rida ja viimane veerg esitavad tunnuste *ääre-* ehk *marginaalsagedusi*. Kahemõõtmelise jaotustabeli viimases reas ja veerus on vastavate tunnuste *ääre-* ehk *marginaaljaotused*. Marginaalsagedused ja marginaaljaotused ühtivad vastavate tunnuste (ühemõõtmeliste) sageduste- või jaotustega.

Alljärgnevad tabelid 17 ja 18 kujutavad kahemõõtmelist sagedus- ja jaotustabelit.

Tabel 17

$X\backslash Y$	y_1	...	y_h	X marginaalsagedus
x_1	n_{11}	...	n_{1h}	$n_{1.}$
...	...	...	...	...
x_k	n_{k1}	...	n_{kh}	$n_{k.}$
Y marginaalsagedus	$n_{.1}$	...	$n_{.h}$	n

Tabel 18

$X\backslash Y$	y_1	...	y_h	X marginaaljaotus
x_1	p_{11}	...	p_{1h}	$p_{1.}$
...	...	...	...	...
x_k	p_{k1}	...	p_{kh}	$p_{k.}$
Y marginaaljaotus	$p_{.1}$	...	$p_{.h}$	p

Kahemõõtmelise sagedus- ja jaotustabeli käsitlus ja tõlgendus sõltub sellest, kas on tegemist kõikse või valikuuringuga.

- Kui uuring on kõikne, siis esitab jaotustabel tunnuspaari (X,Y) ühisjaotust. Vastavalt sellele, kas kasutatakse tunnuste üksikväärtusi (diskreetse tunnuse puhul) või klassifitseeritud väärtusi (juhtum, kui tunnuse väärtused muutuvad pidevalt või on erinevaid väärtusi väga palju), on esitatud jaotus täpne või ligikaudne.
- Kui on tegemist valikuuringuga, siis esitab saadud jaotustabel tunnuspaari *valimi ühisjaotust*. See jaotus on *hinnanguks* tunnuspaari (või juhusliku vektori) (X,Y) ühisjaotusele üldkogumis ehk *teoreetilisele ühisjaotusele.*

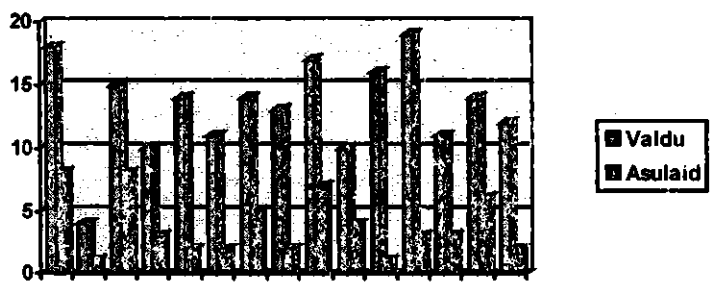
Kui mõlemad juhuslikud suurused X ja Y on pidevad, siis on ka juhuslik vektor (X,Y) pidev ja tema jaotust saab kõige sobivamalt esitada *kahemõõtmelise tihedusfunktsiooni* abil.

Näide 55

Vaatleme Eesti maakondi ning teeme igas maakonnas kindlaks elanike arvu, valdade arvu ja linnaliste asulate arvu. Saame alljärgneva tabeli 19, kus objektiks on maakond ja iga objekti kohta on teada kolm tunnust. Nende hulgast valime analüüsimiseks kaks diskreetset tunnust – valdade arvu X ja asulate arvu Y.

Tabel 19

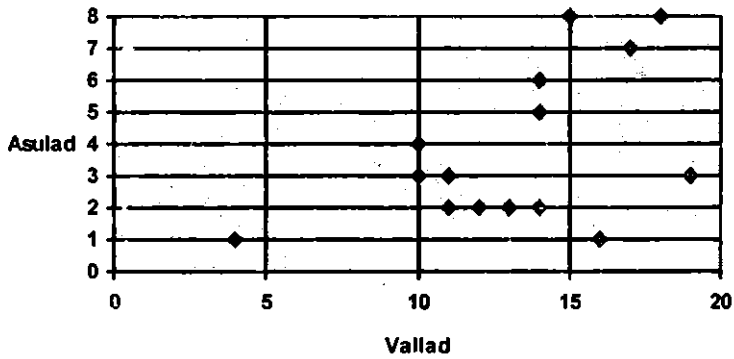
Maakond	Elanikke	Valdu	Asulaid
Harjumaa	551136	18	8
Hiiumaa	11905	4	1
Ida-Virumaa	202903	15	8
Jõgevamaa	41 845	10	3
Järvamaa	43562	14	2
Läänemaa	32228	11	2
Lääne-Virumaa	75 565	14	5
Põlvamaa	36375	13	2
Pärnumaa	99612	17	7
Raplamaa	40 028	10	4
Saaremaa	40512	16	1
Tartumaa	153 228	19	3
Valgamaa	39552	11	3
Viljandimaa	63813	14	6
Võrumaa	44137	12	2



Joonis 14

Joonisel 14 on kujutatud kahe tunnuse (valdade arv ja asulate arv) ühine tulpdiagramm, mis on põhimõtteliselt samane ühe tunnuse tulpdiagrammiga ja millest on üsna raske midagi tunnustevaheliste seoste kohta välja lugeda.

Joonisel 15 on kujutatud nende tunnuste korrelatsiooniväli ehk hajuvusdiagramm, mis põhimõtteliselt erineb varemvaadeldud graafikutest. Siin vastab iga puunkt ühele maakonnale ning koordinaatteljed tähistavad tunnuseid (traditsiooniliselt horisontaaltelg X- ja vertikaaltelg Y-tunnust).



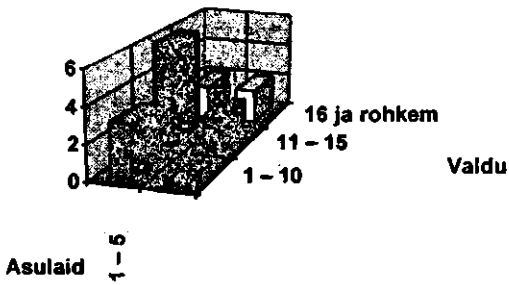
Joonis 15

Andmete kompaktsemaks esitamiseks klassifitseerime tunnused. Saame klassifitseeritud sagedustabeli, kus tunnusel Asulaid on kaks ja tunnusel Valdu kolm väärtusklassi, vt tabel 20.

Tabel 20

Asulaid\Valdu	1 – 10	11 – 15	16 ja rohkem	Kokku
1 – 5	3	6	2	11
6 ja rohkem	0	2	2	4
Kokku	3	8	4	15

Kahemõõtmelise sagedus- ja jaotustabeli illustreerimiseks sobib hästi ruumiline tulpdiagramm (vt joonis 16), millel tulpade paiknemine vastab tabeli lahtritele ja tulpade kõrgused on võrdelised vastavalt sageduste või tõenäosuste ehk suhteliste sagedustega.



Joonis 16

4.1.2. Tinglikud jaotused

Sageli pakub huvi vaadelda ainult üht osa üldkogumist, kusjuures see osa on määratletud mingi tunnuse abil fikseeritud tingimusega. Näiteks huvitagu meid üksnes sellised maakonnad, kus on 11–15 valda, s.o keskmise suurusega maakonnad. Nende andmed on kõik tabeli *20* ühes, keskmises veerus, ning siit me näeme, et selliseid maakondi on kaheksa. Saame ka leida asulate jaotuse keskmise suurusega maakondades: kuues neist on kuni viis asulat, kahes – üle viie. Seega saame leida asulate arvu jaotuse tingimusel, et maakonnas on valdu 11–15. See jaotus on:

$$P(\text{asulaid on 1–5}) = 3/4; \quad P(\text{asulaid on üle 5}) = 1/4.$$

Marginaaltõenäosused saadakse vastavalt rea- või veerutõenäosuste summeerimisel,

$$p_{i\cdot} = \sum_{j=1}^{h} p_{ij}, \qquad p_{\cdot j} = \sum_{i=1}^{k} p_{ij}.$$

Tinglikud jaotused saadakse jaotustabelist marginaaltõenäosuste kaudu. Kui $X = x_i$, siis saame *Y tingliku tõenäosusfunktsiooni*:

$$P(Y = y_j \mid X = x_i) = \frac{p_{ij}}{p_{i\cdot}}, \quad j = 1,\ldots, h, \tag{28}$$

kusjuures tinglike tõenäosuste summa (reasumma) võrdub ühega. Samal viisil saame X tingliku tõenäosusfunktsiooni Y suhtes,

$$P(X = x_i \mid Y = y_j) = \frac{p_{ij}}{p_{\cdot j}}, \quad i = 1,\ldots, k, \tag{28'}$$

millele vastav veerusumma võrdub ühega.

Praktilises töös pakuvad *tinglikud jaotused* (tõenäosusfunktsioonid) tihti suurematki huvi kui ühisjaotus. Oluline on ka see, et tingimuse formuleerimisel pole tabeli päises antud klassi kasutamine ainus võimalus, on võimalik ka klasse ühendada, näiteks vaadelda tunnuse X jaotust tingimusel $Y > y_j$. Kõige üldisem juhusliku suuruse Y abil formuleeritud tingimus on aga $Y \subset A$, kus A on juhusliku suuruse Y mingi väärtuste hulk.

Kui uuritava tunnuse X jaotus üldkogumis on pidev, siis on pidevad ka kõik tunnuse X tinglikud jaotused, kusjuures tingimust määrav tunnus Y ei tarvitse olla pidev. Sel juhul esitavad tinglikke jaotusi tinglikud tihedusfunktsioonid. Kui tingimust määrav tunnus Y on pidev, siis on tavaliselt tingimuseks, et tunnuse Y väärtus kuulub mingisse hulka, näiteks $Y > 0$.

4.1.3. Üldine statistiline sõltuvus tunnuste vahel

Vaatleme tunnuse X tinglikke jaotusi (tihedusfunktsioone) tingimustel $Y = y_j$. Kui need on kõik ühesugused, siis ilmselt tunnus Y ei mõjusta tunnuse X jaotust, ning selle kohta öeldakse, et X on sõltumatu Y-st. Arusaadavalt on siis samasugune ka X marginaaljaotus, ja me saame võrduse $P(X = x_i|Y = y_j) = P(X=x_i)$ ehk $p_{ij}/p_{\cdot j} = p_{i\cdot}$ iga i ja j korral. Viimase seose saame ümber kirjutada kujul

$$p_{ij} = p_{i\cdot}\, p_{\cdot j}, \quad i = 1, \ldots, k\,;\, j = 1, \ldots, h.$$

See on (diskreetsete) juhuslike suuruste X ja Y sõltumatuse tingimus. Et saadud avaldis on indeksite i ja j suhtes sümmeetriline, siis me saame järeldada, et statistiline sõltumatus on vastastikune: kui X ei sõltu Y-st, siis ei sõltu Y ka X-st.

Kui juhuslikud suurused X ja Y ei ole sõltumatud, siis nad on statistiliselt sõltuvad. Tunnuse X sõltuvus Y-st väljendub selles, et leidub kaks hulka A ja B nii, et X tinglik jaotus tingimusel $Y \subset A$ ja X tinglik jaotus tingimusel $Y \subset B$ ei ühti.

Statistiline sõltuvus on vastastikune: kui X sõltub Y-st, siis sõltub Y ka X-st.

4.1.4. Statistilise sõltuvuse (seose) tugevuse mõõtmine seosekordajate abil

Statistilise sõltuvuse puhul pole tähtis mitte üksnes selle olemasolu, vaid ka selle tugevus. Kõige tugevam on statistiline sõltuvus sel juhul, kui ühe juhusliku suuruse väärtust teades on võimalik üheselt kindlaks teha ka teise juhusliku suuruse väärtus. See tähendab, et üks juhuslik suurus on teise kaudu *täielikult prognoositav*. Diskreetse juhusliku suuruse puhul

realiseerub see võimalus näiteks siis, kui ühisjaotus on diagonaalne, st kui kehtivad seosed

$$p_{ij} = \begin{cases} p_{i\cdot} = p_{\cdot j}, & \textit{kui } i = j \\ 0, & \textit{muidu.} \end{cases}$$

Statistiline sõltuvus on kõige nõrgem siis, kui sõltuvus üldse puudub, st kui vaadeldavad tunnused või juhuslikud suurused on *sõltumatud*. Statistilise seose tugevust mõõdetakse seosekordajate abil. Tavaks on seosekordajad defineerida nii, et tugevaima statistilise sõltuvuse korral on seosekordaja väärtus 1, sõltumatuse puhul 0 ning muudel juhtudel nulli ja ühe vahel. Mida tugevam on seos, seda suurem on seosekordaja väärtus. Statistilise seose tugevuse mõõtmiseks kasutatakse mitmesuguseid seosekordajaid. Üks tuntumaid seosekordajaid on *Crameri seosekordaja V*, mille arvutamisvalem sagedustabeli põhjal on alljärgnev:

$$V = \sqrt{\frac{H}{n(\min(k,h)-1)}},$$

$$\text{kus } H = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{(n_{ij} - \frac{n_i n_j}{n})^2}{\frac{n_i n_j}{n}}. \qquad (29)$$

Märgime, et H avaldist saab mitmeti teisendada. Üks sageli esinevaid kujusid on alljärgnev, hästi meeldejääv valem, kus E tähistab *empiirilist* (st mõõdetud) sagedust ja T – *teoreetilist sagedust*:

$$\mathrm{H} = \sum_{i=1}^{k}\sum_{j=1}^{h} \frac{(E_{ij} - T_{ij})^2}{T_{ij}}. \qquad (29')$$

Tunnuste sõltumatuse korral on teoreetiliseks ühisjaotuseks empiiriliste marginaaljaotuste korrutis, st kehtib võrdus $T_{ij} = (n_{i\cdot}\, n_{\cdot j})/n$ ja $E_{ij} = n_{ij}$. Kui juhuslikud suurused X ja Y on statistiliselt sõltumatud, siis on summas H kõigi liidetavate lugejad võrdsed nulliga ja $V = 0$. Mida rohkem erinevad liidetavate lugejad nullist, seda tugevam on sõltuvus juhuslike suuruste vahel. Kordaja V omandab siis aina suuremaid väärtusi. Maksimaalne on sõltuvus siis, kui $k=h$ ja ühisjaotus on diagonaalne. Sel korral saab Crameri V võrdseks ühega.

Näide 56

Jätkame näidet 55 ja mõõdame, kui tugev on sõltuvus valdade arvu ja asulate arvu vahel maakonnas. Arvutame selleks Crameri seosekordaja V. Koondame arvutused alljärgnevasse tabelisse 21.

Tabel 21

E_{ij}	T_{ij}	$E_{ij}-T_{ij}$	$(E_{ij}-T_{ij})^2$	$(E_{ij}-T_{ij})^2/T_{ij}$
3	2,2	0,8	0,64	0,290909
6	5,87	0,13	0,017	0,002896
2	2,93	-0,93	0,865	0,295222
0	0,8	-0,8	0,64	0,8
2	2,13	-0,13	0,017	0,007981
2	1,07	0,93	0,865	0,808411
				2,205419

Viimase veeru summa annabki suuruse H. Et käesoleval juhul min(k,h) =2 ja n= 15, siis saame V väärtuseks:

$$V=\sqrt{\frac{2,2054}{15}}=0,388.$$

Lisaks Crameri kordajale *V* leidub veel rida teisigi seosekordajaid, mis mõõdavad statistilise seose tugevust. Praktiliselt kõik nad muutuvad 0 ja 1 vahel, omandades väärtuse 0 sõltumatuse korral ja 1 täieliku sõltuvuse korral. Nende vahelised erinevused ei ole enamasti tõlgendamise seisukohalt eriti olulised.

Kui oleme seose tugevust mõõtnud valimi ja mitte üldkogumi põhjal, tekib küsimus – kas seos tunnuste vahel eksisteerib ka üldkogumis? Sellele küsimusele vastamiseks tutvume uue jaotuste klassiga.

4.1.5. Hii-ruut (χ^2-) jaotus

Statistiliste hüpoteeside kontrollimisel on reeglina tarvis mingi arvutatud statistiku väärtust võrrelda mingi tuntud jaotuse kriitiliste väärtustega, mis leitakse kas statistikatabelitest või arvutatakse standardsete statistikaprogrammide abil. Kõige sagedamini on nii, et *nullhüpoteesi kehtimise korral peaks vaadeldav statistik olema mingi hästituntud teoreetilise jaotusega, mille täiendkvantiilid on tabuleeritud.* Kui aga nullhüpotees ei kehti, pole arvutatud statistik oodatava jaotusega. Tihti avaldub see mittekooskõla nii, et arvutatud statistiku väärtused on liiga suured, ületades tabelis paiknevaid täiendkvantiile, mida sel juhul ka *kriitilisteks väärtusteks* nimetatakse, sest nende kaudu toimib *hüpoteeside kontrollimise kriteerium.*

Probleemiks on siinjuures see, et iga ülesannete klassi jaoks on tarvis arvutada üldiselt erinev statistik ja ka selle teoreetiline jaotus on üldiselt erinev. Seni oleme tutvunud normaal- ja t-jaotuse tabelitega. Käesolevas peatükis lahendatavate ülesannete jaoks on tarvis teoreetiliste jaotuste varu täiendada hii-ruut- (χ^2-) jaotusega.

Olgu juhuslikud suurused X_1, ..., X_n standardiseeritud normaaljaotusega ja sõltumatud. Siis summa $Y_n = \sum_{i=1}^{n} X_i^2$ on juhuslik suurus, mille kohta öeldakse, et ta on χ^2 -jaotusega vabadusastmete arvuga n. Definitsioonist järeldub, et

- χ^2-jaotusega juhusliku suuruse väärtused on alati positiivsed;
- mida suurem on vabadusastmete arv, seda suuremad on üldiselt ka χ^2-jaotusega juhusliku suuruse väärtused.

χ^2-jaotusega juhusliku suuruse täiendkvantiilid on tabuleeritud (vt tabel 38 Lisas).

4.1.6. Statistilise seose olulisuse kontrollimine

Kui meil on tegemist valimiga üldkogumist ning valimis on uuritavad tunnused statistiliselt sõltuvad, siis pakub huvi kontrollida, kas see *sõltuvus on statistiliselt oluline*, st, kas *vaadeldavad tunnused on sõltuvad ka üldkogumis*.

Meenutame, et kõikse uuringu korral niisugusel protseduuril ei ole mõtet, sest kõikse uuringu puhul on igasugune sõltuvus statistiliselt oluline (kuigi mitte alati praktilist huvi pakkuv).

Statistilise seose olulisuse kontrollimiseks saab kasutada matemaatilisele statistikale omast skeemi: oletame, et juhuslikel suurustel X ja Y on üldkogumis ühisjaotus P_{XY}. Meil on aga teada valim (mahuga n) sellest üldkogumist ning valimi põhjal moodustame sagedustabeli (vt tabel 17), mis annab hinnangu tunnuspaari (X,Y) ühisjaotusele üldkogumis. Vastamaks küsimusele, kas juhuslikud suurused X ja Y on üldkogumis sõltuvad või sõltumatud, tuleb kontrollida järgmist hüpoteesipaari.

H_0: X ja Y on sõltumatud,
H_1: X ja Y on statistiliselt sõltuvad.

Fikseerime olulisuse nivoo α ja leiame statistiku, mida saaks kasutada selle hüpoteesipaari kontrollimiseks. Vaatleme statistikut H, mis on arvutatud valemist (29) või (29'). Valemitest on näha, et statistik H on $k \times h$ ühesuguse

kujuga liidetava summa. Analüüsime neid liidetavaid. Igaühes neist on empiirilise ja teoreetilise sageduse vahe, mida võiksime ka kujutleda empiirilise sageduse hälbena tema keskväärtusest (seda ju teoreetiline sagedus on). See vahe on võetud ruutu ja jagatud omakorda teoreetilise sagedusega. Seda võime vaadelda ka nii, et vahe on jagatud ruutjuurega teoreetilisest sagedusest, st normeeritud ja seejärel võetud ruutu. Seega on meil ligikaudselt tegemist standardiseeritud juhuslike suuruste ruutude summaga. Teoreetiliselt on tõestatud, et niisugune summa läheneb valimi mahu n kasvamisel hii-ruut jaotusele vabadusastmete arvuga $f = (k-1)(h-1)$.

Seda tõsiasja saame kasutada püstitatud hüpoteeside kontrollimise eeskirja tuletamiseks. Statistiku H konstrueerimisel on teoreetiline sagedus arvutatud eeldusel, et tunnused on sõltumatud. Järelikult on statistik H nimelt siis χ^2-jaotusega, kui nullhüpotees kehtib. Tähistame χ^2-statistiku kriitilise väärtuse olulisuse nivoo α ja vabadusastmete arvu f korral sümboliga $h(\alpha,f)$. Seega saame hüpoteeside kontrollimiseks alljärgneva eeskirja.

- Kui $H > h(\alpha, f)$, siis võtame vastu H_1 ja loeme tõestatuks, et juhuslike suuruste X ja Y vahel on üldkogumis statistiline sõltuvus, ehk ütleme, et statistiline sõltuvus on oluline olulisuse nivool α.
- Kui $H \le h(\alpha, f)$, siis jääme nullhüpoteesi juurde. Meil ei õnnestunud tõestada statistilise seose olulisust (antud olulisuse nivool).

4.2. *Jaotuste võrdlemine*

4.2.1. Empiiriliste jaotuste võrdlemine

Tunnuse võrdlemisel kahes üldkogumis on üks võimalus teha kindlaks, kas selle tunnuse keskväärtused on mõlemas kogumis võrdsed. Selle ülesande lahendamisega tegelesime paragrahvis 3.5. Kuid mõnikord võivad tunnused erineda oma jaotuse poolest, kusjuures see keskväärtuste erinevuses ei tarvitsegi kajastuda. Käesolevas punktis soovime kontrollida üldisemat hüpoteesi jaotuste erinevuse kohta: kas tunnuse jaotus erinevates üldkogumites on erinev?

Olgu antud kaks üldkogumit ja mõõdetud sama tunnust nendes üldkogumites. Tunnusel on k väärtust. Esimesest üldkogumist on võetud valim, milles on väärtuste sagedused vastavalt $n^1_1, \ldots, n^1_k$ ja teisest valim sagedustega $n^2_1, \ldots, n^2_k$, kus valimimahud on vastavalt $n_i = n^i_1 + \ldots + n^i_k$, $i = 1$ või 2.

Tähistades uuritava tunnuse tõenäosusfunktsiooni i-ndas üldkogumis sümboliga p^i_j, kus $j = 1, ..., k$, saame meid huvitava küsimuse formuleerida alljärgneva hüpoteesipaarina:

$H_0: p^1_j = p^2_j$ iga j korral, $j = 1, ..., k$.
$H_1: p^1_j \neq p^2_j$ mingi väärtuse j korral, $j = 1, ..., k$.

Sisukas hüpotees osutub tõestatuks niipea, kui selgub, et kasvõi tunnuse ühe väärtuse korral on erinevate valimite põhjal saadud tõenäosushinnangud oluliselt erinevad.

Selle hüpoteesipaari kontrollimiseks kasutame χ^2-statistikut. Teoreetiliseks jaotuseks, mis vastab nullhüpoteesile, loeme jaotuse, mille saame mõlemaid valimeid ühendades. Seega tuleb meil püstitatud hüpoteesipaari kontrollimiseks teha järgmised arvutused.

- Leiame tunnuse iga väärtuse j jaoks tõenäosusfunktsiooni hinnangu nullhüpoteesile vastaval eeldusel,

$$\bar{p_j} = (n^1_j + n^2_j)/(n_1 + n_2).$$

- Leiame kõigi tunnuse väärtuste jaoks tegeliku sageduse ja teoreetilise sageduse erinevused mõlemas valimis: $n^i_j - n_i \bar{p_j}$.
- Võtame leitud vahed ruutu ja jagame teoreetilise sagedusega (normeerime).
- Summeerime saadud liikmed. Tulemusena saame nn hii-ruut-statistiku H:

$$H = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(n^i_j - n_i \bar{p}_j)^2}{n_i \bar{p}_j}. \qquad (30)$$

- Selle statistiku jaotus on juhul, kui nullhüpotees kehtib, χ^2-jaotus vabadusastmete arvuga $k-1$.
- Statistiku H abil saame kontrollida püstitatud hüpoteesipaari standardsel viisil.
- Kui arvutatud statistiku H väärtus on suurem kui χ^2-jaotuse vastav kriitiline väärtus (mille juures arvestame vabadusastmete arvu ja olulisuse nivood), siis kummutame nullhüpoteesi ja loeme tõestatuks sisuka hüpoteesi: *uuritavad tunnused on üldkogumites erineva jaotusega.*
- Kui H väärtus on vastavast protsentpunktist väiksem, jääme nullhüpoteesi juurde: *tunnuste jaotused erinevates üldkogumites ei erine kasutatud andmete põhjal oluliselt.*

Näide 57

Jaotame Eesti vallad elanike arvu poolest kolmeks: väiksed (alla 1500 elaniku), keskmised (1500–3000 elanikku) ja suured (üle 3000 elaniku). Küsime, kas valdade jaotus elanike arvu järgi on Põhja- ja Lõuna-Eestis sama. Kasutame vastuse leidmiseks näites 29 kasutusele võetud valimit. Valimi (tabel 22) põhjal saame elanike arvu jaoks alljärgneva sagedustabeli.

Tabel 22

	Väike	Keskmine	Suur	Kokku
Põhja-Eesti	2	2	7	11
Lõuna-Eesti	2	6	1	9
Kokku	4	8	8	20

Teeme nüüd samasuguse tabeli ka nn teoreetiliste sageduste $\bar{p}_i n_i$ jaoks (sel juhul, kui vallaelanike arvu jaotus oleks Põhja- ja Lõuna-Eestis ühesugune).

Tabel 23

	Väike	Keskmine	Suur	Kokku
Põhja-Eesti	2,2	4,4	4,4	11
Lõuna-Eesti	1,8	3,6	3,6	9
Kokku	4	8	8	20

Järgmiseks teeme tabeli, millesse kanname statistiku H arvutamise jaoks vajalikud vahetulemused üksiklahtrite kaupa.

Tabel 24

Lahter	$\bar{p}_i n_i - n'_i$	$(\bar{p}_i n_i - n'_i)^2$	$(\bar{p}_i n_i - n'_i)^2 / \bar{p}_i n_i$
P.-Eesti väike	0,2	0,04	0,018
P.-Eesti keskmine	2,4	5,76	1,309
P.-Eesti suur	-2,6	6,76	1,536
L.-Eesti väike	-0,2	0,04	0,022
L.-Eesti keskmine	-2,4	5,76	1,600
L.-Eesti suur	2,6	6,76	1,878
Kokku			6,363

Viimase veeru summeerimisel leiame H väärtuse 6,363. See suurus ületab parasjagu χ^2-jaotuse 0,05-täiendkvantiili vabadusastmete arvul 2, mille väärtus on 5,991, nagu näeme tabelist 38. Järelikult võime lugeda tõestatuks, et Lõuna- ja Põhja-Eesti valdade jaotus elanike arvu järgi on erinev.

Lõpuks tuleb veel tõele au andes tunnistada, et tehtud näide ei ole päris korrektne, sest χ^2-jaotus on korrektselt rakendatav üksnes siis, kui teoreetilised sagedused on kõik suuremad kui 4. Näeme tabelist 24, et meie

näites on kõik sagedused liiga väikesed, alles kolm korda suurema valimimahu korral saaksime piisavalt suure andmestiku selleks, et rakendada χ^2-jaotusele tuginevat eeskirja hüpoteeside kontrollimiseks.

Esitatud näites kasutatud lahendusskeem on rakendatav kasutamisjuhisena selliste ülesannete lahendamise puhul, kus on tarvis võrrelda sama tunnuse jaotust kahes üldkogumis või üldkogumi osas, kusjures kummastki osast on olemas valim. Siinjuures on oluline, et *valimjaotused kasutaksid samu klassifitseerimise eeskirju.*

4.2.2. Empiirilise jaotuse võrdlemine teoreetilise jaotusega

Et paljude statistikaprotseduuride puhul eeldatakse, et uuritava tunnuse jaotus üldkogumis on normaaljaotusega, siis pakub huvi selgitada välja eeskiri, kuidas seda oletust kontrollida. Veidi ootamatuna tundub see, et pole olemas lihtsat metoodikat selleks, et *tõestada, et mingi jaotus on normaaljaotusega* (või ka mingi teise etteantud jaotusega), küll aga saab tõestada seda, et jaotus erineb oluliselt antud teoreetilisest jaotusest.
Vaatleme hüpoteesipaari:

H_0: X on antud jaotusega P,
H_1: X ei ole antud jaotusega P.

Olgu antud olulisuse nivoo α ja valim juhusliku suuruse X väärtustest mahuga n.

Kontrollimaks püstitatud hüpoteesipaari teeme alljärgnevad sammud.

- Konstrueerime tunnuse X sagedustabeli, selleks vajaduse korral sobivat klassifitseerimiseeskirja rakendades (vt punkt 1.3.3). Märgime, et klasside arv ei tohi olla liiga väike (üldjuhul mitte alla 4) ja samuti ei tohiks klassides olla liiga vähe vaatlusi (rusikareegliks võiks olla, et mitte alla 5). Klassifitseerimise juures pole tähtis, et klassid oleksid ühepikkused, ka lahtised klassid on lubatud. Tähistame konstrueeritud k-klassilises sagedustabelis esinevad sagedused $n_1, \ldots, n_k$.
- Kasutades teadaolevat jaotust P, arvutame teoreetilised sagedused $n \cdot \bar{p}_j$, $j = 1, \ldots, k$.
- Arvutame statistiku H:

$$H = \sum_{j=1}^{k} \frac{(n_j - n\,\bar{p}_j)^2}{n\,\bar{p}_j}. \qquad (31)$$

- See statistik on nullhüpoteesi kehtimise korral χ^2-jaotusega vabadusastmete arvuga $k-r-1$, kus r on vaadeldava teoreetilise jaotuse valimi põhjal hinnatud parameetrite arv.

Näide 58

Kontrollime, kas elanike arv Eesti valdades on normaaljaotusega. Selle ülesande lahendamiseks peab lahtrite arv tabelis olema vähemalt 4, sest normaaljaotusel on 2 parameetrit. Kasutame selleks eelmise näite valimiga analoogset valimit ja moodustame 6-klassilise sagedustabeli:

Tabel 25

	..1500	1500...2000	2000...2500	2500...3000	3000..4000	4000..
Sagedus	4	4	5	0	3	4
Standardi-seeritud klassipiirid	...-0,794	-0,794... -0,475	-0,475... -0,156	-0,156... 0,163	0,163... 0,802	0,802...
Tõenäosus	0,214	0,103	0,127	0,12	0,308	0,128
Teoreetiline sagedus	4 28	2,06	2,54	2,4	6,16	2,56
Hii-ruut	0,018	1,827	2,383	2,4	1,621	0,81

- *Hindame valimi põhjal normaaljaotuse parameetreid – arvutame valimi keskmise $\bar{x}$ = 2744,4 ja standardhälbe hinnangu s = 1566. Eeldame edaspidi, et teoreetiliseks jaotuseks on normaaljaotus N(2744,4; 1566).*
- *Standardiseerime nende parameetrite abil sagedustabeli klassipiirid: (1500 - 2744)/1566 = -0,794; (2000 - 2744)/1566 = -0,475; (2500 - 2744)/1566 = -0,156; (3000 - 2744)/1566 =0,163; (4000 - 2744)/1566 = 0,802.*
  Saadud arvudega saame täita rea "Standardiseeritud klassipiirid".
- *Leiame standardiseeritud normaaljaotuse tabelist klassipiiridele vastavad tõenäosused: $\Phi(-0,794)=0,214$; $\Phi(-0,475)=0,317$; $\Phi(-0,156)=0,444$; $\Phi(0,163)=0,564$; $\Phi(0,802)=0,872$.*
- *Arvutame nüüd klasside teoreetilised tõenäosused $p_i=P(a_{i-1}<X_0<a_i)$, kus X_0 on standardiseeritud normaaljaotusega juhuslik suurus, lahutamise teel: $P(X_0 < -0,794) = 0,214$; $P(-0,794 < X_0 < -0,475) = 0,317 - 0,214 = 0,103$; ... $P(X_0 > 0,802) = 1 - 0,872 = 0,128$.*
  Nende arvutuste tulemusena saime rea "Tõenäosus".
- *Teoreetilised sagedused saame, korrutades teoreetilised tõenäosused läbi valimi mahuga. Arvutame need ja paigutame tabelisse ritta "Teoreetiline sagedus": $20 \cdot 0,214 = 4,28$; ...*
- *Seejärel arvutame samuti kui eelmises näites teoreetilise ja empiirilise sageduse vahe ruudu ja normeerime tulemuse, st jagame teoreetilise sagedusega. Vastavad arvud on toodud tabeli viimases reas "Hii-ruut".*
- *Liidame kokku tabeli viimases reas paiknevad arvud. Saame 9,059.*
- *Hii-ruut-statistiku vabadusastmete arvuks on 6 – 2 – 1 = 3.*
- *Leiame tabelist kriitilise väärtuse olulisuse nivool 0,05 ja vabadusastmete arvul 3; see on 7,815.*

Et arvutatud statistiku väärtus H on suurem kui tabelist leitud kriitiline väärtus, loeme tõestatuks, et elanike arv Eesti valdades ei ole normaaljaotusega juhuslik suurus.

5. Mudelid ja prognoosimine

5.1. Lineaarne regressioon ja korrelatiivne sõltuvus

5.1.1. Lineaarne sõltuvus arvtunnuste vahel

Vaatleme arvulisi juhuslikke suurusi X ja Y ning eeldame, et nad on *statistiliselt sõltuvad*. Statistilise sõltuvuse olemasolu tähendab ühtlasi seda, et ühe juhusliku suuruse väärtusi teades tekib võimalus teist *prognoosida*. Selleks on aga tarvis leida prognoosiv valem ehk *mudel*. Lihtsaim mudel on lineaarne,

$$Y = a + bX + \varepsilon, \qquad (32)$$

kus X ja Y on juhuslikud suurused, a ja b *mudeli parameetrid* (a – vabaliige ja b – regressioonikordaja) ning ε on mudeli jääkliige ehk viga.

Mudeli parameetrid tuleb määrata nii, et mudeli viga oleks võimalikult väike. Mudeli parameetreid on võimalik hinnata valimi põhjal. Üks võimalus parameetrite hindamiseks on kasutada vähimruutude meetodit, st määrata parameetrid a ja b nii, et summa

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - (a + bx_i))^2 , \qquad (32')$$

mis iseloomustab vaatluste hälbimist mudelist, omandaks minimaalse väärtuse. Selleks, et leida eeskirja a ja b määramiseks, tuleb lahendada *ekstreemumülesanne* parameetrite a ja b suhtes. Selle tulemusena saame a ja b määramiseks valemid:

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{ja} \quad a = \bar{y} - b\bar{x}. \qquad (33)$$

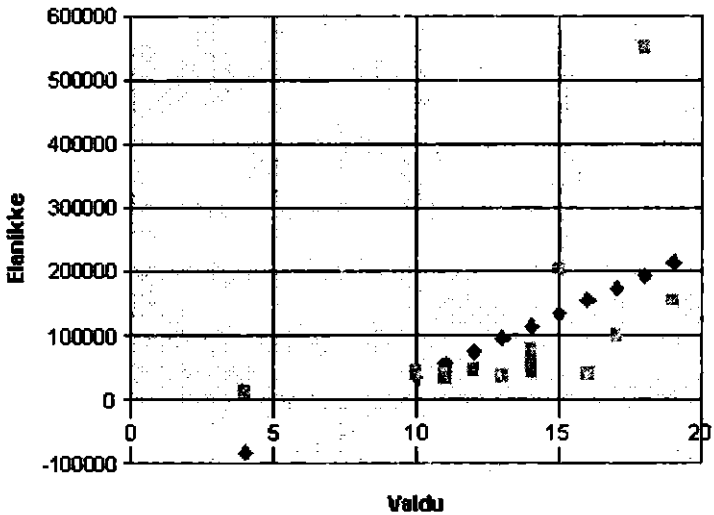
Sirget

$$y = a + bx$$

nimetatakse *regressioonisirgeks*.

Näide 59

Vaatleme tabelis 19 esitatud Eesti n-akondi, võttes tunnuseks X valdade arvu ja tunnuseks Y elanikkude arvu. Saame alljärgneva korrelatsioonivälja (vt joonis 17), kus valimi punkte tähistavad ruudud (mitte rombid):



Joonis 17

Seame enesele ülesandeks leida regressioonisirge, mis kirjeldaks elanike arvu sõltuvust valdade arvust. Selleks tuleb meil leida parameetrid a ja b. Regressioonikordaja b praktiliseks arvutamiseks on meil sobiv kasutada ülalesitatud valemit veidi teisendatud kujul:

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}.$$

Siin on punktideks maakonnad, st n=15. Selle valemi jaoks vajalikud arvutused on koondatud tabelisse 26, kus iga rida vastab ühele maakonnale (tähestikulises järjestuses). Esimeses veerus on maakonna elanike arv (tuhandetes), teises veerus valdade arv ja ülejäänud veergudes arvutusteks vajalikud vahetulemused. Kõige viimases reas on veerusummad.

Kasutades tabeli viimast rida, leiame keskmised $\bar{y} = 98{,}42$ (st maakonna keskmine elanike arv on veidi alla saja tuhande) ja $\bar{x} = 13{,}2$ (keskmiselt on maakonnas 13 valda).

Tabel 26

y_i	x_i	y_i^2	x_i^2	$x_i y_i$
551,1	18	303711,2	324	9919,8
11,9	4	141,61	16	47,6
202,9	15	41168,41	225	3043,5
41,8	10	1747,24	100	418
43,6	14	1900,96	196	610,4
32,2	11	1036,84	121	354,2
75,6	14	5715,36	196	1058,4
36,4	13	1324,96	169	473,2
99,6	17	9920,16	289	1693,2
40	10	1600	100	400
40,5	16	1640,25	256	648
153,2	19	23470,24	361	2910,8
39,6	11	1568,16	121	435,6
63,8	14	4070,44	196	893,2
44,1	12	1944,81	144	529,2
1476,3	198	400960,7	2814	23435,1

Paigutades tabelist leitud väärtused valemisse (33) saame:

$$b = (23435{,}1 - 15 \times 98{,}4 \times 13{,}2)/(2814 - 15 \times 13{,}22) = 3947{,}94/200{,}4 = 19{,}7.$$

See tähendab, et iga lisanduv vald tähendab maakonna elanike arvu kasvu ligemale 20 000 võrra. Teades, et keskmine vallaelanike arv on vaid kaks ja pool tuhat, võib saadav tulemus tunduda ootamatuna.

Vabaliikme a leidmiseks kasutame juba leitud b väärtust:

$$a = 98{,}42 - 19{,}7 \times 13{,}2 = -161{,}6.$$

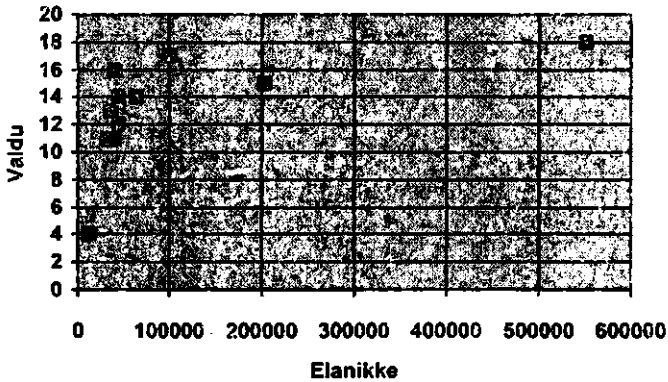
Seega saime maakonna elanike arvu prognoosimiseks valdade arvu kaudu järgmise valemi.

$$\text{Elanike arv} = 19{,}7 \times \text{valdade arv} - 161{,}6.$$

Püüame saadud valemit tõlgendada. Esmapilgul tekitab teatavat hämmingut ka see, et vabaliige on negatiivne. Vabaliikme negatiivsus tähendab üldiselt seda, et väga väike valdade arv (alla kümne) ei ole saadud mudeli seisukohast üldse reaalne. Alates teatavast piirist aga vastab igale lisanduvale vallale ligemale 20 000 elaniku suurune juurdekasv. Siin on oluline aga see, et suuremates maakondades, kus on rohkem valdu, on rohkem ka linnalisi asulaid, ning leitud mudel arvestab lisaks vallaelanikele ka linnaelanikke.

Kokkuvõttes tekib siiski mulje, et saadud mudel ei kirjelda elanike arvu ja valdade arvu vahelist seost kuigi hästi.

Lineaarne regressioonsõltuvus on *vastastikune*. Kui õnnestub moodustada tunnuse X lineaarne mudel tunnuse Y järgi, siis on ka vastupidine võimalik. Alljärgneval joonisel on toodud sama ülesande näitel korrelatsiooniväli, mis illustreerib valdade arvu sõltuvust elanike arvust.



Joonis 18

5.1.2. Lineaarne korrelatsioonikordaja

Lineaarse mudeli headust ehk lineaarse seose tugevust iseloomustab *lineaarne korrelatsioonikordaja r*, mille arvutusvalem on järgmine:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{j=1}^{n}(y_j - \bar{y})^2}}. \qquad (34)$$

Siin $\bar{x}$ ja $\bar{y}$ on tunnuste X ja Y keskmised ning x_i ja y_i on vastavate tunnuste väärtused i-ndas kogumi punktis.

Lineaarse korrelatsioonikordaja väärtused muutuvad erinevalt statistilise seose kordajast V intervallis $[-1, 1]$. Seega iseloomustab lineaarne korrelatsioonikordaja mitte üksnes seose tugevust (seda iseloomustab korrelatsioonikordaja absoluutväärtus), vaid ka seose suunda.

Kui tunnuse X väärtuste suurenedes ka tunnuse Y väärtused keskmiselt suurenevad, siis on juhuslikud suurused X ja Y *positiivselt korreleeritud.* Kui aga tunnuse X suurenedes tunnuse Y väärtused vähenevad ja vastupidi, siis on tunnuste X ja Y vaheline korrelatsioon *negatiivne.*

Sõltumatute juhuslike suuruste vaheline korrelatsioon on null (tunnused on *mittekorreleeritud*). Siiski leidub juhuslikke suurusi, mille vaheline korrelatsioonikordaja on null, kuigi tunnused ise pole sõltumatud. Teisisõnu, *leidub mittekorreleeritud, kuid sõltuvate tunnuste paare.* Lineaarse korrelatsioonikordaja ruut ehk *determinatsioonikordaja* näitab, kui suure osa ühe tunnuse hajuvusest saab kirjeldada teise tunnuse abil. Determinatsioonikordaja väärtused on 0 ja 1 vahel, tema väärtus avaldatakse tihti ka protsentides.

Näide 60

Arvutame eelmises ülesandes leitud lineaarse mudeli headust iseloomustava korrelatsioonikordaja väärtuse. Selleks saame kasutada regressioonikordaja arvutamisel tehtud eeltööd, kuid nimetaja leidmisel tuleb arvestada ka Y hajuvust.

$$r = 3947/(\sqrt{200,4 \times (400960,7 - 15 \times 98,42^2)}) = 3947/7157,9 = 0,55.$$

Seega saame determinatsioonikordaja väärtuseks $0,55^2 = 0,3$ ehk 30%. Valdade arv kirjeldab maakonna elanike arvust umbes 30%. Tulemus on üsna ootuspärane, sest vastab maaelanike osatähtsusele elanikkonna hulgas.

5.1.3. Lineaarse mudeli parandamine teisenduste abil

Mõnikord tekib mulje, et lineaarne mudel ei sobi antud nähtuse kirjeldamiseks kõige paremini. Siis aitab vahel lähtetunnuste teisendamine. Üks sagedamini kasutatavaid võimalusi on Y-tunnuse *logaritmiline teisendus*, mille tulemusena me saame mudeli

$$\log Y = a + bX.$$

Siin võib kasutada suvalist logaritmi alust, kuid kõige sagedamini mõistetakse logaritmina naturaallogaritmi, st logaritmi alusel e või ka kümnendlogaritmi. Selle mudeli tõlgendamiseks võtame mõlemast poolest antilogaritmi, st saame varasemaga samaväärse valemi

$$Y = A \times B^X,$$

kus konstandid A ja B avalduvad lineaarse regressiooni parameetrite a ja b kaudu: $A = 10^a$ ja $B = 10^b$. Niisugune mudel tähendab praktiliselt tunnuse Y väga kiiret kasvu tunnuse X väärtuste kasvades.

Näide 61

Jätkame eelmist näidet. Et siin elanike arv kasvas valdade arvuga võrreldes väga kiiresti, siis proovime kasutada Y logaritmilist teisendust. Saame uue Y-veeru, vt tabel 27:

Tabel 27

x_i	$\log(y_i)$
18	2,74
4	1,076
15	2,307
10	1,62
14	1,64
11	1,51
14	1,88
13	1,56
17	2
10	1,602
16	1,607
19	2,19
11	1,6
14	1,805
12	1,644

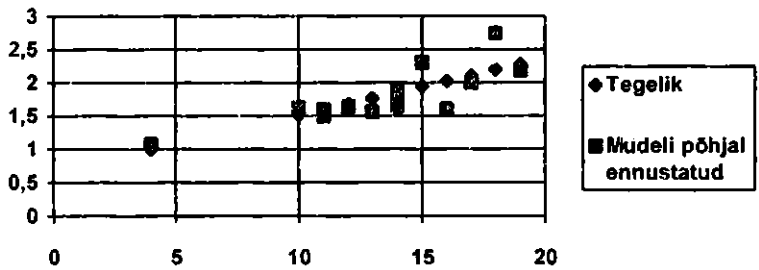
Leitud tabeli põhjal moodustame uue korrelatsioonivälja ja leiame uue mudeli. Saame seose:

$$\log Y = 0{,}6636 + 0{,}085\,X,$$

kusjuures korrelatiivne seos logY ja X vahel on märksa tugevam kui Y ja X vahel. Korrelatsioonikordaja väärtuseks on käesoleval juhul 0,81, millele vastab determinatsioonikordaja 0,66. Seega on logY kirjeldatus X lineaarse funktsioonina üle kahe korra parem kui Y kirjeldatus X lineaarse funktsioonina. Seda näeme ka korrelatsioonivälja lt joonisel 19, kus vertikaalteljel on kujutatud elanike arvu logaritm. Näeme, et siin asuvad prognoosipunktid mõõdetud punktidele märksa lähemal kui joonisel 17.

Loomulikult pole toodud teisendus ainus võimalik. Sageli kasutatakse ka teisendusi $Y \Rightarrow 1/Y$, $Y \Rightarrow e^Y$ jne. Samal viisil on võimalik teisendada ka tunnust X. Teisendamise juures tuleb arvestada aga teisendustega seostuvaid lisanõudeid, millest olulisimad on järgmised.

- Logaritmiline teisendus on võimalik üksnes positiivsete väärtustega tunnuse puhul.
- Pöördarvuks saab tunnust teisendada ainult siis, kui tal ei ole väärtust 0.



Joonis 19

Tähtis on teada ka seda, et teisendamise tulemusena *saadud mudel ei tarvitse olla optimaalne esialgsete tunnuste jaoks*, vaid ta on (vähimruutude mõttes) parim teisendatud tunnuste jaoks. Selletõttu ei ole tunnuste teisendamisel erilist mõtet siis, kui selle tulemusena mudel vaid veidi paraneb. Tihti aga aitab teisendamine väga lihtsate vahenditega mudelit oluliselt parandada, ning nimelt sellisel puhul on teisenduse rakendamine õigustatud.

5.1.4. Lineaarse mudeli ja lineaarse korrelatsioonikordaja olulisus

Kui lineaarne korrelatsioonikordaja on arvutatud valimi põhjal, siis tekib küsimus, kas valimis avastatud lineaarne sõltuvus kehtib ka üldkogumis. Selle kontrollimiseks tuleb meil valida olulisuse nivoo α ja sõnastada hüpoteesipaar

$H_0: r = 0$ üldkogumis;
$H_1: r \neq 0$ üldkogumis.

Kui nullhüpotees on õige, siis öeldakse, et *korrelatiivne seos ei ole statistiliselt oluline*, kui aga õige on sisukas hüpotees, siis öeldakse, et *korrelatiivne seos tunnuste X ja Y vahel on statistiliselt oluline*.

Nagu nägime eelmises peatükis, on hüpoteeside kontrollimiseks tarvis valimi põhjal arvutada mingi statistiku väärtus, kusjuures selle nn *teststatistiku* väärtus nullhüpoteesi kehtivuse korral peab olema teada ja tabuleeritud. Võrreldes arvutatud statistiku väärtust teoreetilise jaotuse tabelist leitavate kriitiliste väärtustega (enamasti täiendkvantiilidega), näeme, kas andmed on kooskõlas nullhüpoteesiga. Kui statistiku arvutatud väärtus on kriitilisest väärtusest suurem, pole andmed nullhüpoteesiga kooskõlas ja sisukas hüpotees loetakse tõestatuks. Vastasel korral jääb sisukas hüpotees tõestamata ning vastu võetakse nullhüpotees, kuid seda muidugi ei loeta tõestatuks.

Lineaarse korrelatsioonikordaja r puhul on olukord väga lihtne: teststatistikuks, mille väärtused on tabuleeritud, on valimi korrelatsioonikordaja. Korrelatsioonikordaja kriitilised väärtused on arvutatud eeldusel, et üldkogum on kahemõõtmelise normaaljaotusega ja sõltumatute komponentidega. Need kriitilised väärtused (täiendkvantiilid) on toodud tabelis 39. Nagu ka t-jaotuse korral, on tabelis eraldi osad ühepoolsete ja kahepoolsete hüpoteeside jaoks (vastavalt vasakpoolne ja parempoolne osa). Ühepoolse hüpoteesi kontrollimiseks olulisuse nivool α kasutatakse, samuti kui t-jaotuse puhulgi, α-täiendkvantiili $q(\alpha)$, mille puhul kehtib võrdus $P(r > q(\alpha)) = \alpha$. Kahepoolse hüpoteesi kontrollimisel kasutatakse statistiku sümmeetrilisust ja arvutatakse väärtus $q'(\alpha)$ nii, et kehtivad võrdused $P(r > q'(\alpha)) = \alpha/2$ ja $P(r < -q'(\alpha)) = \alpha/2$, mis kokku annab võrduse $P(|r| > q'(\alpha)) = \alpha$.

Tabelis paiknevaid väärtusi q ja q' nimetatakse ühise nimetusega *korrelatsioonikordaja kriitilisteks väärtusteks.* Paneme tähele, et korrelatsioonikordaja kriitilised väärtused sõltuvad ka valimi mahust n, kuid traditsiooni mõttes on tabeli rida määravaks argumendiks valitud mitte valimi maht, vaid sellest kahe võrra väiksem *vabadusastmete arv* $f = n - 2$.

Tähtis on teada seda, et *kui korrelatsioonikordaja ei ole oluline, pole oluline ka regressioonimudel tervikuna*, ning selle edaspidine kasutamine on ebakorrektne.

5.1.5. Ühepoolsete hüpoteeside kontrollimine lineaarse korrelatsioonikordaja kohta

Mõnikord pakub huvi ühepoolse hüpoteesi kontrollimine korrelatiivse seose kohta. Huvitagu uurijat olukord, kus korrelatiivne sõltuvus on positiivne. Sel juhul saame püstitada alljärgneva hüpoteesipaari:

$$H_0 : r \leq 0,$$
$$H_1 : r > 0,$$

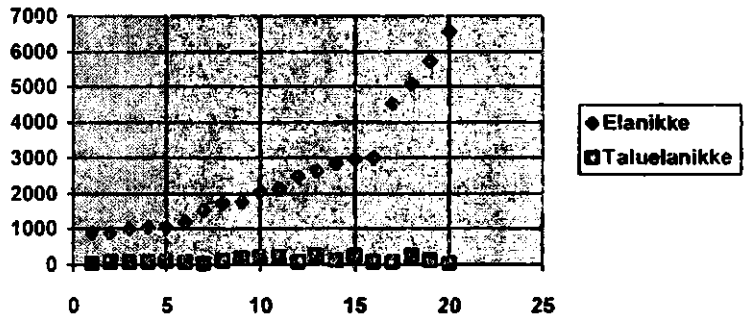
kus r tähistab korrelatsioonikordajat üldkogumis. Hüpoteeside kontrollimine toimub samuti kui eelnevalgi juhul, ainsaks erinevuseks on, et kriitilise väärtuse võtame tabeli sellest osast, mis vastab ühepoolsele hüpoteesile. Seega näeme, et ühepoolse hüpoteesi tõestamine on üldiselt lihtsam (vajab keskmiselt veidi väiksemat valimit), võrreldes kahepoolse hüpoteesi tõestamisega. Selle põhjus on ka arusaadav: ühepoolse hüpoteesi korral me kasutame teatavat eelinformatsiooni, sest meid huvitab ainult ühesuunaline sõltuvus. Igasugune lisateave muudab üldiselt otsustamise hõlpsamaks.

Näide 62

On antud juhuslik valim Eesti valdadest, vt tabel 28. X on elanike arv vallas ja Y – elanike arv taludes. Meid huvitab teada, kas taluelanike arv on korreleeritud elanike üldarvuga vallas. On loomulik, et kui selline sõltuvus on, siis ta on positiivne, seetõttu me kontrollime ühepoolset hüpoteesi H_1: $r > 0$. Valime olulisuse nivooks 0,05. Arvutuste tulemusena saame korrelatsioonikordaja väärtuseks 0,14, mis on märksa väiksem kui tabelis leiduv kriitiline väärtus 0,378, mille leiame, kasutades vabadusastmete arvu 18 tabeli 39 teisest veerust.

Tabel 28

Vald	*Elanikke*	*Taluelanikke*
Lohusuu	*889*	*27*
Laeva	*901*	*81*
Lümanda	*1006*	*67*
Misso	*1054*	*58*
Orava	*1064*	*129*
Illuka	*1191*	*73*
Kõrgessaare	*1537*	*47*
Mäksa	*1705*	*128*
Urvaste	*1767*	*192*
Pühajärve	*2063*	*221*
Kohila	*2126*	*216*
Rakke	*2468*	*86*
Saarde	*2615*	*254*
Vändra	*2842*	*131*
Kanepi	*2947*	*268*
Taebla	*3013*	*79*
Ülenurme	*4523*	*76*
Tarvastu	*5076*	*249*
Kose	*5731*	*111*
Saue	*6562*	*51*



Joonis 20

Järelikult ei õnnestu meil tõestada, et Eesti valdades oleks taluelanike arv üldiselt korreleeritud elanike arvuga vallas. Sama kinnitab joonis 20: taluelanike arv on üsna ühesugune nii suurtes kui väikestes valdades. Tulemus on kooskõlas ka näitega 58, kus järeldasime, et suuremates maakondades mõjustab elanike arvu just linnaliste asulate arvukus. Järelikult pole meil mõtet seda mudelit üldse hakatagi koostama.

5.1.6. Lineaarne prognoos ja prognoosijääk

Kui lineaarne mudel on oluline, siis saame arvutada iga objekti jaoks tema *prognoosi*, kasutades lineaarset regressiooniseost

$$\tilde{y}_i = a + bx_i.$$

Kõik prognoosid asuvad regressioonisirge peal. Mudeli jääkliikme väärtust, st juhusliku suuruse tegeliku väärtuse ja prognoosi erinevust nimetatakse *prognoosijäägiks* ehk *prognoosiveaks*,

$$\varepsilon_i = y_i - \tilde{y}_i.$$

Näide 63

Vaatleme veel üht näidet Eesti valdade kohta. Olgu nüüd uuritavateks tunnusteks elanike arv ja töötajate elanike arv 20 vallas. Andmed on esitatud tabelis 29.

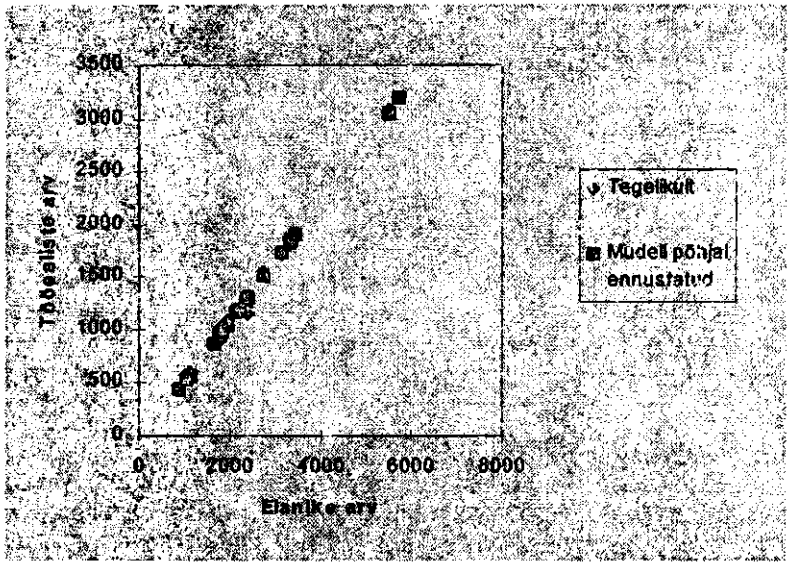
Leiame mudeli töötajate inimeste arvu kirjeldamiseks kõigi inimeste arvu järgi, kasutades selleks punktis 5.1.1 kirjeldatud meetodit. Saadud mudel on alljärgnev:

Töötajate arv = 0,57 × elanike arv - 69.

Tabel 29

Vald	Elanikke	Töötajaid	Töötajate prognoos	Prognoosiviga
Noarootsi	896	458	441	17
Surju	1043	519	524	-5
Hummuli	1122	598	569	29
Pihtla	1657	873	874	-1
Haaslava	1771	909	939	-30
Paistu	1784	1013	946	67
Veriora	1856	917	987	-70
Kaarma	1947	1087	1039	48
Järva-Jaani	2156	1165	1158	7
Puurmani	2214	1201	1199	10
Rakvere	2341	1312	1263	49
Rõuge	2387	1142	1289	-147
Halinga	2397	1308	1295	13
Aseri	2754	1545	1498	47
Anija	3169	1706	1734	-28
Antsla	3341	1805	1832	-27
Rõngu	3376	1860	1852	8
Nissi	3460	1880	1900	-20
Kadrina	5489	3094	3055	39
Jõgeva	5734	3191	3194	-3

Seejärel leiame korrelatsioonikordaja (vt punkt 5.1.2). Korrelatsioonikordaja väärtuseks saime 0,998. Korrelatsioonikordaja on statistiliselt oluline (vt 5.1.4) ja kinnitab väga tugeva lineaarse seose olemasolu. See on ka loomulik, sest tegemist on sisuliselt tihedasti seotud tunnustega. Jälle tekib küsimus: mida näitab negatiivne vabaliige? Nähtavasti on siin tegemist asjaoluga, et väiksemates valdades on töötajate arv suhteliselt väiksem kui suurtes, seetõttu polegi sõltuvus täpselt võrdeline (nii oleks olukord siis, kui vabaliige oleks 0). Tugevat sõltuvust näitab ka joonis 21, millel prognoosipunktid valdavalt kattuvad vaadeldud punktidega korrelatsooniväljal.



Joonis 21

Käesoleva näite puhul on huvitav uurida prognoosivea jaotust, sest sellel on oluline sisuline tähendus. Tabeli 29 viimasest veerust näeme, et absoluutväärtuselt suurim prognoosiviga on Rõuge ja samuti Veriora vallas. Nendes on tööealisi elanikke märksa vähem kui peaks olema elanike üldarvu järgi, tähendab nendest valdadest on tööealised ära rännanud. Seevastu aga Paistu, Rakvere, Kaarma ja Aseri valdades on tööealisi rohkem kui prognoositud. Nähtavasti on neist täisealiste väljaränne olnud väiksem või on toimunud sisseränne.

5.2. Mittearvulise argumendiga mudelid. Dispersioonanalüüs

Praktilistes ülesannetes võib mudeli argumendiks olla mitte ainult pidev arvtunnus, nagu nägime eelmises paragrahvis, vaid ka suvalisse tüüpi kuuluv diskreetne tunnus, nn rühmitav tunnus ehk *faktor*. Käesolevas paragrahvis vaatlemegi selliseid mudeleid.

5.2.1. Rühmakeskmiste hindamine rohkem kui kahe rühma puhul

Vaatleme statistilist kogumit, milles on mõõdetud kaks tunnust.

- Tunnus X on diskreetne tunnus ehk *faktor*, millel on k väärtust ehk *taset*. Neid tasemeid tähistatakse lihtsalt arvudega 1, ..., k.
- Tunnus Y on pidev arvtunnus.

Kui vaadelda nende tunnuste vastastikust mõju, siis üks sagedamini esinevaid skeeme on alljärgnev:

Statistiline kogum on vastavalt faktori X tasemetele jaotatud mitte-lõikuvateks osadeks ehk rühmadeks. *Kui tunnuse Y keskmised neis rühmades on erinevad, siis öeldakse, et faktor X mõjub tunnusele Y.*

Vaatame, kuidas on see ülesanne püstitatav ja lahendatav erinevate statistiliste kogumite korral.

- Kui tegemist on kõikse uuringuga, siis tuleb meil mõjud lihtsalt olemasoleva kogumi põhjal välja arvutada. Eeldame, et meil on vaatlusi kõigist rühmadest, $i = 1, ..., k$, kusjuures vaatlused i-ndast rühmast tähistame $y_{i_1}, y_{i_2}, ..., y_{i n_i}$. On võimalik, et igas rühmas on vaatluste arv erinev. Tähistame i-nda rühma vaatluste arvu tähega n_i. Vaatluste koguarv on sel juhul $n_1 + n_2 + ... + n_k = n$.

Et leida hinnangut tunnuse Y keskväärtusele i-ndas rühmas μ_i, leiame selle tunnuse aritmeetilise keskmise selles rühmas $\bar{y}_i$,

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \ , \ i = 1, 2, ..., k. \qquad (35)$$

Seejärel arvutame ka tunnuse Y üldkeskmise $\bar{y}$, kasutades selleks kõiki vaatlustulemusi,

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{y}_i . \qquad (36)$$

Loomulikult saab üldkeskmise leida ka rühmakeskmiste kaudu, nagu samast valemist näha. Et leida faktori X erinevate tasemete mõju tunnuse Y keskmisele, selleks arvutame vahed $\bar{y}_i - \bar{y}$, $i=1, \ldots, k$. Kui tegemist on kõikse uuringuga, siis ongi ülesanne sellega lahendatud ja faktori X *tasemete mõjud kindlaks tehtud.*

- Kui statistiline kogum on vaadeldav *esindava valimina üldkogumist*, siis on $\bar{y}$ tunnuse Y keskväärtuse hinnang üldkogumis ja $\bar{y}_i$ tunnuse Y keskväärtuse hinnang faktori X tasemele i vastavas üldkogumi osas. Vahed $\bar{y}_i - \bar{y}$ on vaadeldavad kui faktori X tasemete mõjude *hinnangud.* Need hinnangud on nihketa ja neid saab edukalt kasutatada sõltumata sellest, missugune on tunnuse Y jaotus üldkogumis.

Kui on tegemist valikuuringuga, siis võime sõnastada veel ühe küsimuse, mis igasuguse mudeli puhul on ülimalt tähtis: nimelt, kas uuritaval faktoril üldkogumis üldse ongi mõju, või on rühmakeskmiste vahelised erinevused tingitud juhuslikkusest. See on tegelikult *mudeli olulisuse* küsimus. Sellele annab vastuse *ühefaktoriline dispersioonanalüüs*, mida me vaatleme järgmises punktis.

5.2.2. Ühefaktorilise dispersioonanalüüsi ülesanne

Dispersioonanalüüs on metoodika, mis võimaldab uurida faktortunnuse X mõju arvtunnuse Y keskväärtusele, kasutades selleks valimi andmeid. Võtame kasutusele järgmised tähistused.

Olgu tunnuse Y teoreetiline keskväärtus i-ndas rühmas (ehk *rühmas* i) μ_i, $i = 1, \ldots, k$. Meid huvitab kontrollida järgmist hüpoteesipaari:

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$,

H_1: leiduvad rühmad i ja j nii, et $\mu_i \neq \mu_j$.

Viimane väide ei tähenda muidugi seda, et erineksid ainult mingi kahe rühma keskmised, see on nö minimaalne võimalus. Võib esineda ka olukord, kus *kõigi rühmade keskmised on erinevad.*

Sõnastatud hüpoteesipaari tõestamiseks ei sobi ükski varem defineeritud ja tabuleeritud jaotus, me peame selleks tuletama uue jaotuse, mida nimetatakse F-jaotuseks. F-jaotust kasutatakse põhiliselt mudelite olulisuse ja sobivusega seotud hüpoteeside kontrollimiseks. Järgmises punktis defineerimegi F- jaotuse.

5.2.3. F-jaotus

Defineerime juhusliku suuruse F alljärgnevalt.

- Olgu Z_1 juhuslik suurus, mille jaotuseks on χ^2-jaotus vabadusastmete arvuga f_1;
- olgu Z_2 juhuslik suurus, mille jaotuseks on χ^2-jaotus vabadusastmete arvuga f_2;
- olgu Z_1 ja Z_2 sõltumatud.
- Sellisel juhul on suhe

$$F = \frac{Z_1 / f_1}{Z_2 / f_2}$$

F-jaotusega vabadusastmete arvudega f_1 ja f_2, kusjuures esimesena märgitakse lugejas oleva ruutvormi vabadusastmete arv.

F-jaotusega juhusliku suurused moodustavad kahest täisarvulisest parameetrist – vabadusastmete arvudest f_1 ja f_2 – sõltuva jaotuste pere, mille α-täiendkvantiilid on tabuleeritud vastavalt olulisuse nivoo väärtustele 0,01 ja 0,05. Tavaliselt on F-jaotuse abil kontrollitavad sisulised hüpoteesid sõnastatavad ühepoolsete statistiliste hüpoteesidena, seetõttu piisabki ka F-jaotuse tabelites ainult täiendkvantiilide märkimisest (samuti nagu χ^2-tabeli korralgi). F-jaotuse tabel on vaadeldud tabelitest kõige mahukam, sest sõltub kahest vabadusastmete arvust. Selle raamatu lisas, tabelis 40, on toodud fragment F-jaotuse kriitiliste väärtuste tabelist olulisuse nivoo 0,05 korral.

5.2.4. Dispersioonanalüüsi ülesande lahendamine

Dispersioonanalüüsil on kaks matemaatilist eeldust, mille mittetäidetus paneb selle meetodi rakendatavuse olulise küsimärgi alla. Need eeldused on:

- tunnus Y on kõigis üldkogumi osades normaaljaotusega;
- tunnuse Y dispersioon on kõigis üldkogumi osades sama.

Lisaks sellele eeldame, nagu alati, et valimid on esindavad ja valimite punktid on sõltumatud (näiteks ei tohi valimis olla korduvaid mõõtmistulemusi samast punktist).

Esimeseks sammuks dispersioonanalüüsi teostamisel on üldkeskmise hinnangu $\bar{y}$ arvutamine valemist (36) ja rühmakeskmiste hinnangute $\bar{y}_i$ leidmine valemist (35), $i = 1, \ldots, k$.

Siis leiame koguhajuvuse, so hälvete ruutude summa üldkeskmisest:

$$SS = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2.$$

Selle summa lahutame kaheks liidetavaks,

$$SS = SS_1 + SS_0,$$

kus esimene liidetav iseloomustab rühmadevahelist, faktori mõjust tingitud hajuvust, ja teine liidetav rühmasisest ehk juhuslikku hajuvust,

$$SS_1 = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2, \quad SS_0 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2.$$

Nii koguhajuvus kui ka mõlemad liidetavad on nn *juhuslikud ruutvormid*, mida iseloomustab vabadusastmete arv. Siin arvutatakse vabadusastmete arv, lahutades vaatluste arvust (valimi mahust) vaatluste põhjal määratud konstantide arvu (üldisemalt: lineaarsete seoste arvu). Koguhajuvuse vabadusastmete arv on seega n–1. Esimeses liidetavas ruutvormis SS_1 on kasutatud vaatlustena k rühmakeskmist, ning neid ühendab üks lineaarne seos, mis määrab üldkeskmise. Järelikult on vabadusastmete arv k–1. Teises ruutvormis SS_0 on kokku n liidetavat, kuid neid seob k seost. Need seosed on rühmakeskmiste avaldised üksikpunktide kaudu. Järelikult on teise ruutvormi vabadusastmete arv $n - k$.

Dispersioonanalüüsis kasutatavate ruutvormide puhul peavad ruutvormide summeerimisel liituma ka vabadusastmete arvud; see seaduspärasus aitab kontrollida arvutuste õigsust. Nii on see ka praegu ruutvormide SS, SS_1 ja SS_0 puhul: n–1 = n–k + k–1. Kui dispersioonanalüüsi eeldused on täidetud ja nullhüpotees on õige, siis on suhe

$$F = \frac{(n-k)SS_1}{(k-1)SS_0}$$

F-jaotusega vabadusastmete arvudega k–1 ja n–k. Kui aga nullhüpotees ei kehti, siis on F-statistiku väärtus üldiselt liiga suur, sest rühmadevaheline hajuvus on tingitud faktori mõjust. See asjaolu ilmnebki arvutatud F-statistiku väärtuse võrdlemisel tabeliväärtusega. Seega saame püstitatud hüpoteesipaari kontrollimiseks alljärgneva eeskirja.

- Kui $F > F(\alpha, f_1, f_2)$, siis on sisukas hüpotees H_1 tõestatud olulisuse nivool α;
- kui $F \le F(\alpha, f_1, f_2)$, siis ei õnnestu sisukat hüpoteesi tõestada ning tuleb vastu võtta nullhüpotees H_0.

5.2.5. Dispersioonanalüüsi tabel

Dispersioonanalüüsiga seotud arvutused koondatakse tavaliselt alljärgnevasse nn dispersioonanalüüsi tabelisse, vt tabel 30.

Dispersioonanalüüsile on iseloomulik suhteliselt väike arvutusmahukus, mistõttu selle teostamine on täiesti võimalik ka käsitsi või taskuarvuti abil.

Tabel 30

Mõju	Ruutvorm	Vabadus-astmeid f	SS/f	F – statistik	Järeldus
Faktor	SS_1	$k-1$	$SS_1/(k-1)$	$(n-k)SS_1/((k-1)SS_0$	
Juhus	SS_0	$n-k$	$SS_0/(n-k)$		
Summa	SS	$n-1$			

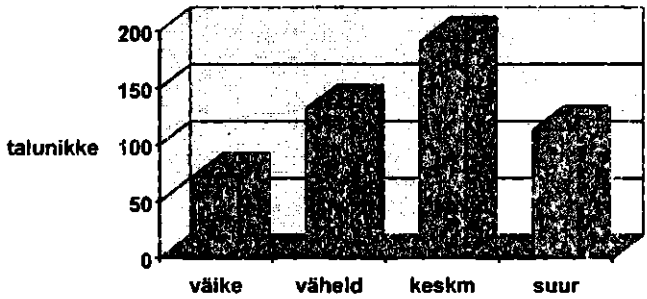
Dispersioonanalüüsi tabeli viimane rida võimaldab arvutusi kontrollida: viimases reas peavad olema eelnevate ridade vastavate veergude summad.

Näide 64

Pöördume tagasi näite juurde, kus vaatlesime taluelanike arvu sõltuvust vallaelanike arvust ning kus lineaarne mudel praktiliselt ei töötanud. Jaotame vallad elanike arvu järgi neljaks rühmaks: väiksed vallad (elanikke alla 1100), väheldased vallad (elanikke 1100 – 2100), keskmisest veidi suuremad vallad (elanikke 2100 –3000) ja suured vallad (elanikke üle 3000). Loeme selliselt määratletud valla suuruse faktoriks ja võtame Y-tunnuseks talunike arvu.

Üksikutes vallatüüpides saime järgmised keskmised talunike arvud: 72,4; 132,2; 191; 113,2; vt ka alljärgnevat joonist. On selge, et lineaarset seost valla suuruse ja talunike arvu vahel pole, suurtes valdades on keskmiselt vähem talunikke kui keskmistes, nähtavasti on neis teisi võimalikke elatusallikaid rohkem.

Teeme nüüdi dispersioonanalüüsi kontrollimaks nullhüpoteesi: kõigis vallatüüpides on keskmiselt sama arv talunikke.



Joonis 22

Tulemuste dispersioonanalüüsi tabel on alljärgnev:

Tabel 31

Mõju	*SS*	*Vabadus-astmeid*	*Suhe*	*F*	*F tabeli-väärtus*
Faktor	*36472,4*	*3*	*12157,47*	*2,499132*	*3,238867*
Juhus	*77834,8*	*16*	*4864,675*		
Summa	*114307,2*	*19*			

Näeme, et ka dispersioonanalüüsi tulemusena peame võtma vastu nullhüpoteesi: arvutatud F väärtus on väiksem vastavast kriitilisest väärtusest. Kuigi jooniselt paistab, et erinevused eri tüüpi valdade keskmiste talunike arvude vahel on üsna suured, on siiski selle näitaja rühmasisene hajuvus nii suur, et sisukat hüpoteesi pole võimalik tõestada. Tõenäosliselt aga saaks see võimalikuks valimi mahu suurendamisel.

6. Aegridade analüüsi põhimõisted ja -ülesanded

6.1. Aegrea mõiste. Aegrea komponendid

6.1.1. Aegrida

Tihti pakuvad uurimustes huvi ajas kulgevate protsesside mõõtmistulemused. Niisuguste mõõtmistulemuste käsitlemist järelduste tegemiseks nimetatakse *aegridade analüüsimiseks*, sellega tegelev matemaatika haru on *aegridade teooria*.

Olgu mingit juhuslikku suurust (arvtunnust) X mõõdetud n ajamomendil, kusjuures ajamomentide vahed on võrdsed. Tähistame ajahetked järjestikuste naturaalarvudega 1, ..., n. Siis mõõtmistulemused x_1, x_2, ..., x_n moodustavad *aegrea*. Aegrida tähistatakse ka tähega X_t. Väärtuse x_i kohta öeldakse ka, et see on aegrea X_t väärtus ajahetkel i.

Aegrea käsitlemiseks sobivad põhimõtteliselt kahe tunnuse analüüsi meetodid, kusjuures argumendiks t loetakse *aega* (mille väärtusteks on 1, ..., n) ja mõõdetavaks tunnuseks tunnust X.

Aegrea uurimisel tekib aga rida uusi momente, mis on seotud selle eripäraga. Juhusliku suuruse X muutumine ajas võib olla kirjeldatav mitme komponendi abil. Neist olulisemad on:

- sesoonne komponent v_i, mis muutub perioodiliselt ette teada oleva perioodiga, mille pikkuseks on tavaliselt aasta;
- muud perioodilised komponendid, mille tähiseks on p_i;
- süstemaatiline komponent ehk trend t_i, mis esindab aegrea determineeritud muutust sõltuvalt ajast;
- juhuslik komponent e_i.

Seega on üks võimalikke aegrea esitamise vorme komponentide kaudu alljärgnev nn *aditiivne esitus*, mille tunnuseks on see, et aegrida avaldub oma komponentide summana:

$$x_i = v_i + t_i + e_i, \qquad (37)$$

kus i=1, ..., n. Selles aegreas puudub mittesesoonne perioodiline komponent.

6.1.2. Aegrea analüüs ja prognoosimine

Aegridade teoorias eeldatakse harilikult, et aegrida on teatava *teoreetilise juhusliku protsessi X realisatsioon*. See teoreetiline juhuslik protsess iseloomustab uuritavat nähtust üldkogumis. Sellise eelduse korral moodustavad vaatluste tulemusena saadud aegrea väärtused x_t koos vastavate ajahetkedega t_i *valimi* selle juhusliku protsessi (öeldakse ka: teoreetilise aegrea) väärtuste hulgast. Uurija eesmärgiks on:

- leida mudel, mis kirjeldaks aegrea muutumist ajas, kirjeldada selle üksikkomponente, eriti trendi;
- siluda aegrida, st kõrvaldada aegrea väärtustest juhuslik komponent e_t;
- leida eeskiri f aegrea järgnevate liikmete prognoosimiseks eelnenud (teadaolevate) liikmete põhjal:

$$x_{n+s} = f(x_1, \ldots, x_n),\ s = 1, \ldots, m.$$

Kõik nimetatud ülesanded on statistilised, sest aegrida käsitletakse üldiselt kui valimit. Seetõttu tekib ka küsimus leitud mudelite *olulisusest* ja tehtud prognooside *usaldusväärsusest*: kui mudel ei ole oluline, siis pole ka tema alusel tehtud prognoos usaldusväärne.

Eriti problemaatiline on prognooside tegemine pikkade ajavahemike peale ette, st olukord, kui m on suhteliselt suur.

6.1.3. Aegreast tuletatud aegread

Sageli saab olemasoleva aegrea põhjal tuletada uusi, samuti huvipakkuvaid aegridu. Tuntuimad näited on alljärgnevad:

- Rea X_t *juurdekasvude* Δ_t aegrida, kus
$$\Delta_t = x_t - x_{t-1}.$$
- Defineerides rea sesoonse juurdekasvu valemiga
$$\Delta_t^s = x_t - x_{t-s},$$
saame rea X_t *sesoonsete juurdekasvude* Δ_t^s aegrea. Siin s tähistab sesooni pikkust.
- Arvutades rea X_t liikmete summad
$$S_t = x_1 + x_2 + \ldots + x_t,$$
saame rea *summade* aegrea S_t.
- Aegrea X_t keskmist
$$\bar{x} = 1/n\,(x_1 + x_2 + \ldots + x_n)$$
nimetatakse *keskmiseks ajas* (keskmiseks tendentsiks). Seda võib käsitleda ka *konstantse aegreana*, mille kõik liikmed võrduvad keskmisega $\bar{x}$.
- Keskmistatud aegrida, mis on saadud aegreast X_t tema keskmise lahutamisel.

6.1.4. Trendi hindamine ja prognoosimine

Trendi hindamiseks ja prognoosimiseks on põhimõtteliselt kasutatavad kõik kahe tunnuse mudelid, kus argumendiks on aeg (vaadeldavad ajamomendid $t_1, \ldots, t_n$) ja hinnatavaks funktsiooniks on trend. Trendiks võib olla näiteks tõusev sirge (siis on tegemist *lineaarse trendiga*), paraboolne, logaritmiline või eksponentsiaalne kõver, neil juhtudel kõneldakse vastavalt *paraboolsest, logaritmilisest* või *eksponentsiaalsest trendist*. Eksponentsiaalne trend sobib kirjeldama aegrea *järjest kiirenevat kasvu*.

Kui perioodilised kõikumised on trendiga võrreldes suhteliselt väikesed (või puuduvad hoopis) ja vaatlused hõlmavad täisarvu perioode, saab trendi t_i, $i=1, \ldots, n$, arvutada vahetult aegreast. Sõltuvalt tehtavaist eeldustest on selleks olemas järgmised võimalused.

- Eeldatakse, et perioodilistel komponentidel (sh sesoonsel komponendil) puudub süstemaatiline mõju ja juhuslikud *hälbed* e vead (juhusliku komponendi väärtused) e_i on sõltumatud. Neil eeldustel võetakse trend võrdseks mõõdetud aegrea väärtustega: $t_i \equiv x_i$; $i=1, \ldots, n$.
- Eeldatakse, et hälbed on sõltumatud, kuid perioodiline komponent on olemas (ei ole konstantselt võrdne nulliga). Sel juhul kõrvaldatakse perioodiline komponent (näiteks punktis 6.1.5 kirjeldatava meetodiga); $t_i \equiv x_i - v_i$; $i=1, \ldots, n$ ja edasi toimitakse samuti kui eelnevalgi juhul.
- Kõrvaldatakse nii perioodiline kui ka juhuslik komponent mingi silumismeetodi abil (vt paragrahv 6.2) ja saadud silutud aegrea väärtused loetakse trendiks.

Kui trendi hindamiseks ja kirjeldamiseks on leitud hästi sobiv mudel (funktsioon), siis on enamasti põhimõtteliselt võimalik sama funktsiooni ekstrapoleerides seda kasutada ka aegrea tulevase käitumise ennustamiseks ehk *prognoosimiseks*. Sageli on aga niisuguse "hea" mudeli, st trendi pikema ajavahemiku jooksul hästi kirjeldava analüütilise funktsiooni leidmine raske. Kui mudelit soovitakse kasutatada ekstrapoleerimiseks, tuleb eriti hoolega jälgida trendi muutusi vaatlusperioodi lõpuosas ($i= 1, \ldots, n-1, n$).

6.1.5. Perioodilise komponendi hindamine

Kui aegrea trendi on õnnestunud hinnata, siis võime selle esialgsest reast lahutada ning edaspidi oletada, et tegemist on null-trendiga aegreaga

$$y_i' = v_i + e_i.$$

Sellisest reast on suhteliselt lihtne leida aegrea *perioodikeskmisi* tasemeid.

Kui perioodi pikkus on p ja perioode on q, $n=p\cdot q$, siis saame *prioodikeskmised* tasemed m_j leida lihtsalt

$$m_j = \frac{1}{q}\sum_{h=0}^{q-1} x_{j+hp},$$

kus j omandab väärtused 1, ..., p.

Loomulikult on võimalik ka *perioodikeskmisi* m_j omakorda teatavate funktsioonidega lähendada. Kõige paremini sobib selleks sinusoid. Üks võimalusi sinusoidiga lähendamiseks on see, et leitakse näiteks vähimruutude meetodil kordajad a ja b ning algnurk c nii, et kehtiks seos

$$m_j = a\sin(c+j\pi/p) + b\cos(c+j\pi/p), \quad j=1,\ldots,p.$$

Põhimõtteliselt on võimalik kasutada perioodilise liikme m_j lähendamiseks ka rohkem liikmeid. Sellega võib saada parema lähendi, kuid mudelisse tuleb lülitada rohkem valimi põhjal hinnatavaid konstante, mis omakorda suurendab juhuse mõju tulemusele.

Näide 65

Vaatleme lihtsat näidet selle kohta, kuidas leida aegrea perioodilist komponenti.

Võtame aluseks igakuise sündide arvu Eestis 1995. aastal. Andmed on esitatud alljärgneva tabeli kahes esimeses veerus. Ilmselt on tegemist ühe perioodiga. Trendi loeme konstantseks, seega $t_i = 1126$, $i = 1, \ldots, 12$.

Tabel 32

Kuu	Sünde	Erinevus keskmisest	$a_i = (i/n)360$	$\cos a_i$	$\sin a_i$
Jaanuar	1116	-10	0	1	0
Veebruar	1064	-62	30	0,886	0,5
Märts	1258	132	60	0,5	0,886
Aprill	1174	48	90	0	1
Mai	1192	66	120	– 0,5	0,886
Juuni	1182	56	150	– 0,886	0,5
Juuli	1206	80	180	– 1	0
August	1152	26	210	– 0,886	– 0,5
September	1122	-4	240	–0,5	– 0,886
Oktoober	994	-132	270	0	– 1
November	1061	-65	300	0,5	– 0,886
Detsember	990	-136	330	0,886	– 0,5

Püüame kolmanda veeru väärtusi, mis esindavad perioodilise ja juhusliku komponendi summat $v_i + e_i$, lähendada võimalikult lihtsa perioodilise mudeliga $b\cos(c+i/n\times 360°)$, kus nurk c on tundmatu, kuid nurk $i/n \times 360°$

omandab väärtused 0°, 30° jne. Vastavad nurga väärtused ja funktsioonid ongi kirjutatud tabeli 32 neljandasse, viiendasse ja kuuendasse veergu.

Vaadeldavas mudelis on kaks tundmatut parameetrit b ja c, mille hindamiseks vähimruutude meetodil saame kasutada aegrea kahtteist punkti. Kõigepealt teisendame mudelis olevat trigonomeetrilist avaldist, kasutades selleks valemit

$$\cos(\alpha + \beta) = \cos\alpha \cos\beta - \sin\alpha \sin\beta;$$

Lähendamiseks kasutame nüüd teisendatud valemit

$$v_i = b \cos(c + a_i) = b(\cos a_i \cos c - \sin a_i \sin c),$$

milles on kaks tundmatut parameetrit – kordaja b ja nurk c. Nende määramiseks kasutame samaselt lineaarsele regressioonanalüüsile vähimruutude meetodit ja leiame:

- $b \cos c = -55,0;$
- $b \sin c = -77,2.$

Arvestades trigonomeetriliste funktsioonide omadusi saame:

- $b^2 = 55^2 + 77,2^2 = 8984,84$ *ja* $b = 94,79.$
- *Nurk c paikneb ilmselt kolmandas veerandis, ning arvestades, et temale vastav siinus on –0,8144 ja koosinus on – 0,58, järeldub siit, et nurga c suurus on* $234^\circ,5.$

Tabel 33

Kuu	Sünde	$a_i = (i/n)360$	$c+a_i$	$\cos(c+a_i)$	$b(c+a_i)$	e_i
Jaanuar	1116	0	234,5	-0,581	-55,04	45,04
Veebruar	1064	30	264,5	-0,096	-9,08	52,91
Märts	1258	60	294,5	0,415	39,31	92,69
Aprill	1174	90	324,5	0,814	77,17	-29,17
Mai	1192	120	354,5	0,995	94,35	28,35
Juuni	1182	150	24,5	0,910	86,26	30,26
Juuli	1206	180	54,5	0,581	55,04	24,96
August	1152	210	84,5	0,096	9,08	16,92
September	1122	240	114,5	-0,414	-39,31	35,31
Oktoober	994	270	144,5	-0,814	-77,15	-54,84
November	1061	300	174,5	-0,995	-94,35	-29,35
Detsember	990	330	204,5	-0,910	-86,26	-49,74

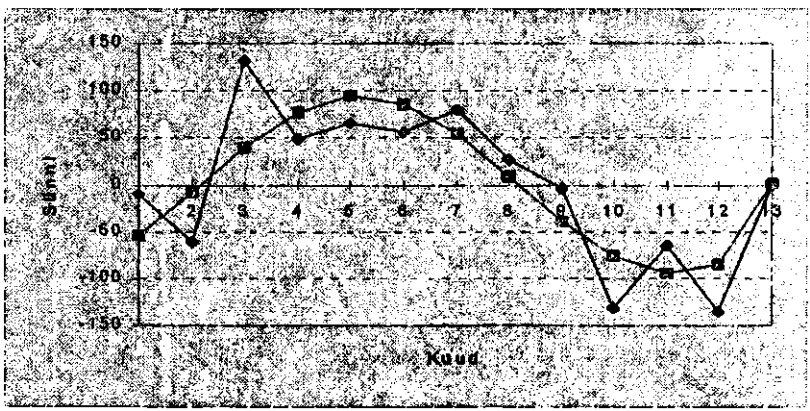
Lisaks arvutame järgmised suurused (vt. tabel 33):

- *nurgad* $c+a_i$,
- *selle nurga koosinused,*
- *prognoositud väärtused* $b \cos(c+a_i)$,

- *prognoosijäägid, e_i, mis käesoleval juhul on arvutatud valemist $e_i = x_i - \bar{x} - v_i$.*

Illustreerimaks esialgset aegrida ja saadud lähendit, esitame veel joonise 23.

Saadud lähendi headust iseloomustab ka esialgse hälvete ruutude summa $SS = 7127$ ja prognoosijääkide ruutude summa $SS_0 = 2226$ võrdlus (vt punkt 5.2.4).



Joonis 23

Näide 66

Vaatleme sündide arvu Eestis aastail 1945–1995 (vt tabel 34). Sünnid moodustavad aegrea, mida kujutab graafik joonisel 24.

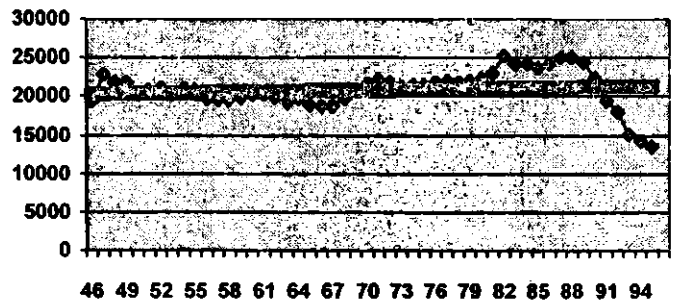
Püüame lähendada selle aegrea trendi lineaarse regressiooni teel. Saadud mudel ei ole kuigi hea, sest korrelatsioonikordaja väärtus on kõigest 0,120. Saadud lineaarne mudel on alljärgnev:

$$\text{sündide arv} = 19317 + 20{,}9 \times \text{aastaarv},$$

siin vaadeldakse aastaarvuna üksnes kümnelisi ja ühelisi.

Tabel 34

Aasta	Sünde	Aasta	Sünde	Aasta	Sünde
45	14968	62	19959	79	21879
46	19408	63	19275	80	22204
47	22721	64	19629	81	22937
48	21777	65	18909	82	25128
49	21770	66	18629	83	24155
50	20279	67	18671	84	24234
51	20730	68	19782	85	23630
52	21111	69	20781	86	24105
53	20146	70	21552	87	25086
54	20909	71	22118	88	25060
55	20786	72	21757	89	24292
56	19660	73	21239	90	22308
57	19509	74	21461	91	19320
58	19598	75	21360	92	18006
59	19938	76	21801	93	15170
60	20187	77	21977	94	14178
61	20230	78	21842	95	13560



Joonis 24

6.2. *Aegridade silumine*

Aegridade käsitlemisel on, sõltuvalt eesmärgist, võimalik kõrvaldada niihästi süstemaatiline komponent – sesoonsus või/ja trend – kui ka juhuslik komponent. Juhusliku komponendi kõrvaldamise meetodeid nimetatakse üldiselt *silumismeetoditeks*.

6.2.1. Libiseva keskmise meetod aegridade silumiseks. Lineaarne silumine

Libiseva keskmise meetodi idee seisneb selles, et aegrea üksikud lõigud lähendatakse polünoomidega, ning aegrea uueks väärtuseks võetakse leitud polünoomi väärtus lähendataval ajahetkel.

Lihtsaim libiseva keskmise silumismeetod on lineaarne silumine. Kirjeldame alljärgnevas lineaarset silumist kolme punkti abil.

Olgu mõõdetud n punkti, x_1, ..., x_n. Iga punkti asemele vahemikust 2, ..., n–1 arvutatakse uus punkt x_i' eeskirjaga $x_i' = (x_{i-1} + x_i + x_{i+1})/3$. Samal viisil võib iga punkti asendada ka viie, seitsme ja rohkemagi punkti aritmeetilise keskmisega. Sellise keskmistamise tulemusena saadakse üldiselt märksa siledama (laugema) kujuga kõver, kuid samal ajal tähendab see, et ka osa sisukat informatsiooni võib kaotsi minna. Suurema arvu punktide kasutamine silumiseks tähendab ka suurema arvu rea otstes paiknevate punktide kaotsiminekut silutud reast.

6.2.2. Polünoomiga lähendamine libiseva keskmisega silumisel

Polünomiaalse silumise korral leitakse aegrea r järjestikust punkti lähendav s-astme polünoom, kusjuures loomulikult peab punktide arv olema suurem kui polünoomi järk. Lähend arvutatakse tavaliselt vähimruutude meetodil.

Kasutades sobivaid eeldusi polünoomi orientatsiooni kohta, on võimalik saavutada seda, et polünomiaalne lähend realiseerub lähtepunktide lineaarkombinatsioonina. Iga erinev punktide arv ja polünoomi järgu kombinatsioon määrab esialgse rea liikmete jaoks üldiselt erinevad kaalud, mille järgi uus punkt arvutatakse,

$$x_i' = \sum_{j=i-u}^{i+u} w_j x_j .$$

Kaalude w_j summa võrdub alati ühega ja liikmete arv r on $2u+1$. Enamiku libiseva keskmise tüüpi valemite puhul on tüüpiline see, et hinnatavale punktile lähemad esialgse rea punktid on suuremate, kaugemad aga üldiselt väiksemate kaaludega. Libiseva keskmise avaldises on kaalud sümmeetrilised, st x_{i-j} ja x_{i+j} on samade kordajatega. Tuntud on näiteks Spenceri 15- ja 21-punktilised valemid. Esimene nendest kirjutatakse skemaatiliselt üles alljärgnevalt

$$1/320\ \{-3, -6, -5, 3, 21, 46, 67, 74\}.$$

Siin on kaaludeks vastavalt -3/320, -6/320, jne, kusjuures poolpaksult trükitud liige vastab samale indeksile, millele vastavat väärtust lähendatakse.

Silumise eesmärgiks võib olla juhuslike vigade eemaldamise kõrval ka perioodilise komponendi eemaldamine. See on võimalik üksnes siis, kui silumisvahemik $2u+1$ on vähemalt sama pikk kui periood.

Näide 67

Vaatleme kõigepealt aegrea silumist, kasutades näites 65 kasutusele võetud sündide andmestikku. Seame enesele eesmärgiks siluda seda aegrida viiest punktist läbi pandud parabooli abil. Selle metoodika rakendamisel läheb kaotsi kaks punkti aegrea kummastki otsast ja silumine tuleb läbi teha kaheksa korda. Vaatame kõigepealt viit esimest aegrea punkti, vt joonis 25. Nende viie punkti abil arvutame uue väärtuse kolmandale punktile. Esialgses andmestikus iseloomustab kolmandat punkti (märtsikuud) järsk hüpe ülespoole, mille suhtes eeldame, et on tegemist juhusliku hälbega.

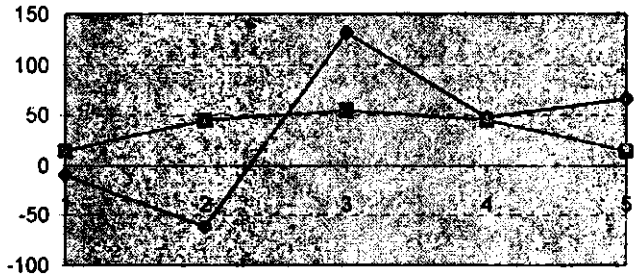
Silumise käigus otsime esialgse viie punkti abil parima ruutparabooli, mis lähendab neid viit punkti. Lihtsuse mõttes oletame, et tegemist on paraboliga, mille telg on vertikaalne, siis on ta kirjeldatud kahe parameetriga a ja b:

$$y = a + b(x - \bar{x})^2$$

Parameetrite väärtused on vähimruutude meetodil lihtsalt hinnatavad, käesoleval juhul saame:

- $a = 54,5$
- $b = 9,86$
- *ja otsitava punkti uueks väärtuseks on 54,5, vt ka joonist.*

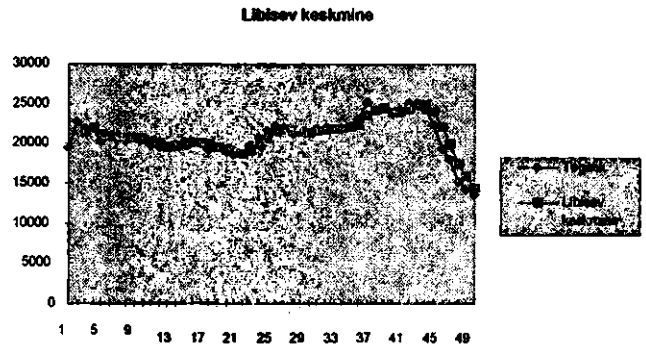
Järgmisel sammul võtame viieks punktiks punktid 2–6 aegreast jne.



Joonis 25

Näide 68

Libiseva keskmise meetodil silutud sündimusandmed esitatakse joonisel 27.



Joonis 26

6.2.3. Eksponentsilumine

Libiseva keskmise meetod sobib küll olemasoleva aegrea analüüsimiseks, kuid ei sobi hästi prognoosimiseks, sest kasutab iga ajamomendi jaoks nii sellele eelnevaid kui ka järgnevaid vaatlusi. Selletõttu on libiseva keskmise meetodit edasi arendatud alljärgneva mõttekäigu kaudu. Kuigi libiseva keskmise meetodi puhul kasutatakse realiikmete uute väärtuste arvutamisel ainult esialgse rea väärtusi x_i, on võimalik lihtsa teisenduse abil rea uute väärtuste x_i' avaldamine ka teisendatud rea eelmiste liikmete kaudu, kasutades alljärgnevat avaldist:

$$x_i' = x_{i-1}' + (x_i - x_{i-m})/m,$$

kus x_i' tähistab silumise tulemusena saadavat aegrea liiget. Seega saab teisendatud aegrea järgmiste liikmete arvutamiseks kasutada uue rea juba arvutatud liikmeid. See mõte ongi aluseks eksponentsilumise (ka eksponentsiaalse silumise) meetoditele. Lihtsaima eksponentsilumise meetodi üldvalem on alljärgnev

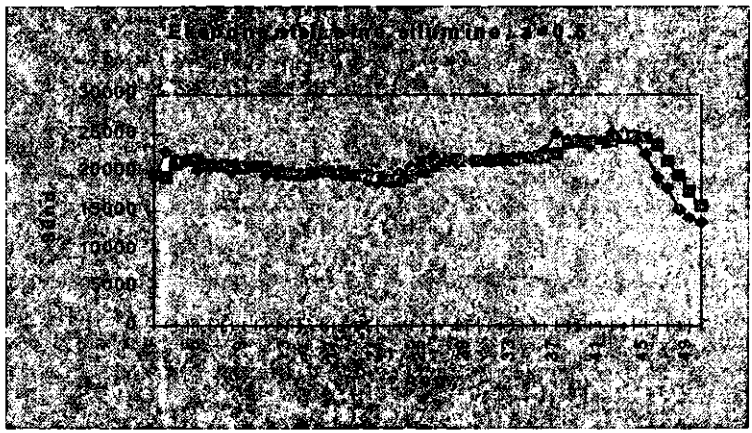
$$x_i' = \alpha x_i + (1 - \alpha)\, x'_{i-1}$$

Mida suurem on parameetri α väärtus, seda rohkem jälgib aegrida esialgset aegrida. Väikese parameetri väärtuse korral aga mõjutab tegelik rea väärtus antud hetkel silumisel saadavat väärtust vähe, ning niisugusel juhul on oht, et silutud rida läheb nö "metsa", jääb tegelikest väärtustest kaugele, ning selle kasutamine prognoosimiseks võib anda väga vildakaid tulemusi.

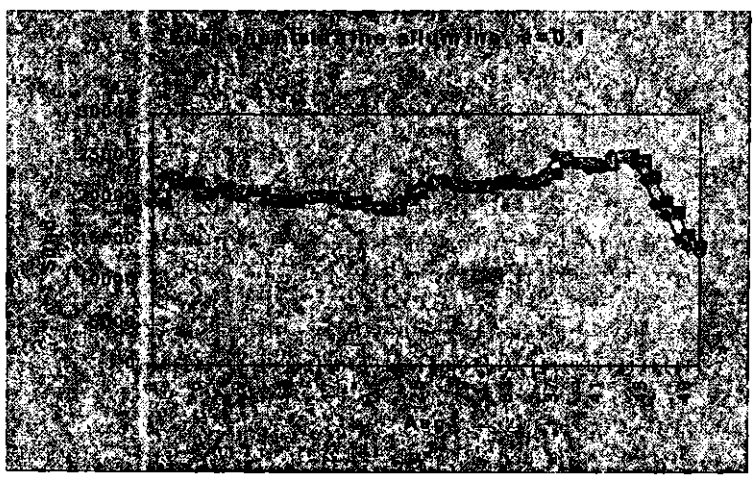
Märgime, et vahel kasutatakse tähist α ka liikme x_{i-1}' kordaja märkimiseks. Sellisel juhul on järeldused vastupidised eelmises lõigus toodutega – mida suurem on α väärtus, seda enam silutakse esialgset aegrida. Joonistel 27-28 ongi kasutatud viimatimainitud tähistust. Eksponentsilumine sobib statsionaarsete aegridade korral (vt punkt 6.3.1). Mittestatsionaarsete aegridade korral kasutatakse selle meetodi edasiarendust, nn Holt-Wintersi meetodit.

Näide 68

Kasutame eelmises näites vaadeldud aegrea silumiseks eksponentsilumist, kusjuures varieerime parameetrit α:



Joonis 27



Joonis 28

Erinevate α väärtuste korral saadud tulemused on esitatud joonistel 27 ja 28.

6.3. Autokorrelatsioonifunktsioon. Aegrea juhusliku komponendi prognoosimine

6.3.1. Statsionaarne aegrida

Kui aegreast

$$x_i = t_i + v_i + \varepsilon_i$$

on süstemaatilised (juhusest mitte sõltuvad) komponendid trend t_i ja sesoonne komponent v_i eraldatud, jääb järele aegrea juhuslik komponent, mille kohta me edaspidi ütleme, et see on *puhtjuhuslik* aegrida ε_i. Ka selle osa puhul on võimalik püstitada prognoosimisülesanne – tänu sellele, et aegrea üksikväärtused, sh. ka ε_i väärtused erinevatel ajahetkedel on üldjuhul sõltuvad juhuslikud suurused. Selle poolest erinebki aegrida teistest statistika valdkondades käsitletavatest valimitest.

Fikseerime edaspidiseks eeldused:

- Olgu $t_i = v_i = 0$, st. süstemaatiline osa on aegreast kõrvaldatud;
- olgu $EX_i = E\varepsilon_i = 0$ ja $DX_i = D\varepsilon_i = \sigma^2$ konstantsed.
- Eeldame veel, et aegrea liikmete X_i ja X_{i+j} omavaheline korrelatsioon on ühesugune sõltumata i ja j väärtustest, $i=1, ..., n-1, j=1, ..., n-i$.

Niisuguseid tingimusi rahuldavat aegrida nimetatakse *statsionaarseks*.

Statsionaarse aegrea puhul iseloomustab aegrea liikmete omavahelisi sõltuvusi autokorrelatsioonifunktsioon $r(i)$. *Valimi autokorrelatsioonifunktsiooni* väärtus juhul $i=1$ arvutatakse alljärgnevast valemist:

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

kus $\bar{x}$ on aegrea üldkeskmine (*keskmine ajas*) ja s^2 on dispersiooni hinnang,

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

Autokorrelatsioonikordaja $r(1)$ suurus iseloomustab aegrea kahe järjestikuse liikme omavahelise korrelatiivse seose tugevust.

Samal viisil võime arvutada ka aegrea liikme korrelatsioonikordaja temast g sammu kaugusel oleva liikmega, g=2, 3, ..., n–1.

$$r(g) = \frac{1}{n-g-1}\sum_{i=1}^{n-g}(x_i - \bar{x})(x_{i+g} - \bar{x})/s^2, \qquad (38)$$

Seega on autokorrelatsioonifunktsioon ehk korrelatsioonifunktsioon $r(i)$ naturaalarvude hulgal määratud funktsioon. Kui aegrida x_i on puhtjuhuslik selles mõttes, et tema süstemaatiline osa on null (kõrvaldatud), siis on funktsioon $r(g)$ üldiselt kahanev ja tema väärtused lähenevad g suurenedes nullile.

Kui aga autokorrelatsioonifunktsioon alguses väheneb, siis jälle hakkab suurenema, viitab see kõrvaldamata jäänud perioodilisusele.

Autokorrelatsioonifunktsiooni graafikut nimetatakse ka *korrelogrammiks*.

6.3.2. Autokorrelatiivne rida

Kui oletada, et tegemist on statsionaarse aegreaga, mille puhul autokorrelatsioonifunktsiooni väärtuseks kohal 1 on α, kus α on positiivne suurus 0 ja 1 vahel, siis saame kirjutada seose

$$x_i = \alpha x_{i-1} + u_i,$$

kus u_i on suurusest x_{i-1} sõltumatu juhuslik suurus. Sama seost saame rakendada ka liikme x_{i-1} jaoks, ja kokkuvõttes saame

$$x_i = \alpha^2 x_{i-2} + u_i + \alpha u_{i-1}$$

jne.

Sellist rida nimetatakse autokorrelatiivseks reaks, ning sellise rea puhul on võimalik prognoosida ka rea juhuslikku komponenti, kuid ainult piiratud ulatuses.

Viimasele mõõtmistulemusele järgneva komponendi x_{n+1} prognoosi headust mõõdab korrelatsioonikordaja α, järgmise prognoosi headust α^2 jne. Seega enamasti peale mõnda sammu prognoositavuse hinnang läheneb 0-le.

Näide 69

Vaatleme autokorrelatsioonifunktsiooni arvutamist näite 65 andmetel, kusjuures kasutame lihtsuse mõttes hälbeid keskmisest. Oletame, et meil ei ole alust arvata, et sellel aegreal oleks olemas perioodiline komponent ja loeme aegrida statsionaarseks. Selle aegrea üldkeskmine on 0 ja dispersioon (vt näidet 65) 7127. Arvutuste vahetulemused on toodud järgmises tabelis 35.

Tabel 35

aegrida (hälbed keskmisest)	*korrutised* $x_i\ x_{i+1}$	*korrutised* $x_i\ x_{i+2}$	*korrutised* $x_i\ x_{i+3}$
-10	*620*	*-1320*	*-480*
-62	*-8184*	*-2976*	*-4092*
132	*6336*	*8712*	*7392*
48	*3168*	*2688*	*3840*
66	*3696*	*5280*	*1716*
56	*4480*	*1456*	*-224*
80	*2080*	*-320*	*-10560*
26	*-104*	*-3432*	*-1690*
-4	*528*	*260*	*-544*
-132	*8580*	*17952*	*-4642*
-65	*8840*	*28300*	
-136	*30040*		

Selleks, et leida korrelatsioonifunktsiooni väärtust r(1), tuleb meil iga aegrea väärtus korrutada järgmise väärtusega ja kõik saadud 11 kahekordset korrutist summeerida. Seda on tehtud tabeli teises veerus, kus viimane lahter on liidetavate summa. Kasutades valemit (38) ja teadaolevat dispersiooni hinnangut, saame

- *r(1) = 30040/(11×7127) = 0,38.*

Samal viisil leiame järgmistest tabeli veergudest korrelatsioonifunktsiooni järgmised väärtused:

- *r(2) = 28300/(10×7127) = 0,40,*
- *r(3) = – 4642 / (9×7127) = – 0,07.*

Esimesel pilgul tundub, et saadud tulemust on üsna keeruline tõlgendada. Proovime siiski.

- *Seos aegrea järjestikust liikmete vahel on küllaltki nõrk, vaid 14 - 15%.*
- *Iga liikme seos ülejärgmise liikmega on sama suurusjärku kui seos järgmise liikmega (punkthinnangu tasemel isegi pisut suurem). Sellist seose kuju näeme ka jooniselt, kus aegreal on mitu üles-alla liikuvat sakki. Osalt võib selle põhjuseks olla kuude erinev pikkus.*
- *Korrelatsioonifunktsioon muutub i väärtusel 3 nii lähedaseks nullile, et võib kinnitada: aegrea liikmetel puudub nii kaugele ulatuv vastastikune mõju.*

Näeme, et eelmises näites vaadeldud rida ei ole hästi kooskõlas autokorrelatiivse rea mudeliga, sest sellise rea puhul peaks korrelatsioonifunktsioon järjest vähenema.

6.3.3. ARMA mudelid ja Box-Jenkinsi meetod nende lahendamiseks

Ka aegridade mudelite loomisel võime kasutada matemaatilisele statistikale omast lähenemist: kujutleme, et olemasolev konkreetne aegrida on valim teoreetilise aegrea kõikvõimalike realisatsioonide hulgast. Selle realisatsiooni, valimi, põhjal võime teha järeldusi meile tundmatu teoreetilise aegrea kohta. Niisuguste järelduste hulgast on olulisim aegrea edasise käigu prognoosimine. See saab võimalikuks just siis, kui õnnestub leida mudel, mis on niivõrd hästi vaadeldava aegreaga kooskõlas, et on alust lugeda seda aegrida nimetatud mudeli realisatsiooniks.

Üks lihtsamaid võimalikke mudeleid on autokorrelatiivne aegrida, tema kõrval pakub huvi ka lihtne libiseva keskmise rida, mille puhul eeldame, et iga aegrea liige on tugevasti korreleeritud ainult oma naaberliikmega, kuid mitte kaugemate elementidega, st. et kehtib seos

$$r(1) = \begin{cases} B > 0 & , \ kui \ \ i = 1 \\ 0 & \ \ muidu. \end{cases}$$

Seda, kas konkreetne andmestik kummagagi neist mudelitest kooskõlas on, saab lihtsalt kontrollida, kasutades selleks korrelatsioonifunktsiooni väärtust (milleks saab kasutada korrelatsioonikordaja olulisuse kontrollimise teste).

Mõlemaid mudeleid ühendav ja üldistav mudel on tuntud ARMA-mudelina. Siin kasutatakse ingliskeelseid lühendeid: AR – autoregressiivne, MA – libisev keskmine. Eesti keeles kasutatakse nende kohta ARLIK- või ARLK-mudelite nimetust.

Üldjuhul kõneldakse ARMA(p,q) mudelitest, kus p on autoregressiooni ja q libiseva keskmise mudeli järk.

ARMA-mudelite sobitamiseks aegridadele ja nende mudelite parameetrite hindamiseks on metoodika välja töötanud Box ja Jenkins. Mudelite sobitamisel kasutatakse mitmesuguseid esialgse aegrea teisendusi. Neist olulisimad on alljärgnevad:

- vahede arvutamine (diferentsimine).
  Mittestatsionaarsete aegridade hulgas moodustavad olulise klassi nn *statsionaarsete juurdekasvudega aegread.* Selline aegrida pole küll statsionaarne, statsionaarne on selle rea juurdekasvudest, st üksikliikmete vahedest moodustatud (nn *diferentsitud*) aegrida. Mõnikord juhtub ka nii, et aegrida tuleb korduvalt diferentsida, enne kui jõutakse statsionaarse aegreani.

- Üksikväärtuste teisendamine.
  Mõnikord on otstarbekas rakendada esialgse aegrea igale liikmele x_i mingi (üldiselt ühesugune) teisendus $u_i = f(x_i)$, $i = 1, ..., n$. Saadava aegrea liikmetevahelised sõltuvused jm statistilised omadused säilivad, kuid sellise rea jaoks on mõnikord lihtsam leida sobivaid mudeleid. Üks selliseid näiteid on *multiplikatiivse aegrea* teisendamine aditiivseks.

  Aegrida on multiplikatiivne siis, kui tema järgmised liikmed avalduvad eelmistest korrutamise ja astendamise tehte kaudu. Niisugune on olukord näiteks siis, kui on tegemist eksponentsiaalse trendiga, vt punkt 6.1.4. Samuti kui trendi mõju, võib ka sesoonne mõju olla *multiplikatiivne*. Et niisuguseid multiplikatiivseid aegridu käsitleda aditiivsete mudelite abil, kasutatakse nende üksikväärtuste logaritmimist ja üldtuntud asjaolu, et korrutise logaritm võrdub tegurite logaritmide summaga, seega taandub multiplikatiivne aegrida logaritmimise tulemusena aditiivseks.

  Multiplikatiivsete aegridade näiteks võib tuua mitmesuguseid kasvuprotsesse kirjeldavaid aegridu. Kasvamise protsess võib toimuda nii looduses kui ka ühiskonnas selliselt, et iga järgmise ajavahemiku jooksul suureneb uuritav objekt *mingi arv kordi.*

Lisa

Jaotustabelid

Tabel 36. Normaaljaotuse jaotusfunktsioon

x	P(X<x)
-3	0,001
-2,9	0,002
-2,8	0,003
-2,7	0,003
-2,6	0,005
-2,5	0,006
-2,4	0,008
-2,3	0,011
-2,2	0,014
-2,1	0,018
-2,0	0,023
-1,9	0,029
-1,8	0,036
-1,7	0,045
-1,6	0,055
-1,5	0,067
-1,4	0,081
-1,3	0,097
-1,2	0,115
-1,1	0,136
-1,0	0,159
-0,9	0,184
-0,8	0,212
-0,7	0,242
-0,6	0,274
-0,5	0,309
-0,4	0,345
-0,3	0,382
-0,2	0,421
-0,1	0,460
0	0,500

x	P(X<x)
0,1	0,540
0,2	0,579
0,3	0,618
0,4	0,655
0,5	0,691
0,6	0,726
0,7	0,758
0,8	0,788
0,9	0,816
1,0	0,841
1,1	0,864
1,2	0,885
1,3	0,903
1,4	0,919
1,5	0,933
1,6	0,945
1,7	0,955
1,8	0,964
1,9	0,971
2,0	0,977
2,1	0,982
2,2	0,986
2,3	0,989
2,4	0,992
2,5	0,994
2,6	0,995
2,7	0,997
2,8	0,997
2,9	0,998
3,0	0,999

Tabel 37. Studenti t-jaotuse täiendkvantiilide q väärtused (nn kriitilised väärtused)

abadusastmet arv	**Ühepoolne hüpotees:** $P(X > q) = \alpha$		**Kahepoolne hüpotees** $P(\|X\| > q) = \alpha$	
$f = n - 1$	α **=0,05**	α **=0,01**	α **=0,05**	α **=0,01**
1	6,31	31,82	12,71	63,66
2	2,92	6,97	4,30	9,92
3	2,35	4,54	3,18	5,84
4	2,13	3,75	2,78	4,60
5	2,01	3,37	2,57	4,03
6	1,94	3,14	2,45	3,71
7	1,89	3,00	2,36	3,50
8	1,86	2,90	2,31	3,36
9	1,83	2,82	2,26	3,25
10	1,81	2,76	2,23	3,17
12	1,78	2,68	2,18	3,05
14	1,76	2,62	2,14	2,98
16	1,75	2,58	2,12	2,92
18	1,73	2,55	2,10	2,88
20	1,73	2,53	2,09	2,85
25	1,71	2,49	2,06	2,79
30	1,70	2,46	2,04	2,75
40	1,68	2,42	2,02	2,70
60	1,67	2,39	2,00	2,66
120	1,66	2,36	1,98	2,62
∞	1,64	2,33	1,96	2,58

Tabel 38. χ^2-jaotuse täiendkvantiilide h väärtused (kriitilised väärtused)

Vab.-astmeid	P(X>h) = 0,05	P(X>h) = 0,01
1	3,841	6,635
2	5,991	9,210
3	7,815	11,345
4	9,488	13,277
5	11,070	15,068
6	12,592	16,812
7	14,067	18,475
8	15,507	20,090
9	16,919	21,666
10	18,307	23,209
12	21,026	26,217
14	23,685	29,141
16	26,296	32,000
18	28,869	34,805
20	31,410	37,566
25	37,652	45,624
30	43,773	50,892
35	49,802	57,342
40	55,758	63,691
45	61,656	69,957
50	67,505	76,154
60	79,082	88,379
70	90,531	100,425
100	124,32	135,807

Tabel 39. Lineaarse korrelatsioonikordaja kriitilised väärtused.

Vab. Astmete arv	Ühepoolne hüpotees		Kahepoolne hüpotees	
f	α=0,05	α=0,01	α=0,05	α=0,01
1	0,988	0,999	0,997	0,9998
2	0,900	0,980	0,950	0,990
3	0,805	0,934	0,878	0,959
4	0,729	0,882	0,811	0,917
5	0,669	0,833	0,754	0,875
6	0,621	0,789	0,707	0,834
7	0,582	0,750	0,666	0,798
8	0,549	0,715	0,632	0,765
9	0,521	0,685	0,602	0,735
10	0,497	0,658	0,576	0,708
12	0,457	0,612	0,532	0,661
14	0,426	0,574	0,497	0,623
16	0,400	0,543	0,468	0,590
18	0,378	0,516	0,444	0,561
20	0,360	0,492	0,423	0,537
25	0,323	0,445	0,381	0,487
30	0,296	0,409	0,349	0,449
35	0,275	0,381	0,325	0,418
40	0,257	0,358	0,304	0,393
45	0,243	0,338	0,288	0,372
50	0,231	0,322	0,273	0,354
60	0,211	0,295	0,250	0,325
70	0,195	0,274	0,232	0,302
80	0,183	0,257	0,217	0,283
90	0,173	0,242	0,205	0,267
100	0,164	0,230	0,195	0,254

Tabel 40. F-jaotuse kriitilised väärtused olulisuse nivool 0,05

f_1 / f_2	1	2	3	4	5	6	7	8	9	10	12	14	16	20
1	161	200	216	225	230	234	237	239	241	242	244	245	246	248
2	18,5	19,0	19,2	19,3	19,3	19,3	19,4	19,4	19,4	9,39	19,4	19,4	19,43	19,44
3	10,1	9,55	9.28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,74	8,71	8,69	8,66
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,87	5,84	5,80
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,68	4,64	4,60	4,56
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,96	3,92	3,87
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,57	3,52	3,49	3,44
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	3,28	3,23	3,20	3,15
9	5,12	4,26	3,96	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,07	3,02	2,98	2,93
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,91	2,86	2,82	2,77
12	4,75	3,88	3,49	3,26	3,20	3,00	2,92	2,85	2,80	2,76	2,69	2,64	2,60	2,54
14	4,60	3,74	3,34	3,11	3,11	2,58	2,77	2,70	2,65	2,60	2,53	2,48	2,44	2,39
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,37	2,33	2,28
20	4,35	3,49	3,10	2,87	2,71	2,60	2,52	2,45	2,40	2,35	2,28	2,23	2,18	2,12

Aineregister

—K—

—L—

—M—

—N—

—O—

—P—

—R—

—S—

—T—

—U—

—V—

—Ä—

—Ü—