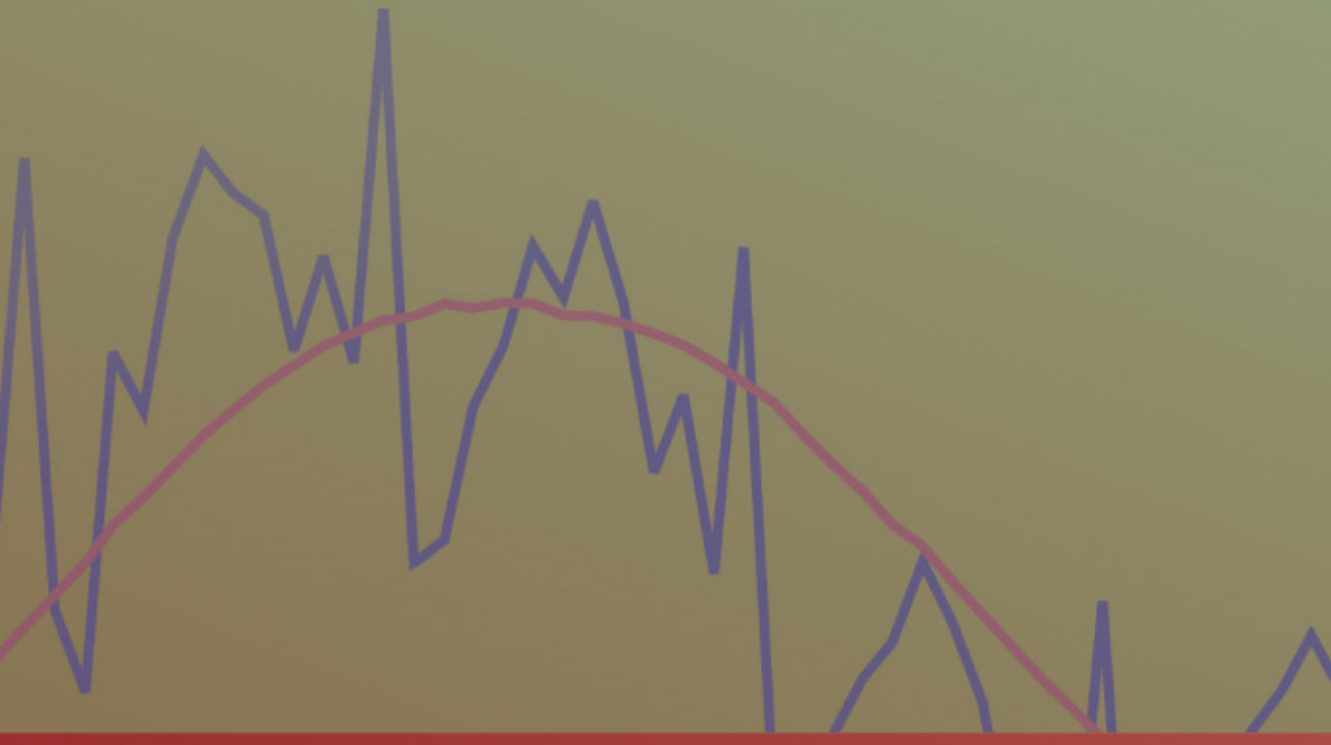
Ako Sauga

STATISTIKA



Statistika õpik majanduseriala üliõpilastele

Ako Sauga

TTÜ
KIRJASTUS

Tallinn 2017

Ilmunud riikliku programmi
„Eestikeelsed kõrgkooliõpikud 2013–2017“ toetusel

Õpiku väljaandmist toetasid

HARIDUS- JA TEADUSMINISTEERIUM

ARCHIMEDES

Retsenseerinud Kalev Pärna ja Elmo Tempel

Kaane kujundanud Tiia Eikholm

Keelekorrektuur: Jane Võrk

ISBN 978-9949-83-082-4
ISBN 978-9949-83-218-7 (pdf)

Sisukord

Eessõna

Eesti keeles on ilmunud mitmeid statistikaõpikuid, mis on mõeldud erinevate erialade üliõpilastele. Seni pole aga ilmunud õpikut, millega oleksid kaasas andmefailid ülesannete jaoks. Statistikat ei saa tulemuslikult õppida ilma praktilise andmeanalüüsita ja käesolev õpik püüabki seda lünka täita.

Õpiku esimestes peatükkides käsitletakse statistika põhimõisteid, andmete esitamist ja kirjeldava statistika suurusi. Seejärel on lühike ülevaade tõenäosusteooria algtõdedest eesmärgiga tuletada meelde koolis õpitu. Tõenäosusteooria põhimõisted on vajalikud järgnevate peatükkide läbimiseks. Pikemalt peatutakse juhuslike suuruste jaotusseadustel. Erinevatel jaotusseadustel põhinevad järeldava statistika meetodid, kuid neil on ka otsene rakendus mitmesuguste äri- ja majandusprobleemide lahendamisel. Järeldava statistika peatükid käsitlevad valikvaatlusi, hüpoteeside kontrollimist, korrelatsioon- ja regressioonanalüüsi ning aegridade silumist ja prognoosimist. Õpiku viimane, 11. peatükk on pühendatud indeksite kasutamisele kui ühele olulisele majandusstatistika valdkonnale. Kui peatükid 1–10 tuleks läbida järjest, siis indeksite peatükk ei ole teistega seotud ja sellega tutvumine võib toimuda ükskõik millisel õppetöö etapil.

Autor ei ole seadnud eesmärgiks esitada statistika teooria matemaatilist käsitlust. Loomulikult päris ilma matemaatikata ei saa. Statistiliste meetodite tulemusrikas kasutamine eeldab nendest arusaamist. Keerulised arvutused teeb arvuti, kuid arvutustulemuste tõlgendamine jääb inimesele. Seepärast on vajalik aru saada, kuidas üht või teist suurust leitakse, mis mõjutab selle väärtust ja mida see meile räägib. Neile, kes matemaatikat ei karda, on mõningate valemite tõestused toodud lisades.

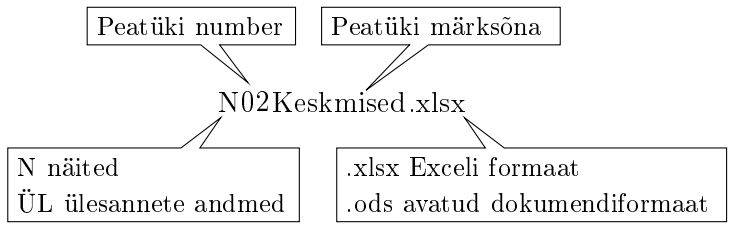
Statistiliste arvutuste metoodikat ja tulemuste tõlgendamist on püütud selgitada õpikus olevate näidetega ning kinnistada iga peatüki lõpus olevate ülesannetega. Kõikide ülesannete vastused on toodud õpiku lõpus. Iga ülesande lõpus on ka viide leheküljele, kus asub vastus. Kuna viited on hüperlingina, siis õpiku pdf-versioonis võimaldab see kiiresti vastuse juurde hüpata. Näidete ja ülesannete koostamisel on võimalikult palju tuginetud reaalsetele andmetele. Andmed pärinevad

avalikest andmebaasidest (Eesti Statistikaamet, Eurostat jt), erialaajakirjadest, mitmesugustest avalikult kättesaadavatest uuringuaruannetest, veebiportaalidest ja õpiku autori juhendatud üliõpilaste aine- ning lõputöödest. Näidete juures kasutatud andmed on võimaluse korral esitatud õpikuga kaasasolevates tabelarvutuse failides. Lisaks on nendes failides tehtud läbi näidetes käsitletud arvutused ning lisatud kommentaare ja nõuandeid tabelarvutuse kasutamiseks. See võimaldab õppijal paralleelselt lugeda õpikut ning uurida näitefaile.

Statistika mõistete ja meetodite paremaks tundmaõppimiseks on õpikus hulgaliselt jooniseid ja graafikuid. Lisaks on soovitav kasutada interaktiivseid demosid õpiku autori kodulehel `http://www.sauga.pri.ee/cdf/`. Demod on Wolfram CDF formaadis (*Computable Document Format*) ning nende vaatamiseks on vaja arvutisse installeerida vabalt levitatav Wolfram CDF Player[1].

Ülesandeid on õpikus kahte tüüpi. Esimese osa moodutavad sellised, mille lahendamiseks pole suuri andmemahtusid vaja ja piisab ülesande tekstis toodud informatsioonist. Enamasti saab neid lahendada kalkulaatori abil ja mõeldud on need valemitega tutvumiseks. Kuid mõningatel juhtudel, näiteks normaaljaotuse kasutamisel, tuleb ka nende ülesannete juures pöörduda tabelarvutuse poole. Alternatiivne võimalus on kasutada õpiku lõpus olevaid tabeleid. Teist liiki ülesanded, mille number algab tähega A, on lahendamiseks tabelarvutusprogrammis ja vajalikud andmed on tabelarvutuse failides.

Õpiku juurde kuulub 10 faili näidetega ja 11 faili ülesannete andmetega, failide nimekiri on lisas D. Failid saab alla laadida TTÜ Raamatukogu digikogust aadressilt `https://digi.lib.ttu.ee/` jaotisest „Õpikud ja õppevahendid“. Failid on kahes formaadis: MS Exceli formaat (*.xlsx) ja avatud dokumendiformaat (*.ods). Viimast kasutavad näiteks vabavaralised tabelarvutusprogrammid LibreOffice Calc[2] ja OpenOffice Calc[3]. Näited ja ülesannete andmed on failidesse grupeeritud peatükkide kaupa. Faili nimi koosneb mitmest osast, nagu on näidatud järgneval skeemil.



Õpiku lõpus lisas C.11 on esitatud vajalike tabelarvutusfunktsioonide loetelu koos viidetega tekstis esinevatele juhistele. Tekstis tähistab

[1] Wolfram CDF Player `https://www.wolfram.com/cdf-player/`

[2] `https://et.libreoffice.org/`

[3] `https://www.openoffice.org/`

vastavat kohta lehe servas olev ruudustikuga pisipilt. Kasutatud tabelarvutusfunktsioonid vastavad Exceli versioonile 2010 ja LibreOffice Calc versioonile 5.1.0.3 ning on neis programmides ühesugused. Exceli eelis seisneb selles, et sel on kaasas andmeanalüüsi vahendite komplekt *Data Analysis*, mis oluliselt lihtsustab hüpoteeside kontrollimist ning regressioonanalüüsi läbiviimist.

Lugemise hõlbustamiseks on tähtsamad tekstis esinevad mõisted toodud eraldi välja lehekülje serval. Tekstis on olulisematele terminitele sulgudes lisatud ka ingliskeelsed vasted.

Tähtis mõiste

Olulised definitsioonid ja valemid on paigutatud varjutatud kastidesse.

Oluline definitsioon või valem.

Näited on ümbritsetud raamiga. Kui näites esitatud arvutused on olemas ka tabelarvutuse failis, asub lehe serval pisipilt, mille all on kirjas vastava faili nimi ja lehe nimi, kus näide asub. Need näited, mille juures pisipilt puudub, on vaid õpikust lugemiseks.

Fail
Leht

Näide 0.1. Mingi näide

Näide

Õpiku kirjutamisel on kasutatud kirjastamispaketti LaTex. Joonised on loodud programmidega Wolfram Mathematica[4] ning Incscape[5]. Pisipiltide loomisel on kasutatud portaalis Pixabay[6] jagatud avalikku omandisse (*public domain*) kuuluvaid pilte. Näidete arvutused on tehtud programmis MS Excel (versioon 2010).

Autor on tänulik õpiku retsensentidele Tartu Ülikooli professorile Kalev Pärnale ning Tartu Observatooriumi vanemteadurile Elmo Tempelile, kelle väärtuslikud nõuanded ja kommentaarid aitasid õpikus esitatud materjali oluliselt parandada.

Käesolev õpik on ilmunud riikliku programmi „Eestikeelsed kõrgkooliõpikud 2013–2017“ toetusel.

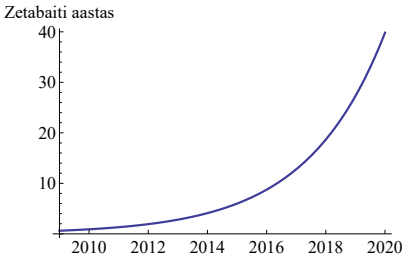
Ako Sauga
Tallinn, jaanuar 2017

[4] Wolfram Research, Inc., Mathematica, Ver. 10.4, `http://www.wolfram.com`
[5] The Inkscape Project, `https://inkscape.org`
[6] `https://pixabay.com`

Peatükk 1

Sissejuhatus

Seoses infotehnoloogia arenguga on oluliselt suurenenud statistilise informatsiooni kasutamine ja selle põhjal järelduste tegemine ning otsuste vastuvõtmine. Infotehnoloogia areng on statistika kasutamist mõjutanud kahel põhjusel. Esiteks on oluliselt arenenud mitmesuguste andmete kogumise võimalused. Ajakirjas „Imeline teadus“ ilmunud artikli põhjal lisandub maailmas iga päev ligikaudu 16,4 miljardit gigabaiti andmeid (Poulsen, 2015). Teiseks on andmete töötlemine muutunud võimalikuks kõigile, kes omavad personaalarvutit või ka nutitelefoni. Andmehulga plahvatuslik kasv pakub üha rohkem võimalusi neile, kes suudavad sellest kaosest vajalikku informatsiooni välja noppida.



Maailmas koguneva andmehulga plahvatuslik kasv (Poulsen, 2015). 1 zetabait (ZB)= 10^{12} gigabaiti (GB)

1.1. Põhimõisted

Mõistel „statistika“ on mitu tähendust:

- andmestik;
- tegevus, mis haarab info hankimist;
- teadus.

Statistika on teadus massnähtuste kvantitatiivse uurimise meetoditest.

Statistika kui teaduse definitsioonis on olulised kaks momenti:

- statistika käsitleb nähtuste kogumeid, mis koosnevad paljudest üksiknähtustest (massnähtused);
- statistika uurib massnähtusi kvantitatiivsest küljest, uuringu tulemused esitatakse arvnäitajate abil ja need sõltuvad vähe uurija tõlgendusest.

Sõltuvalt analüüsi sügavusest jaguneb statistika kaheks.

Kirjeldav ja järeldav statistika

Kirjeldava statistika (*descriptive statistics*) eesmärgiks on statistilise informatsiooni kompaktne ja ülevaatlik esitamine ning kokkuvõtete tegemine. Kasutatakse sobivaid näitarve, tabeleid, diagramme.

Järeldav statistika (*inferential statistics*) on järelduste tegemine spetsiaalseid, tõenäosusteoorial põhinevaid statistilisi meetodeid kasutades. Vaatlustulemusi kasutatakse hinnangute ja prognooside tegemiseks (veel) vaatlemata objektide ja situatsioonide jaoks. Järeldava statistika alla kuuluvad näiteks statistiliste hüpoteeside kontrollimine, korrelatsioon- ja regressioonanalüüs, aegridade analüüs ja prognoosimine.

Järgnevalt tutvume mõningate statistikas kasutatavate põhimõistetega.

Põhimõisted

Element (*element*) on indiviid, ettevõte, nähtus vms, mille kohta kogutakse informatsiooni, mida mõõdetakse, vaadeldakse, küsitletakse.

Tunnus (*variable*) on näitaja, mida mõõdetakse ja mis võib erinevatel elementidel olla erineva väärtusega. Tunnused võivad olla uuritavad tunnused ja tausttunnused.

Varieerumine, hajumine (*variation*) tähendab, et tunnus omandab erinevaid väärtusi.

Variatsioonrida (*ordered sample*) on korrastatud statistiline rida, mille elemendid on järjestatud mingi tunnuse alusel.

Üldkogum (*population*) on elementide hulk, mille kohta soovitakse saada informatsiooni, et lahendada püstitatud probleemülesannet. Sõltuvalt probleemist võib üldkogumiks olla näiteks Harjumaa ettevõtted, Eesti ettevõtted või Euroopa Liidu ettevõtted; Harjumaa elanikud, Eesti elanikud või Euroopa Liidu elanikud.

Osakogum (*subpopulation*) on üldkogumi alamhulk, mis on fikseeritav tausttunnuse või uuritava tunnuse väärtuse järgi ja mida soovitakse eraldi uurida. Näiteks kui üldkogumiks on Eesti elanikkond, siis osakogumiteks võivad olla mehed ja naised või linna- ja maaelanikud.

Valim (*sample*) on üldkogumi alamhulk, mille elemente mõõdetakse ja mille põhjal tehakse järeldusi üldkogumi kohta.

Maht (*size*) on üldkogumi või valimi elementide arv.

Mõõtmismeetod (*measurement method*) on meetod, mille abil saadakse informatsiooni uuritavate objektide kohta.

Mõõtmisvahend (*measurement instrument*) on vahend, mille abil mõõdetakse tunnuse väärtusi. Mõõtmisvahendiks võib olla ankeet või mingi füüsiline instrument (kaalud, kell).

Probleemülesanne (*problem task*) on vastava ainevaldkonna ülesanne, mille lahendamisele saab kaasa aidata statistiline uuring. Probleemülesande analüüs peab selgitama vajaduse statistilise ülesande järele.

Statistiline ülesanne (*statistical task*) on ülesanne, mida saab lahendada statistiliste meetoditega. Statistiline ülesanne püstitatakse probleemülesandest lähtudes. Fikseeritakse üldkogum, uuritavad tunnused ja abitunnused, valimi koostamise meetod, kasutatavad statistilised näitajad, väljastatavad tabelid, lisatavad usaldushinnangud. Valikumeetodi ja valimi mahu planeerimisel tuleb arvestada olemasoleva aja, majanduslike ja tehniliste ressurssidega.

Mingi probleemi statistilisel analüüsimisel tuleb alati läbida järgmised **tööetapid**:

Tööetapid

1. Andmete kogumine.
2. Andmete kokkuvõte ja töötlemine. Kasutatakse erinevat tarkvara: tabelarvutusprogrammid MS Excel, OpenOffice, LibreOffice Calc või spetsiaalsed statistikakapaketid SPSS, SAS, R (vabavara).
3. Tulemuste analüüs ja tõlgendamine (interpreteerimine), järelduste ning üldistuste tegemine.
4. Tulemuste esitamine: tabelid, diagrammid, valemid.

1.2. Mõõtmine ja mõõteskaalad

Mõõtmine on objektide võrdlemine. Korraga saab võrrelda ainult kaht objekti omavahel; kui objekte on rohkem, võrreldakse neid paarikaupa. Näiteks võime võrrelda Juku ja Kalle pikkust ning saada tulemuseks, et nende pikkused on erinevad. Siis võime võrrelda Juku ja Mati pikkust ning võrdlustulemuseks saada, et need poisid on ühepikkused. Sellisel võrdlemisel piirdusime võimaliku ekvivalentsuse või mitteekvivalentsuse määramisega.

Täiuslikuma võrdlemise korral määratakse objektide järjestus võrreldava tunnuse järgi. Väide „Juku on Kallest pikem“ sisaldab rohkem informatsiooni kui väide „Juku ja Kalle on erineva pikkusega“, seepärast on järjestust määrav võrdlemine informatiivsem. Joonisel 1.1 on toodud viie poisi järjestus pikkuse järgi, vasakul on kõige lühem.

Veel täiuslikuma võrdlemise korral määratakse tunnuse väärtuste arvuline suhe. Näiteks Juku ja Kalle pikkuse võrdlemisel saame tulemuseks, et Juku on Kallest 1,03 korda pikem ehk Juku pikkus on 1,03 Kalle pikkust. Priidu ja Toomase pikkuste võrdlustulemuseks on,

Kalle Juku Priit Toomas
Mati

Joonis 1.1. Viie poisi järjestus pikkuse järgi, vasakul on kõige lühem

Mõõtühik

et Priidu pikkus on 0,97 Toomase pikkust. Mõistlik on aga valida üks poiss välja ning võrrelda ülejäänud poiste pikkusi tema omaga. Sellist erilist objekti, millega ülejäänud objekte võrreldakse, nimetatakse mõõdetava suuruse etaloniks. Etaloniga kindlaks määratud mõõdetava suuruse väärtus on **mõõtühik**. Tabeli 1.1 teises ja kolmandas veerus on poiste pikkused, kui ühel juhul on etaloniks võetud Kalle ja teisel juhul Toomas. Paraku ei ütle need arvud midagi inimesele, kes pole näinud Kallet või Toomast ja ei tea, kui pikad need poisid on. Seepärast on mõistlik kasutada üldlevinud standardseid mõõtühikuid, pikkuse mõõtmisel näiteks meetrit (tabel 1.1 viimane veerg).

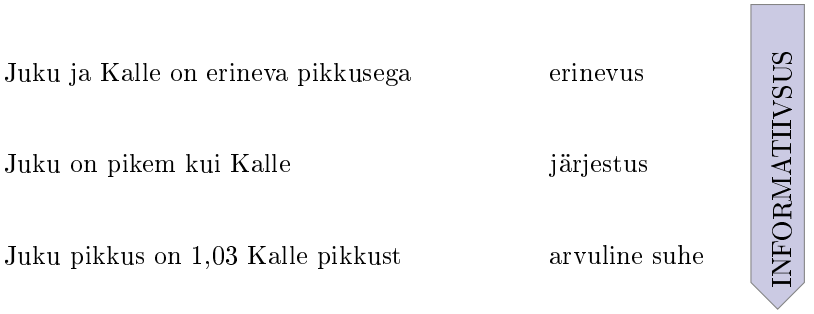
Tabel 1.1. Poiste pikkused, kui etaloniks on Kalle, Toomas ja standardühik meeter

Õpilane	Õpilase pikkus, kui mõõtühikuks on Kalle pikkus	Õpilase pikkus, kui mõõtühikuks on Toomase pikkus	Õpilase pikkus, kui mõõtühikuks on meeter
Juku	1,03	0,92	1,65
Kalle	1	0,89	1,60
Mati	1,03	0,92	1,65
Priit	1,09	0,97	1,75
Toomas	1,13	1	1,80

Näeme, et õpilase pikkust väljendav arv sõltub sellest, millist mõõtühikut kasutatakse. Seepärast tuleb mõõdetud suuruse väärtusele alati **lisada mõõtühik**. Mõõtühik näitab, millise objektiga võrdlemisel on mõõtesuurus saadud. Mis kasu on meil näiteks teadmisest, et 2015. aasta Mazda CX-5 hind on 3 miljonit, kui me ei tea, mis rahaühikutes see hind on antud? Või kuidas suhtuda informatsiooni, et 2014. aastal oli USA SKP 17,4 triljonit, aga Ungari SKP 38,5 triljonit? Kummalgi juhul pole arvude taga ühikuid ning me ei tea, millega on võrreldud Mazda väärtust, USA SKP-d või Ungari SKP-d.

Mazda hind 3 miljonit saadakse Mazda väärtuse võrdlemisel jeeni väärtusega, järelikult Mazda hind on 3 miljonit jeeni. USA SKP oli 17,4 triljonit USA dollarit ning Ungari SKP 38,5 triljonit forintit. Kui me tahame USA ja Ungari SKP-d võrrelda, tuleb nende mõõtmiseks kasutada samu ühikuid. Teades, et 1 forint on 0,00357 USA dollarit, saame Ungari SKP väärtuseks 0,137 triljonit USA dollarit. Nüüd saame järeldada, et USA SKP oli ligikaudu 127 korda suurem kui Ungari SKP.

Erinevad mõõtmistasemed on kokkuvõtlikult esitatud joonisel 1.2. Nägime, et pikkuse mõõtmisel on võimalik kasutada kõiki kolme taset. Kõikide tunnuste korral see nii aga pole. Näiteks rahvuste, ettevõtete tegevusalade, ametite korral saame kasutada vaid eristamise taset.



Joonis 1.2. Erinevad mõõtmistasemed

Mõõtmistasemetele vastavad **mõõteskaalad**, mida on kolme tüüpi.

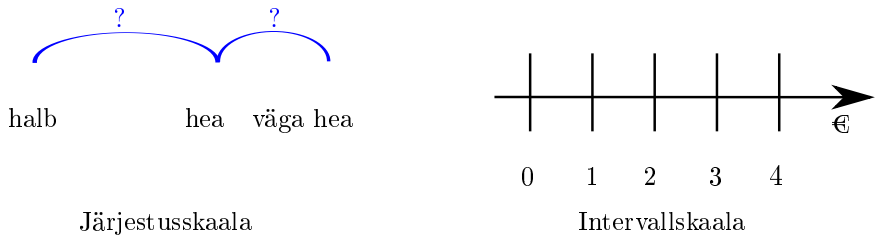
Mõõteskaalad

1. **Nimiskaalat** ehk nominaalskaalat (*nominal scale*) kasutatakse objektide eristamiseks. Näiteks rahvus, huvid, tegevusalad, sportlaste numbrid, telefoninumbrid, kaupade koodid. Kuigi viimaste näidete korral kasutatakse numbreid, siis nende numbritega ei ole mõtet teha arvutusi, sest need ei ole arvud. Eraldi võib välja tuua kaheväärtuselised ehk binaarsed tunnused, s.o tunnused, millel on ainult kaks väärtust. Näiteks sugu: „mees“/„naine“, vastusevariandid: „jah“/„ei“.
2. **Järjestusskaala** (*ordinal scale*) korral saab väärtusi reastada kasvavas või kahanevas järjestuses, kuid intervallid skaalajaotiste vahel pole määratud. Näiteks ettevõtete jaotamine väikesteks, keskmisteks ja suurteks. Haridustasemed on algharidus, põhiharidus, keskharidus. Vastusevariandid ankeedis võivad olla: „poolt“, „pigem poolt kui vastu“, „pigem vastu kui poolt“, „vastu“. Järjestusskaala väärtused ei näita, kui palju üks väärtus teisest erineb.
3. **Intervallskaalal** (*interval scale*) on skaalajaotiste intervallid ühepikkused. Intervallskaalas on näiteks inimese vanus, sissetulek, testimisel saadud õigete vastuste arv, töötajate arv, kauba hind või kaal. Intervallskaalad jagunevad kaheks:
   - **vahemikskaalal** on nullpunkti asukoht kokkuleppeline (näiteks Celsiuse skaala temperatuuri mõõtmiseks, ajaskaala). Vahemikskaala korral võib leida vahesid, ei tohi leida suhteid;
   - **suhteskaala** korral on nullpunkt fikseeritud absoluutselt (näiteks pikkus, kaal, hind).

Intervallskaala võib olla

- diskreetne (loendamisel saadud naturaalarvud);

- pidev (näiteks ajaskaala).



Joonis 1.3. Järjestusskaala jaotiste vahemikud pole teada. Intervallskaala korral on vahemikud teada ja need on kõik ühepikkused, joonisel näiteks 1 euro

Intervallskaala on kõige **informatiivsem** skaala ja selles skaalas mõõdetud tunnuste töötlemiseks on kõige rohkem statistilisi meetodeid. Järjestusskaala on vähem informatiivne, sest puudub informatsioon skaalajaotiste vahemike kohta. Kõige vähem informatsiooni annavad nimiskaalas mõõdetud tunnused.

Arvutusi saab teha ainult intervallskaalas mõõdetud tunnuste korral, sest selle skaala väärtusteks on arvud. Järjestusskaala korral arvutusi teha ei tohiks. Selle reegli vastu eksitakse aga tihti, kui järjestusskaala väärtused kodeeritakse numbriliselt.

Näide 1.1. Hinded kõrgkoolis

Haridus- ja teadusministri määruse „Ühtne hindamissüsteem kõrgharidustasemel, koos diplomi kiitusega (*cum laude*) andmise tingimustega“ alusel tuleb eristava hindamise korral hinnata õppijate õpiväljundite saavutatuse taset järgmise skaala alusel:

1) „suurepärane“ — silmapaistev ja eriti laiapõhjaline õpiväljundite saavutamise tase, mida iseloomustab väga head taset ületav teadmiste ja oskuste vaba ning loov kasutamine;
2) „väga hea“ — väga heal tasemel õpiväljundite saavutamine, mida iseloomustab teadmiste ja oskuste eesmärgipärane ja loov kasutamine. Spetsiifilisemate ja detailsemate teadmiste ja oskuste osas võivad ilmneda mittesisulised ja mittepõhimõttelised eksimused;
3) „hea“ — heal tasemel õpiväljundite saavutamine, mida iseloomustab teadmiste ja oskuste eesmärgipärane kasutamine. Spetsiifilisemate ja detailsemate teadmiste ja oskuste osas avaldub ebakindlus ja ebatäpsus;
4) „rahuldav“ — piisaval tasemel õpiväljundite saavutamine, mida iseloomustab teadmiste ja oskuste kasutamine tüüp-

olukordades, erandlikes olukordades avalduvad puudujäägid ja ebakindlus;

5) „kasin“ — minimaalselt lubataval tasemel olulisemate õpiväljundite saavutamine, mida iseloomustab teadmiste ja oskuste kasutamine tüüpolukordades piiratud viisidel, erandlikes olukordades avalduvad märgatavad puudujäägid ning ebakindlus;
6) „puudulik“ — õppija on omandanud teadmised ja oskused miinimumtasemest madalamal tasemel.

Selline hinnete kirjeldamine on piisav, et järeldada: hinne „suurepärane“ on parem kui hinne „väga hea“, hinne „väga hea“ parem kui „hea“ jne. Kas on aga garanteeritud see, et hinde „suurepärane“ saanud õppur on hinde „väga hea“ saanud õppurist sama palju targem, kui hinde „väga hea“ saanud õppur „hea“ saanud õppijast, kas hinnete vahekaugused on võrdsed? Ilmselt mitte. Tegemist on järjestusskaalaga.
Mitmetes ülikoolides on nendele hinnetele seatud vastavusse numbrilised koodid:

Hinne	Kood
suurepärane	5
väga hea	4
hea	3
rahuldav	2
kasin	1
puudulik	0

Kas nüüd võime väita, et hinde „5“ saanud õppur on hinde „4“ saanud õppurist sama palju targem, kui hinde „4“ saanud õppur „3“ saanud õppijast? Kas hinde „4“ saanud üliõpilane on kaks korda targem kui hinde „2“ saanud üliõpilane? Ei, kodeerimine ei muuda skaalaväärtuste sisu ega taga seda, et hinne „4“ on kaks korda parem kui hinne „2“. Veerus „Kood“ toodud numbrid ei ole arvud, vaid sümbolid (*characters*).

Kui on raske määrata, kas tegemist on järjestus- või intervallskaalaga, tuleks endale esitada küsimus: „**Kas 3 ja 2 erinevus on sama suur kui 2 ja 1 erinevus?**“ Kui see on garanteeritud, siis on tegemist intervallskaalaga. Kui see aga garanteeritud pole, siis on tegemist järjestusskaalaga. Lisaks võib küsida: „**Kas 4 on kaks korda suurem kui 2?**“ Jaatav vastus tähendab, et tegemist on suhteskaalaga. Kui aga esimesele küsimusele vastame jaatavalt ja teisele eitavalt, siis on tegemist vaheskaalaga (näiteks ajaskaala).

Tunnuste tüübid

Tunnuse mõõtmisel kasutatav skaala määrab vastava **tunnuse tüübi**. Tunnuste liigitus tüübi järgi:

- nominaaltunnused (nimiskaala);
- järjestustunnused (järjestusskaala);
- arvtunnused (intervallskaala).

Nominaaltunnuseid ja järjestustunnuseid nimetatakse ka **kvalitatiivseteks** tunnusteks, arvtunnuseid aga **kvantitatiivseteks** tunnusteks. Arvtunnuse saab teisendada järjestustunnuseks või nominaaltunnuseks. Vastupidine teisendus ei ole võimalik.

Näide 1.2. Üleminek intervallskaalalt järjestusskaalale

Ettevõtted jaotatakse töötajate arvu järgi järgmistesse gruppidesse:

mikroettevõte	0 kuni 9 töötajat;
väikeettevõte	10 kuni 49 töötajat;
keskmine ettevõte	50 kuni 249 töötajat;
suurettevõte	250 ja enam töötajat.

Ettevõtte töötajate arv on intervallskaalas. Töötajate arvu põhjal määratud suurusgrupid on aga järjestusskaalas. Kui on teada kolme ettevõtte töötajate arv, siis võime määrata, millisesse gruppi iga ettevõte kuulub.

Ettevõte	Töötajate arv (intervallskaala)	Grupp (järjestusskaala)
A	5	mikroettevõte
B	20	väikeettevõte
C	40	väikeettevõte

Intervallskaalas mõõdetud tunnuse „Töötajate arv“ põhjal võime öelda, et ettevõte C on ettevõttest B kaks korda suurem. Järjestusskaalas mõõdetud tunnuse „Grupp“ väärtus on neil ettevõtetel aga ühesugune. Ettevõtete A ja B võrdlemisel intervallskaalas mõõdetud tunnuse alusel saame öelda, et ettevõte B on ettevõttest A neli korda suurem. Järjestusskaalas mõõdetud tunnuse põhjal aga võime vaid järeldada, et ettevõte B on suurem kui ettevõte A.

Nagu näitest 1.2 nägime, läheb üleminekul intervallskaalas mõõdetud tunnuselt järjestusskaalas mõõdetud tunnusele informatsiooni kaduma. Seepärast ei ole selline üleminek üldiselt soovitatav. Seda võib kasutada kirjeldavas statistikas, kui objekte on palju ja soovime anda ülevaatlikku pilti objektide jaotusest näiteks diagrammi kujul.

Statistiliste näitarvude ning analüüsimeetodite valik sõltub sellest, millist skaalat vastava tunnuse mõõtmisel kasutatakse.

Seepärast on vaja iga mõõtmise korral kindlaks teha kasutatud skaala tüüp. Madalama taseme (vähem informatiivsema) skaala korral kasutatavaid meetodeid võib rakendada ka kõrgema taseme skaala korral, vastupidine pole võimalik.

Otsene ja kaudne mõõtmine

Otsese mõõtmise korral määratakse mõõtmisele kuuluva tunnuse väärtus muid tunnuseid mõõtmata. Nii saab määrata näiteks kauba kaalu, inimese vanust, ettevõtte töötajate arvu. **Kaudse mõõtmise** korral mõõdetakse otseselt teisi, mõõtmisele kuuluva tunnusega mingil viisil seotud tunnuseid ja arvutatakse uuritava tunnuse väärtus. Näiteks tööjõu tootlikkuse mõõtmiseks on vaja otseselt mõõta töötajate arv ning toodangu maht (tk). Tööjõu tootlikkus on siis toodangu maht jagatud töötajate arvuga, ühikuks tükki ühe töötaja kohta.

1.3. Andmekogumismeetodid ja statistiliste vaatluste liigitus

Eksperiment ja vaatlus

Andmete kogumise meetoditeks on eksperiment ja vaatlus. **Eksperimendi** käigus asetatakse uurimisobjekt erilistesse tingimustesse: manipuleeritakse ühe või mitme tunnusega (sõltumatud tunnused), kõrvaldatakse võimalikud kõrvalmõjud (muud tunnused hoitakse konstantsetena) ja mõõdetakse sõltuva tunnuse väärtusi. Eksperimenti kasutatakse peamiselt loodusteadustes, samuti psühholoogias.

Majandus- ja äristatistika korral on põhiliseks andmekogumismeetodiks vaatlus. **Vaatlust** kasutatakse siis, kui huvipakkuva tunnusega manipuleerimine on praktilistel või eetilistel kaalutlustel võimatu, kui uuritavat objekti ei saa paigutada uurija poolt valitud tingimustesse. Ka vaatluse käigus võib uurija objekti mõjutada, kuid see pole eesmärgiks. Näiteks küsitluse läbiviimisel võib küsitleja käitumine ja hääletoon mõjutada vastajat.

Seda, kas uuritava tunnuse muutumise põhjuseks on mingi seletava tunnuse muutumine, saab kindlaks teha eksperimendiga. Enamasti mõjutavad uuritavat tunnust paljud erinevad tegurid ning eksperimendi käigus hoitakse muud tegurid konstantsetena. Vaatluse korral pole selline kontrollitavus võimalik. Seepärast pole enamasti ainult vaatlusele tuginedes võimalik otsustada põhjuse ja tagajärje vahelise seose üle.

Vaatluste liigitus

Statistilisi vaatlusi saab liigitada mitmel moel. Liigitus vaatluse **otstarbe** järgi:

1) esmane ehk primaarne vaatlus korraldatakse konkreetselt antud uuringu jaoks. Näiteks viiakse läbi vastav küsitlus;

2) teisesel ehk sekundaarsel vaatlusel kasutatakse varem muul eesmärgil kogutud andmeid.

Uurimis- või lõputöö kirjutamisel võib üliõpilane kasutada andmeid Eesti Statistikaameti või Eurostati andmebaasist, sellisel juhul on tegemist teisese vaatlusega. Kui ta aga viib läbi eraldi küsitluse ning kasutab oma töös neid andmeid, on tegemist esmase vaatlusega.

Liigitus **andmete hankimisviisi** järgi:

1) **otsese** vaatluse puhul registreerib vaatleja oma silmaga või vastava tehnilise vahendi abil vaatluse ajal toimunud tõsiasju. Näiteks poekülastajate registreerimine, autode loendus maanteel;
2) **küsitluse** ajal registreeritakse küsitletavatelt isikutelt saadud vastused:
   (a) **suuline küsitlus** (vestlus, intervjuu), kus vaatlusalusega vesteldakse ja saadud vastused registreeritakse varem koostatud küsitluslehele;
   (b) **ankeetvaatlus**, mille puhul lastakse uuritavatel vastata kirjalikult või elektroonselt ankeetlehe küsimustele. Tänapäeval kasutatakse sagedasti mitmesuguseid veebipõhiseid ankeetvaatlusi;
   (c) **korrespondentvaatlus**, kus korrespondendid (eraisikud või ettevõtted) regulaarselt koguvad ja saadavad andmeid varem koostatud programmi või küsitluslehe alusel;
3) **dokumentaalvaatlus** kujutab endast kirjalikul või elektroonsel kujul olevate allikate uurimist. Nendeks allikateks võivad olla ettevõtte aruandlusdokumendid, mitmesugused andmebaasid, arhiividokumendid.

Veebis on hulgaliselt elektroonseid andmepankasid, mida saab kasutada dokumentaalvaatluse läbiviimiseks:

- Eesti Statistikaameti andmebaas `http://pub.stat.ee`;
- Eesti Pank `http://www.eestipank.ee`;
- Euroopa Liidu Statistikaamet Eurostat `http://ec.europa.eu/eurostat`;
- Rahvusvaheline Valuutafond (IMF) `http://dsbb.imf.org`;
- Maailma Kaubandusorganisatsioon `http://www.wto.org`;
- Maailmapank `http://data.worldbank.org`;
- Ühinenud Rahvaste Organisatsioon `http://data.un.org`;
- OECD riikide majandusstatistika `http://stats.oecd.org`.

Statistilise andmestiku kogumisel on kaks võimalust: mõõta kõiki üldkogumi objekte (kõikne statistika) või ainult teatavat osa nendest. Liigitus **vaatlusobjekti hõlmamise ulatuse** järgi on järgmine:

1) **monograafilise vaatluse** korral jälgitakse ainult ühte elementi ja tehakse selle alusel järeldus terve kogumi kohta;
2) **võrdlev-monograafilise vaatluse** puhul võetakse kogumist välja kaks erinevat, enamasti kaks äärmist elementi (parim

ja halvim või suurim ja väikseim ettevõte). Näiteks audiitorfirma KPMG uuringus ettevõtete aruandlusest osales 90 ettevõtet kümnest riigist ja viiest tegevusvaldkonnast. Igast tegevusvaldkonnast võeti üks suurettevõte ja üks väikeettevõte;

3) **põhimassi vaatlusega** hõlmatakse peamine osa vaatlusobjektist, kõrvalise tähtsusega osa jäetakse välja (turu-uuringud). Näiteks, kui kollektiivis on mõned liikmed teistega võrreldes väga erinevad oma võimetelt, võivad nad vaatlustulemusi äärmuslikult mõjutada ja nad jäetakse tavaliselt vaatlustest välja;
4) **valikvaatluse** puhul hõlmatakse objektist ainult suhteliselt väike osa. Mitmesuguste meetoditega valitakse üldkogumist välja **valim**. Mida ühtlasem on üldkogum, seda väiksem võib olla valim;
5) **kõikne statistika**. Vaadeldakse kogu üldkogumit. Näiteks rahvaloendus, mitmesugused registrid.

Liigitus **vaatluse sageduse** järgi:

1) pideva vaatluse korral jälgitakse mõnd kindlat tegevust teatud aja jooksul pidevalt;
2) korduv vaatlus toimub perioodiliselt kindla ajavahemiku järel (näiteks 1 nädal);
3) ühekordne vaatlus.

Suurandmed

Eraldi tuleks nimetada **suurandmeid** (*Big Data*). Need on ettevõtete, organisatsioonide ja valitsuste valduses olevad hiiglaslikud digitaalsed andmestikud. Näiteks maailma suurima kaubandusketi Walmarti serveritesse koguneb iga tund 2,5 miljonit gigabaiti andmeid klientide tehingute kohta (McAfee ja Brynjolfsson, 2012). Igas minutis saadetakse maailmas 204 miljonit e-kirja, tehakse 4 miljonit Google'i otsingut, Instagrammi laetakse 216 tuhat fotot ja Twitteris tehakse 277 tuhat säutsu (Knoblauch, 2014).

Suurandmetel on kolm põhilist omadust, mida inglise keeles tähistatakse tähekombinatsiooniga VVV:

- suur maht (*volume*);
- pidev ja kiire andmevoog (*velocity*);
- andmeformaatide paljusus (*variety*).

Suurandmete tekkimiseks vajaminev andmevoog pärineb peamiselt kolmest allikast: internet, info- ja kommunikatsioonitehnoloogia ning mitmesugused sensorid. Mõningaid näiteid suurandmete tekkimiskohtadest:

- kaardimaksed,
- veebikaubandus,
- sotsiaalvõrgustikud,
- turvakaamerad,
- mobiilirakendused,
- GPS-rakendused.

Kui riiulite vahel kõndiva ostja eelistustest saab poeomanik teada alles ostu sooritamisel, siis veebipoe omanikuni jõuab info ka selle kohta, milliseid kaupu klient lihtsalt vaatas. Amazon kasutab seda infot personaalsete pakkumiste tegemisel. Massachusetsi Tehnoloogiainstituudi uurimisgrupp prognoosis USA kaubandusketi Macy's käivet, enne kui kaubanduskett ise müügitehingud registreeris, kasutades selleks mobiilpositsioneerimise andmeid (McAfee ja Brynjolfsson, 2012). Otsingumootoritesse sisestatavate märksõnade analüüsimisel võib saada informatsiooni finantsturgude käitumise kohta (Preis, Reith ja Stanley, 2010), prognoosida inflatsiooni (Guzmán, 2011) või muutusi töötuse määras (Ettredge, Gerdes ja Karuga, 2005). Rakenduse HealthMap seiresüsteemid analüüsivad pidevalt terviseraporteid, blogisid ja sotsiaalmeediat, et leida viiteid mõnele uuele nakkushaiguse puhangule. 2014. aastal prognoositi nõnda ebola epideemia puhkemist (Poulsen, 2015). Eesti on Euroopas esimene riik, kes kasutab suurandmeid riikliku statistika tegemisel: mobiilpositsioneerimise andmete põhjal tehakse turismistatistikat. Mõttekoja demosEUROPA (*Centre for European Strategy Foundation*) arvutuste kohaselt mõjutavad suurandmed potentsiaalselt enam kui poolt Euroopa Liidu majandust.

Ettevõttes võib suurandmete kasutamine mõjutada näiteks kolme valdkonda (Rõõm, 2014):

1) ressursside parem kasutamine ning toodangu müügi- ja turustusprotsesside efektiivsemaks muutumine;
2) toodete ja protsesside paranemine innovatsiooni kaudu, mis on lihtsam tänu pidevale jälgimisele ja tarbijate tagasisidele;
3) juhtimise kvaliteedi parandamine tõenduspõhiste otsuste abil.

Suurandmetega seoses tuleks nimetada veel järgmisi andmetüüpe:

- **avaandmed** (*open data*) on masinloetavas formaadis andmed, mis on antud kõigile vabalt ja avalikult kasutamiseks;
- **linkandmed** (*linked data*) on sellisel viisil avaldatud struktureeritud andmed, mis lubab neid automaatselt seostada;
- **metaandmed** (*metadata*) on andmeid kirjeldavad andmed. Need annavad teada, kes, mida, kus, millal ja kuidas tegi. Näiteks e-kirja saatmise ja avamise aeg, telefonikõne tegemise aeg, pikkus ja osapooled, mingi dokumendi loomise aeg. Metaandmed aitavad infoobjekte idendifitseerida ja luua otsisüsteeme.

Järgnevalt mõningad portaalid, kust võib leida vabalt kasutatavaid suurandmeid ja neil põhinevaid rakendusi:

- *Amazon Web Services Public Data Sets* `http://aws.amazon.com/datasets/` võimaldab ligipääsu erinevatele andmetele astronoomia, bioloogia, geograafia, keemia, kliima, majanduse ja matemaatika vallast;

- Eesti Avaandmete portaal `https://opendata.riik.ee/` sisaldab Eesti avaliku sektori juurdepääsupiiranguteta andmeid;
- *Data Portals* `http://dataportals.org/` on kokku kogunud viited erinevate riikide avaandmete portaalidele;
- *Google Trends* `https://trends.google.com/trends/` võimaldab analüüsida valitud otsingusõnade esinemissagedust maailma erinevates piirkondades ning keeltes;
- *Quandl* `https://www.quandl.com/` vahendab miljoneid majandus- ja finantsaegridu;
- *Wolfram Alpha* `https://www.wolframalpha.com/` võimaldab teha päringuid ja arvutusi andmetega, mis pärinevad paljudest erinevatest andmebaasidest.

Suurandmed liigituvad pidevalt kogutavate otseste vaatlusandmete alla. Üheks oluliseks suurandmete hüveks on nende objektiivsus. Küsitluse korral võib inimene ennast teadlikult või ebateadlikult näidata paremana, kui ta tegelikult on. Suurandmete korral seda probleemi ei ole, sest inimese käitumine registreeritakse automaatselt.

Enamasti ei ole suurandmed kogutud statistilise analüüsi tegemiseks, selleks tuleb need andmestikud süstematiseerida ja töödelda. Vaja on uusi analüüsimeetodeid ja spetsiaalset tarkvara (Morton, Runciman ja Gordon, 2014; Varian, 2014). Ettevõtted peavad investeerima inimestesse, kel on võime näha erinevate andmete kooskasutamise potentsiaali ning oskus neid analüüsida.

1.4. Mõõtemääramatus, valiidsus ja mõõtevead

Mõõtmist kui protsessi võib iseloomustada mitmesuguste kriteeriumidega.

Mõõtemääramatus on tingitud sellest, et tihti pole meil võimalik uuritava tunnuse väärtust täpselt määrata. Me saame vaid kindlaks teha, et see väärtus jääb teatud vahemikku. Näiteks soovime määrata mingi tunnuse väärtust üldkogumis ja kasutame selleks valikvaatlust. Sellisel juhul on tulemuseks, et üldkogumi väärtus jääb vahemikku $x \pm \Delta x$, kus x on valikvaatlusel saadud väärtus ja Δx mõõtemääramatus. Mõõtemääramatust võib põhjustada ka mõõtevahend: mõõtes objekti pikkust millimeeterjoonlauaga, saame objekti pikkuseks $x \pm 0{,}5\,\mathrm{mm}$, sest täpsemalt me selle joonlauaga mõõta ei saa. Ka mõõdetav objekt ise võib põhjustada mõõtemääramatust. Objekti pikkuse mõõtmisel võib probleemiks olla otspindade ebatasasus. Isiku kuusissetuleku mõõtmisel on probleemiks sissetuleku kõikumine erinevatel kuudel. Probleemiks võib olla ka mõõdetava suuruse puudulik defineerimine. Näiteks palume inimestel ankeedis märkida oma kuusissetulek,

aga me ei täpsusta, kas bruto- või netosissetulek. Hiljem me ei tea, kes on märkinud oma brutosissetuleku ja kes netosissetuleku. Tekibki mõõtemääramatus, mis on bruto- ja netosissetulekute vahe. Võib öelda, et mõõtemääramatus tekib nii mõõdetava objekti kui ka mõõtmise ebatäiuslikkusest.

Valiidsus ehk kehtivus näitab, kas me mõõdame ikka seda, mida oleme tahtnud mõõta. See probleem tekib näiteks mitmesuguste ankeetküsitluste korral. Oletame, et tahame mõõta töötajate rahulolu. Selleks tuleb meil koostada vastav küsimustik. Probleem on selles, milliseid küsimusi esitada, et vastused peegeldaksid võimalikult täpselt töötajate rahulolu. Valiidsuse suurendamiseks esitatakse erineva sõnastusega küsimusi ning siis võrreldakse nende vastuseid. Mõningatel juhtudel saab mõõtmise valiidsust hinnata, kui võrrelda antud mõõtmismeetodi kasutamisel saadud tulemusi mõnel teisel meetodil mõõdetuga. Kui mõõtmistulemus võrdub tegeliku väärtusega, siis öeldakse, et mõõtmine on valiidne.

Usaldusväärsus ehk reliaablus näitab, kuivõrd stabiilsed on mõõtmistulemused. Usaldusväärsust saab kontrollida kordusmõõtmiste abil. Mõõtes kirjutuslaua laiust, võime üks kord saada tulemuseks 120,1 cm, teine kord 119,9 cm ja kolmas kord 120,0 cm. Sentimeetri täpsusega langevad need tulemused kokku: 120 cm. Kui kordusmõõtmiste tulemused on oluliselt erinevad, siis pole mõõtmine usaldusväärne ehk reliaabel. Ei saa usaldada mõõtmisprotseduuri, mis ühel juhul annab laua laiuseks 120 cm, teisel juhul 100 cm ja kolmandal mõõtmisel 150 cm.

Ökonoomsus iseloomustab seda, kui palju ressursse (aeg, raha) nõuab mõõtmine.

Erinevusi tegelike väärtuste ja mõõtmise tulemusena saadud väärtuste vahel nimetatakse **mõõtevigadeks**. Vead võivad olla tahtlikud ja mittetahtlikud. Tahtlike vigade korral moonutab andmete esitaja või koguja andmeid meelega. Mittetahtlikud vead jagunevad kolmeks: süstemaatilised vead, juhuslikud vead, eksed.

Mõõtevead

Süstemaatilise vea põhjuseks võib olla ebatäpne mõõtmisvahend (kell, kaal), halvasti sõnastatud küsimus ankeedis. Süstemaatilisi vigu on võimalik avastada kordusmõõtmisel, kui kasutada teist mõõtmisvahendit või -meetodit.

Juhuslik viga mõjutab mõõtmistulemust kord ühes, kord teises suunas. Näiteks võib eksida testil saadud punktide loendamisel. Mingi tegevuse jaoks kulunud aja mõõtmisel tekib reaktsioonikiirusest tingitud viga.

Ekse on jäme viga ja enamasti on põhjustatud inimlikest eksimustest. Näiteks mõõtmistulemus kirjutati kogemata valedes ühikutes, jäeti sisestamata üks arvus esinev 0 (või lisati ülearune 0), vahetati ära tunnuste järjekord, unustati ühe tunnuse väärtus üles kirjutada ja seetõttu kirjutati teiste tunnuste väärtused valedesse kohtadesse jms.

Ekset ei tohi segi ajada erindiga. Erind on selline tunnuse väärtus, mis on saadud korrektse mõõtmise tulemusena, aga erineb oluliselt ülejäänud väärtustest. Et otsustada, kuidas erindiga käituda (kas jätta sisse või eemaldada andmestikust), tuleb lähtuda ülesande püstitusest.

1.5. Tunnuste kodeerimine

Arvutis teostatavaks andmetöötluseks on mittearvuliste ehk kvalitatiivsete tunnuste väärtused vaja **kodeerida** — igale väärtusele tuleb seada vastavusse mingi arv.

Nimiskaala korral pole üldiselt oluline, millistele väärtustele millised arvud vastavad. Kahe vastusevariandi korral on soovitav kodeerida „ei“ — 0 ja „jah“ — 1.

Järjestusskaala korral tuleks kodeerimisel jälgida, et koodid säilitaksid järjestuse, vt näiteks hinnete kodeerimist näites 1.1. Ei tohiks kasutada järgmist kodeeringut:

hea	1
halb	2
ei oska öelda	3

Tuleks kasutada kodeeringut:

hea	1
ei oska öelda	2
halb	3

Eraldi tuleb kodeerida puuduvad vastused. See võimaldab hiljem kontrollida, kas kõik vastused on ikka sisestatud. Kui puuduvad vastused pole spetsiaalselt kodeeritud, ei saa kindlaks teha, kas vastaja jättis sellele küsimusele vastamata või unustas andmete sisestaja vastuse sisestamata. Puuduva vastuse koodiks võib valida „0“, kui see ei saa olla vastusevariant. Kui aga küsitakse näiteks kuusissetulekut või laste arvu, siis koodi „0“ ei saa puuduva vastuse kodeerimiseks kasutada. Sellisel juhul tuleb valida selline arv, mis ei saa olla vastuseks. Näiteks sissetuleku korral „-1“ või laste arvu korral „100“.

Näide 1.3. Kodeerimine leibkonna eelarve uuringus

Näitena on toodud Eesti Statistikaameti poolt teostatavas iga-aastases leibkonna eelarve uuringus kasutavad kodeeringud (Leibkonna eelarve uuring 2012). Kvalitatiivsed tunnused on siin „Sugu“, „Rahvus“, „Leibkonna elukoht piirkonna järgi“ ning „Eluruumi tüüp“. Nende tunnuste väärtused on kodeeritud. Lisaks on eraldi koodid puuduvate väärtuste tähistamiseks. Puuduvad väärtused võivad olla tingitud sellest, et vastaja ei tea näiteks maja ehitusaastat või ei soovi vastata.

Tunnuse selgitus	Väärtuste selgitus	Ei tea või keeldub vastamast
Sugu	1 mees	
	2 naine	
Rahvus	1 eesti	
	2 muu	9
Leibkonna elukoht piirkonna järgi	1 Põhja-Eesti	
	2 Kesk-Eesti	
	3 Kirde-Eesti	
	4 Lääne-Eesti	
	5 Lõuna-Eesti	9
Maja ehitusaasta		999
Eluruumi tüüp	1 eramaja	
	2 mitmepereelamu	
	3 korterelamu	
	4 muu	9
Eluruumi pind liikme kohta		999

1.6. Andmete esitamine

Andmete säilitamiseks ja töötlemiseks kasutatakse kas spetsiaalset andmetöötlustarkvara (näiteks tasulised SPSS, SAS, Stata, vabalt kasutatavad PSPP, Gretl, R) või tabelarvutust (MS Excel, OpenOffice.org Calc, LibreOffice Calc, Google arvutustabel). Andmed on struktureeritud tavaliselt nii, et objektidele vastavad read (kirjed) ning tunnustele veerud (väljad). Tabelarvutuses saab andmeid ka pöörata ehk transponeerida.

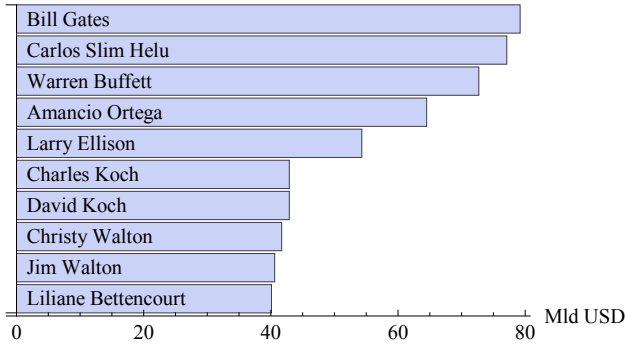
Andmete visuaalseks esitamiseks kasutatakse mitmesuguseid tabeleid ja diagramme. Kui me soovime näidata meid huvitavate tunnuste väärtusi kõikide objektide korral ning objekte ei ole palju, võib kõik need väärtused eraldi välja tuua nagu näiteks tabelis 1.2. Väärtuste võrdlemiseks on aga sobivam esitada need diagrammil, kus erinevused on paremini märgatavad (joonis 1.4.) Diagrammil saab esitada tabelis 1.2 toodud tunnustest korraga vaid ühe, sest näiteks vara ja vanus on erinevates mõõtühikutes.

Sagedustabel

Kui soovime anda ülevaadet suuremast arvust objektidest, kasutatakse **sagedustabelit**, kus on esitatud iga väärtuse esinemissagedus antud kogumis. Tabelis 1.3 on toodud riikide nimed, kust pärinevad maailma 50 kõige rikkamat inimest. Iga riigi korral on esitatud selle esinemise sagedus vastavas andmekogumis. Sageduste summa on võrdne kogumi mahuga.

Tabel 1.2. Maailma kümme kõige rikkamat inimest 2015. aasta algul (Kroll ja Dolan, 2015)

Jrk nr	Nimi	Puhasvara, mld USD	Vanus	Riik
1	Bill Gates	79,2	59	USA
2	Carlos Slim Helu	77,1	75	Mehhiko
3	Warren Buffett	72,7	84	USA
4	Amancio Ortega	64,5	78	Hispaania
5	Larry Ellison	54,3	70	USA
6	Charles Koch	42,9	79	USA
6	David Koch	42,9	74	USA
8	Christy Walton	41,7	60	USA
9	Jim Walton	40,6	67	USA
10	Liliane Bettencourt	40,1	92	Prantsusmaa



Joonis 1.4. Maailma kümne rikkama inimese varandus 2015. aasta algul tabelist 1.2

Tabel 1.3. Millistest riikidest pärinevad 50 maailma kõige rikkamat inimest (Kroll ja Dolan, 2015)

Riik	Sagedus	Riik	Sagedus
USA	26	Brasiilia	1
Saksamaa	5	Hispaania	1
Hiina	3	Jaapan	1
India	3	Kanada	1
Hongkong	2	Mehhiko	1
Itaalia	2	Rootsi	1
Prantsusmaa	2	Saudi Araabia	1

N01Sissejuhatus
T1.2-5

Sagedustabeleid saab koostada nimiskaalas, järjestusskaalas ja intervallskaalas mõõdetud tunnuste korral. Intervallskaala puhul on sagedustabel ülevaatlik siis, kui erinevaid väärtusi ei ole eriti palju.

Kui intervallskaalas mõõdetud tunnusel on palju erinevaid väärtusi, ei ole kasulik neid kõiki sagedustabelis üles lugeda — tabel muutub ebaülevaatlikuks. Üksikväärtuste asemel valitakse sobivad väärtuste vahemikud ehk sagedusklassid (intervallid, *bin*) ning variatsioonrida esitatakse intervallreana (vt tabelid 1.4 ja 1.5). Seda protseduuri nimetatakse **intervallimiseks**. Rida lüheneb tunduvalt ja on ülevaatlikum, kuid osa informatsiooni läheb kaduma.

Intervallimine

N01Sissejuhatus T1.2-5

Tabel 1.4. Maailma 50 kõige rikkama inimese vanuseline jaotus (Kroll ja Dolan, 2015). Kuna ühe isiku vanus puudus, on sageduste summa 49

Vanus	Sagedus
kuni 40	1
41–50	6
51–60	11
61–70	9
71–80	14
81–90	6
91 ja rohkem	2

Tabel 1.5. Maailma 50 kõige rikkama inimese varanduse suuruse jaotus (Kroll ja Dolan, 2015). Intervallid on ligikaudu piiratud

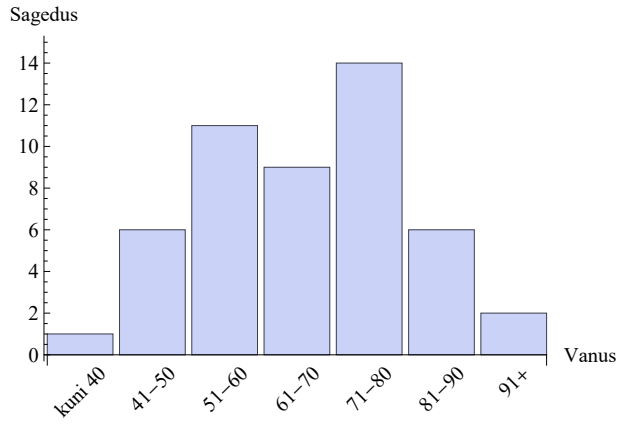
Puhasvara, mld USD	Sagedus
kuni 20	7
20–30	25
30–40	8
40–50	5
50–60	1
60–70	1
70–80	3

Sagedusklassi iseloomustavad **alumine piir**, **ülemine piir** ja **laius**. Klassi laius arvutatakse kas sama vahemiku ülemise ja alumise piiri vahena või kahe naabervahemiku ülemiste väärtuste vahena. Tabelis 1.4 on intervallid **rangelt piiratud** (41–50, 51–60, 61–70) ning nende klasside laius on 10 aastat. Varanduse suurus (tabel 1.5) on aga pidev tunnus ning siin ei saa võtta näiteks klassile „kuni 20“ järgneva klassi alumiseks piiriks väärtust 21, sest vara suurus võib olla ka 20,3 mld USD. Selles tabelis on intervallid **ligikaudu piiratud**. Väärtus, mis täpselt võrdub klassi ülemise piiriga, kuulub sellesse klassi. Näiteks kui varanduse väärtus on 30 mld USD, siis see kuulub klassi „20–30“. Äärmised intervallid võivad olla ka **lahtised** ehk **avatud** („kuni 20“, tabelis 1.4 „kuni 40“, „91 ja rohkem“).

Diagrammi, mis iseloomustab meid huvitava tunnuse jaotust ja on koostatud intervallrea põhjal, nimetatakse **histogrammiks** (joonis 1.5).

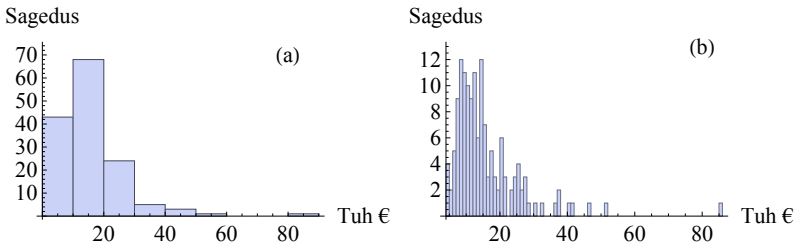
Histogramm

Sagedusklassidesse ehk intervallidesse jagamisel tuleb määrata klasside arv ja laius. Kui tunnusel on väärtusi väga palju, võib sobivate intervallide leidmine olla üpris problemaatiline. Joonisel 1.6 on toodud



Joonis 1.5. Maailma 50 kõige rikkama inimese vanuseline jaotus. Histogramm on loodud tabeli 1.4 põhjal

ühe ja sama kogumi põhjal koostatud histogrammid, kus klassi laius on erinev.



Joonis 1.6. Poe päevakäibe jaotus, 145 päeva. Minimaalne väärtus 4,1 tuhat eurot, maksimaalne väärtus 85,3 tuhat eurot. (a) Klassi laius on 10000 €, klasside arv 9. (b) Klassi laius on tuhat eurot, klasside arv 90

Sturges (1926) soovitas sagedusklasside arvu k määramiseks kasutada järgmist valemit:

Sturgesi valem

$$k = 1 + \log_2 n, \tag{1.1}$$

kus n on elementide arv. Kuna klasside arvuks saab olla vaid täisarv, tuleb valemi (1.1) põhjal leitud k ümardada täisarvuni. Seda valemit kasutab histogrammi konstrueerimisel enamik tarkvarapakette. Tabelis 1.6 on toodud Sturgesi valemi põhjal leitud soovitatav klasside arv mõningate n väärtuste korral.

Võrdsete laiustega sagedusklasside korral on klassi laius Sturgesi valemi kasutamisel:

$$d = \frac{x_{\max} - x_{\min}}{k} = \frac{x_{\max} - x_{\min}}{1 + \log_2 n}, \tag{1.2}$$

Tabel 1.6. Sturgesi valemi (1.1) põhjal leitud klasside arv

Elementide arv n	Klasside arv k
10	4
20	5
50	7
100	8
200	9

kus $x_{\max}$ ja $x_{\min}$ on vastavalt kogumi maksimaalne ja minimaalne väärtus. Valemi (1.2) alusel leitud klassi laiust võib kohandada nii, et klasside piirid oleksid sobivad. Võimaluse korral valitakse piirid nii, et arvu lõpus on 0 või 5.

N01Sissejuhatus J1.6

Joonisel 1.6 toodud poe päevakäivete korral, kui $n = 145$, $x_{max} = 85{,}3$ tuhat eurot ja $x_{min} = 4{,}1$ tuhat eurot, annab valem (1.1) tulemuseks

$$1 + \log_2 145 \approx 8{,}18,$$

ja klassi ligikaudne laius

$$d = \frac{85{,}3 - 4{,}1}{8{,}18} \approx 9{,}92.$$

Mõistlik on klassi laiuseks võtta 10 tuhat eurot ja klasside arvuks 9. See vastab joonisel 1.6(a) toodud diagrammile.

Sturgesi valem sobib hästi siis, kui $n < 200$ ja jaotus on ligikaudu sümmeetriline. Asümmeetria korral on vajalike klasside arv tavaliselt suurem. Kuna joonisel 1.6 toodud jaotus ei ole sümmeetriline, siis klasside arv võiks olla suurem kui Sturgesi valemi abil leitud 9. Optimaalne klasside arv sõltub ka tunnuse väärtuste hajumisest ning seetõttu on erinevate autorite poolt väljapakutud valemeid, mis võtavad hajumist arvesse (vt ptk 3.11).

Paljudes statistikapakettides on võimalik jaotuse iseloomustamiseks histogrammi asemel kasutada ka pidevat kõverjoont, mis saadakse spetsiaalse silumismeetodi abil (kerneli ehk tuumaga silumine).

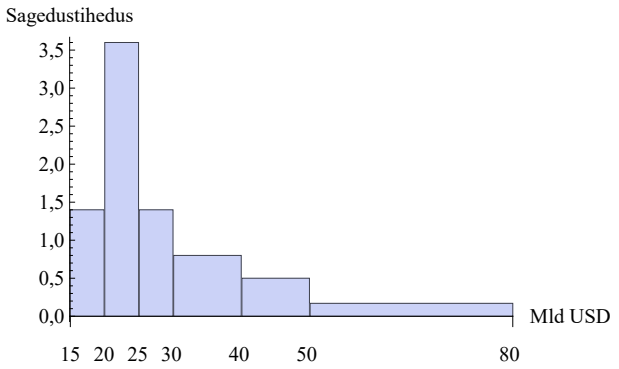
Kui võrdsete laiustega klasside korral on sagedused väga erinevad, võib kasutada muutuva laiusega klasse: piirkonnas, kus väärtusi on rohkem, võtta klassi laius väiksem. Tabelis 1.5 on klasside laiused võrdsed ning ühe klassi sagedus on 26, aga kahe klassi sagedus vaid üks. Tabelis 1.7 on varade jaotuse iseloomustamiseks kasutatud ebavõrdsete laiusega klasse. Sellisel juhul ei saa me aga võrrelda klasside sagedusi.

Sagedustihedus

Võrdlemiseks kasutatakse **sagedustihedust**, mis on sagedus jagatud klassi laiusega. Vastava histogrammi koostamisel kantakse püstteljele sagedustihedus (joonis 1.7). Sagedus on siis sagedustiheduse ja klassi laiuse korrutis, mis on võrdne tulba pindalaga.

Tabel 1.7. Maailma 50 kõige rikkama inimese varanduse suuruse jaotus. Klassid on varieeruva laiusega, erinevate klasside võrdlemiseks kasutatakse sagedustihedust

Puhasvara, mld USD	Klassi laius	Sagedus	Sagedus-tihedus
15–20	5	7	$7/5 = 1{,}4$
20–25	5	18	$18/5 = 3{,}6$
25–30	5	7	1,4
30–40	10	8	0,8
40–50	10	5	0,5
50–80	30	5	0,17



Joonis 1.7. Maailma 50 kõige rikkama inimese varanduslik jaotus. Histogramm on loodud tabeli 1.7 põhjal, tulpade kõrgus vastab sagedustihedusele. Klassi sageduse saame, kui vastava tulba kõrguse korrutame tulba laiusega, s.t sagedus on võrdne tulba pindalaga

Tabelarvutuses kasutatakse väärtuste loendamiseks ja sageduste saamiseks vastavaid funktsioone. Üksikute väärtuste loendamiseks sobib funktsioon **COUNTIF**(*range*;*criteria*). See väljastab tingimusele *criteria* vastavate piirkonnas *range* olevate väärtuste sageduse. Väärtusteks võivad olla nii arvud kui tekst. Selle funktsiooni abil on saadud tabel 1.3.

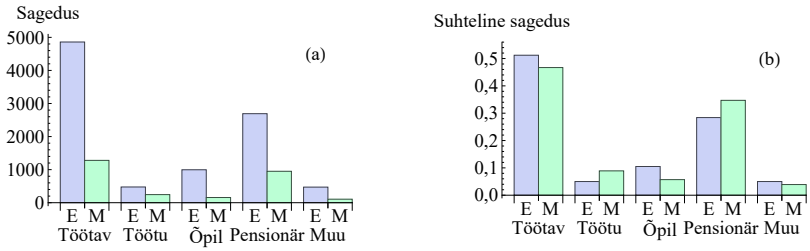
Intervallitud variatsioonrea korral tuleb sageduste saamiseks kasutada tabelarvutusfunktsiooni **FREQUENCY**(*data_array*;*bins_array*). Selle abil on saadud tabel 1.5. Funktsioon väljastab sagedused korraga kõikidesse eelnevalt valitud lahtritesse ning selle sisestamine erineb oluliselt tavaliste funktsioonide sisestamisest. Seda tüüpi funktsioone nimetatakse tabelarvutuses massiivifunktsioonideks (*array function*). Funktsiooni FREQUENCY kasutamisõpetust võib vaadata õpiku

autori kodulehel olevalt ekraanivideolt[1]. Seda funktsiooni on kasutatud näiteks tabelite 1.4 ja 1.5 koostamiseks.

Tabelis 1.8 on toodud Eesti Statistikaameti poolt aastal 2013 läbiviidud sotsiaaluuringus osalejate jaotus sotsiaal-majandusliku seisundi järgi, eraldi eestlased ja muust rahvusest isikud. Eestlasi oli kokku 9493 ja muust rahvusest vastajaid 2738.

Tabel 1.8. Eestlaste ja muust rahvusest vastajate jaotus sotsiaal-majandusliku seisundi järgi (*Eesti sotsiaaluuring* 2013)

Seisund	Sagedus		Suhteline sagedus	
	Eestlased	Muu rahvus	Eestlased	Muu rahvus
Töötav	4859	1280	51,2%	46,7%
Töötu	475	245	5,0%	8,9%
Õpilane	996	157	10,5%	5,7%
Pensionär	2692	950	28,4%	34,7%
Muu mitteaktiivne	471	106	5,0%	3,9%
Kokku	9493	2738	100%	100%



Joonis 1.8. Eestlaste „E“ ja muust rahvusest vastajate „M“ jaotus sotsiaal-majandusliku seisundi järgi, loodud tabeli 1.8 põhjal. (a) Sagedused, (b) suhtelised sagedused

Suhteline sagedus

Kui me soovime võrrelda sotsiaal-majandusliku seisundi jaotust nendes kahes kogumis, siis on seda raske teha, sest kogumite mahud on väga erinevad (joonis 1.8 (a)). Seepärast on leitud **suhtelised sagedused**: sagedus jagatud vastava kogumi mahuga. Suhtelised sagedused on erineva mahuga kogumite korral võrreldavad. Jooniselt 1.8 (b) näeme, et sotsiaal-majandusliku seisundi jaotus on eestlaste ja muust rahvusest vastajate hulgas ligikaudu ühesugune.

Diagrammid

Andmete visualiseerimiseks kasutatavate arvjooniste ehk diagrammide tüüpe on palju. Nende liigitus geomeetriliste kujundite järgi on järgmine:

- tulpdiagramm (*Column*),

[1] `http://www.sauga.pri.ee/statistika_excelis`

- lintdiagramm (*Bar*),
- joondiagramm (*Line*),
- sektordiagramm (*Pie*),
- punktdiagramm ehk hajumisdiagramm (*XY Scatter*),
- sõõrdiagramm (*Doughnut*),
- radardiagramm (*Radar*),
- pinddiagramm (*Area*),
- silinder (*Cylinder*),
- koonus (*Cone*),
- püramiid (*Pyramid*),
- muud.

Rõhutada tuleb, et ainult punkt- ehk hajumisdiagrammil on horisontaalteljeks arvtelg. Ülejäänud diagrammitüüpidel on see mõeldud kvalitatiivse tunnuse esitamiseks (*Category axis*).

Diagrammide liigitus kasutusotstarbe järgi:

- võrdlusdiagrammid — kahe või enama nähtuse võrdlemine;
- struktuuridiagrammid — iseloomustatakse nähtuse koostist;
- dünaamikadiagrammid — iseloomustatakse nähtuse muutumist ajas (aegread);
- seosediagrammid — iseloomustatakse nähtuste muutumises ilmnevaid seoseid (hajumisdiagramm);
- jaotusdiagrammid — iseloomustatakse kogumi üksikelementide jaotust mingi tunnuse järgi (histogramm);
- levikudiagrammid — statistilised kaardid, mille abil kirjeldatakse uuritava nähtuse territoriaalset levikut;
- organisatsioonidiagrammid — organisatsiooni struktuuri, alluvusvahekordade kujutamine;
- protsessidiagrammid — mõne protsessi või tegevuse kulg (kasutatakse näiteks projektijuhtimises).

Erinevate diagrammitüüpide kasutamisega ning diagrammide kujundamisega saab tutvuda õpikuga kaasasolevas failis ÜL01Arvjoonised. Failis on erinevad andmetabelid, mille põhjal tuleb koostada sobiv diagramm. Lisaks leiab ülevaate erinevatest diagrammidest ja nende kasutamisest järgmistest õpikutest: August Aarma „Arvjoonised“, Silvi Roomets „Arvjoonised“ või Uno Mereste ja Maimu Saarepera „Arvjoonised“.

1.7. Ülesanded

1.1. Juku kohta on olemas järgmised andmed:

a) sugu: mees;
b) perekonnaseis: abielus;
c) laste arv: 2;

d) haridus: kesk;
e) vanus: 25-aastane;
f) amet: lukksepp;
g) kategooria: III;
h) telefoni nr: 532-1421.

Milliseid mõõteskaalasid on nende andmete korral kasutatud? VASTUS lk 655.

1.2. Ettevõtte iseloomustamisel võib kasutada mitmesuguseid näitajaid:

a) ettevõtte registrinumber äriregistris;
b) töötajate arv;
c) omandivorm: eraettevõte, munitsipaalettevõte, riigiettevõte;
d) krediidireiting: AAA, AA, A, BBB, BB, B, C.

Milliste skaaladega on tegemist nimetatud näitajate korral? VASTUS lk 655.

1.3. 1995. aastal Eestis läbiviidud tööjõu-uuringul esitati muuhulgas järgmised küsimused. Millist skaalat iga küsimuse juures kasutatakse?

1. Kui kergesti Te arvate end tööd leidvat?
   (a) Töö leidmine ei valmistaks erilisi raskusi;
   (b) töö leidmine valmistaks mõningaid raskusi;
   (c) töö leidmine valmistaks suuri raskusi.
2. Teil on praegu töö olemas. Kui tõenäoliseks Te peate, et lähema paari aasta jooksul oma praegusest töökohas lahkute?
   (a) Väga tõenäoliseks;
   (b) üsna tõenäoliseks;
   (c) üsna ebatõenäoliseks;
   (d) väga ebatõenäoliseks.
3. Millise vanuseni Te töötada sooviksite, kui eeldada, et tööd jätkub?
4. Milline on Teie kodune keel?
5. Kas Teie leibkond kasutab praegust eluaset
   (a) omanikuna;
   (b) kooperatiivi või ühistu liikmena;
   (c) üürnikuna;
   (d) allüürnikuna?
6. Mitu tuba on Teie leibkonnal kasutada?
7. Kas Teie leibkond on viimase aasta jooksul saanud eluasemetoetust?
   (a) Jah;
   (b) ei.

VASTUS lk 655.

1.4. Milliste näitajate korral võib leida kahe väärtuse suhte?

1. Ettevõtte A asutamisaasta 2000, ettevõtte B asutamisaasta 2005.
2. Ettevõtte A vanus 15 aastat, ettevõtte B vanus 10 aastat.
3. Toiduaine A säilitustemperatuur +10 °C, toiduaine B säilitustemperatuur −18 °C.
4. Toote A kaal 10 kg, toote B kaal 12 kg.

Vastus lk 655.

Ülesanded arvjooniste kohta on failis ÜL01Arvjoonised

ÜL01Arvjoonised

Failis on erinevad andmetabelid, mille põhjal tuleb koostada sobiv diagramm. Lisatud on pildid, milline peaks olema õige diagramm, ning nõuanded selleni jõudmiseks.

Järgmiste ülesannete andmed on failis ÜL01Sagedustabelid

ÜL01Sagedus-tabelid

A.1.1. Eesti Statistikaamet viib igal aastal läbi leibkonna eelarve uuringut. 2012. aastal osales uuringus 9080 isikut. Üheks uuringuankeedis esitatud küsimuseks oli: „Kuidas tuleb leibkond vajalike kulutuste tegemisel ots otsaga kokku?“ Vastusevariante oli kokku kuus. Tabelis on toodud vastuste sagedus, eestlased ja muust rahvusest vastajad eraldi (*Leibkonna eelarve uuring* 2012).

Leida suhtelised sagedused ja nende põhjal konstrueerida tulpdiagramm, kus on esitatud nii eestlaste kui muust rahvusest vastanute vastuste jaotus. Vastus lk 655.

A.1.2. Seisuga 31.12.2015 oli Eestis arvel 101 769 veoautot[2]. Tabelis on andmed 101 732 veoauto kohta: mark, mudel, keretüüp, väljalaskeaasta ja mootori võimsus. Välja on jäetud need, mille korral väljalaskeaasta või mootori võimsus olid määramata.

1. Koostada sagedustabel erinevate keretüüpide esinemissagedustega.
2. Milline on kõige vanema veoauto väljalaskeaasta?
3. Mitme veoauto väljalaskeaasta on varasem kui 1970?
4. Koostada sagedustabel, mis annaks ülevaate vahemikus 1970–2015 erinevat väljalaskeaastat omavate veoautode esinemise kohta. Lisada tulpdiagramm.
5. Koostada kaks erineva klasside arvuga sagedustabelit mootori võimsuse jaotuse kohta ja lisada vastavad tulpdiagrammid:
   (a) 5 sagedusklassi, klasside laius 100, esimese klassi ülemine piir 100, viimane klass lahtine: > 400;

[2] Allikas: Maanteeameti veebileht `http://www.mnt.ee/`. Sõidukite ja juhilubade statistika. Arvel olevad veoautod seisuga 31.12.2015 (mootori maht ja võimsus).

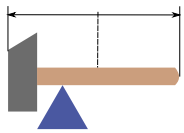
(b) 15 sagedusklassi, klasside laius 50, esimese klassi ülemine piir 50, viimane klass lahtine: > 375.

VASTUS lk 655.

A.1.3. Tabelis on toodud Harju Elektri aktsia sulgemishind 16.08.2013–14.02.2014. Koostada vastav sagedustabel ja hindade jaotust illustreeriv histogramm. Klasside arvu leidmiseks kasutada Sturgesi valemit. VASTUS lk 655.

Peatükk 2

Statistilised keskmised



Mitu keskpunkti on haamril? Ilmselt rohkem kui üks. Võime jagada pikkuse pooleks ja leida geomeetrilise keskpunkti. Võime leida masskeskme, mille suhtes on haamer tasakaalus. Kas me võime öelda, et üks keskmine on tähtsam kui teine? Ilmselt ei või, sest need keskmised väljendavad erinevat informatsiooni.

Samamoodi võib ka mingi tunnuse väärtuste hulga iseloomustamiseks kasutada erinevaid keskmisi. Tuntumad neist on aritmeetiline keskmine, mediaan ja mood. Teatud juhtudel kasutatakse ka harmoonilist, geomeetrilist ja ruutkeskmist.

2.1. Aritmeetiline keskmine

Kõige tuntum statistiline keskmine on aritmeetiline keskmine. Olgu näiteks antud viie inimese vanused: 21, 19, 25, 19, 23. Vanuste aritmeetiline keskmine on

$$\bar{x} = \frac{21 + 19 + 25 + 19 + 23}{5} = 21{,}4.$$

Olgu üldkogumi maht n ja selle elementidel mingi kvantitatiivse tunnuse väärtused x_i. Üldkogumi **aritmeetiline keskmine** on

$$\bar{x} = \frac{\sum x_i}{n}. \tag{2.1}$$

Aritmeetiline keskmine

Valemis (2.1) tähistab kreeka täht „sigma“ $\sum$ summeerimist üle kõigi väärtuste x_i:

$$\sum x_i = x_1 + x_2 + \ldots + x_n. \tag{2.2}$$

Summa (2.2) jagamine arvuga n tähendab, et me jagame selle summa võrdselt n objekti vahel. Näiteks kui me kasutame keskmise palga leidmiseks aritmeetilist keskmist, siis me jagame kogu palkadeks makstud summa võrdselt kõigi n palgasaajate vahel.

Teiselt poolt, teades aritmeetilist keskmist $\bar{x}$ ja elementide arvu n, saame leida summa

$$\sum x_i = n\bar{x}. \tag{2.3}$$

Ühes vanas IV sajandil kirjapandud India loos prognoosis loo peategelane Rtuparna puu küljes olevate puulehtede arvu nii, et loendas üle ühe oksa küljes olevad lehed ning korrutas selle arvu puuokste arvuga (Bakker ja Gravemeijer, 2006). Tulemuseks saadud 2095 oli väga lähedal arvule, mille ta sai peale päevapikkust kõigi lehtede üleloendamist. Nii saab juhtuda, kui uuritud oksa küljes olevate lehtede arv vastab ligikaudu aritmeetilisele keskmisele, see oks esindab kõiki ülejäänud oksi. Aritmeetiline keskmine on selles mõttes esinduslik, et seda ja elementide arvu teades on võimalik leida summat (2.3).

Kuna aritmeetilise keskmise valemis tuleb tunnuse väärtused liita, saab seda keskmist kasutada ainult intervallskaala korral, kui tunnuse väärtusteks on arvud. Järjestusskaala korral ei ole tegemist arvväärtustega ja järelikult ei saa neid liita ega leida aritmeetilist keskmist. Ei saa ju sooritada järgmist tehet:

$$\frac{\text{väga hea} + \text{hea} + \text{hea}}{3}$$

Ka siis, kui me kasutame järjestusskaala väärtuste numbrilist kodeerimist, ei muuda see skaalaväärtuste sisu ja meil ei ole tegemist arvudega. Sama kehtib nimiskaalas mõõdetud tunnuste kohta.

> Aritmeetilist keskmist saab kasutada ainult intervallskaala korral.

Kuid ka intervallskaalas mõõdetud suuruste korral ei ole aritmeetilise keskmise kasutamine alati sobiv.

Näide 2.1. Keskmine vanus ja aritmeetiline keskmine

Ettevõttes töötab neli inimest, kelle vanused on 21, 22, 24 ja 28 aastat. Vanuste aritmeetiline keskmine on 23,75 aastat. Võetakse tööle uus inimene, kelle vanus on 50 aastat. Nüüd on vanuste aritmeetiline keskmine 29 aastat. Üks ekstreemne vanus mõjutas aritmeetilist keskmist nii, et viie inimese keskmine vanus on suurem kui nelja inimese vanus nendest viiest! Kas niimoodi leitud keskmine vanus on esinduslik, s.t kas annab meile õiget informatsiooni ettevõtte töötajate keskmise vanuse kohta?

Aritmeetiline keskmine on tundlik ekstreemsete väärtuste suhtes.

Sellest tulenevalt ei sobi aritmeetilist keskmist kasutada kogumite korral, kus on üksikuid ekstreemselt suuri või väikesi väärtusi.

Näide 2.2. Omavalitsusametnike palgad

30. sept 2005. aastal ilmus Postimehes artikkel „Tallinna linnaametnike palgad on tõusnud üle 13 000 krooni“. Artiklis võrreldi omavalitsusametnike keskmisi palkasid Valga- ja Hiiumaal ning Tallinnas. Võrdlemisel kasutati palkade aritmeetilist keskmist. „Selgus, et kõige suuremat palka, keskmiselt 13136 krooni kuus, saavad Tallinna omavalitsusametnikud. Kõige vähem teenivad Valga- ja Hiiumaa omavalitsuste ametnikud, kus keskmine palk on 5000–6000 krooni.“ Millest siis nii suur keskmise palga erinevus? Põhjus selgub artikli lõpus: „Tallinna linnavalitsuse keskmiste palkade puhul tuleb samas arvesse võtta seda, et keskmist palgataset tõstavad oluliselt Tallinna linnajuhid, kelle põhipalgad ulatuvad sõltuvalt ametikohast 17 000 kuni 29 000 kroonini.“

Andmete esitamiseks kasutatakse tihti sagedustabelit. Kui kogumis on näiteks 100 objekti, siis ülevaatliku pildi saamiseks loendatakse iga väärtuse esinemise sagedus. Sagedustabelis esitatakse tunnuse väärtused ja nende sagedused. Sellisel juhul ei saa me aritmeetilise keskmise leidmiseks kasutada valemit (2.1).

Näide 2.3. Keskmine tubade arv pere kohta

Elamistingimuste uurimiseks küsitleti sadat peret nende kasutuses olevate tubade arvu kohta. Andmed on esitatud grupeeritult sagedustabelina. Leiame keskmise tubade arvu ühe pere kohta.

Variant i	Tubade arv (variantide arvväärtused x_i)	Perede arv (esinemissagedus ehk kaal f_i)
1	1	12
2	2	30
3	3	23
4	4	19
5	5	9
6	6	7

Keskmise tubade arvu ühe pere kohta saame, kui summaarse tubade arvu jagame perede arvuga. Summaarse tubade arvu saamiseks leiame algul, kui palju tube on kokku nendel peredel, kellel on üks tuba ($12 \cdot 1$), nendel peredel, kellel on kaks tuba ($30 \cdot 2$) jne. Seejärel liidame saadud korrutised. Arvutuste organiseerimisel on otstarbekas lisada veerg „Korrutised“.

Variant i	Tubade arv (variantide arvväärtused x_i)	Perede arv (esinemissagedus ehk kaal f_i)	Korrutised $f_i x_i$
1	1	12	12
2	2	30	60
3	3	23	69
4	4	19	76
5	5	9	45
6	6	7	42
KOKKU		100	304

Aritmeetiline keskmine on summaarne tubade arv jagatud perede arvuga:

$$\frac{304}{100} = 3{,}04. \tag{2.4}$$

Vastus: keskmine tubade arv ühe pere kohta on 3,04. Loomulikult saab mingil perel olla vaid täisarv tube. Kuid arv 3,04 iseloomustab keskmist peret sel moel, et korrutades seda perede arvuga 100, saame tubade koguarvu.

Näites 2.3 leidsime summaarse tubade arvu 304 summana korrutistest $f_i x_i$: $\sum f_i x_i = 304$, kus f_i oli vastava variandi x_i esinemissagedus kogumis. Selle jagasime perede koguarvuga 100, mis on leitav sageduste summana $\sum f_i = 100$. Arvutuse (2.4) võib esitada ka pikemalt:

$$\frac{12 \cdot 1 + 30 \cdot 2 + 23 \cdot 3 + 19 \cdot 4 + 9 \cdot 5 + 7 \cdot 6}{12 + 30 + 23 + 19 + 9 + 7} = \frac{304}{100} = 3{,}04. \qquad (2.5)$$

Viimast arvutust üldistades saame kaalutud aritmeetilise keskmise valemi.

Kaalutud aritmeetiline keskmine

Kui on antud erinevate väärtuste x_i esinemissagedused f_i, siis aritmeetilise keskmise leidmiseks kasutatakse **kaalutud aritmeetilise keskmise** valemit

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}. \qquad (2.6)$$

Variandi x_i esinemissagedust f_i nimetatakse ka selle variandi **kaaluks**.
Kaalutud aritmeetilise keskmise valemi võib esitada ka osakaalude $p_i = f_i / \sum f_j$ abil:

$$\bar{x} = \sum p_i x_i. \qquad (2.7)$$

Valem (2.7) saadakse valemist (2.6), kui lugejas olev summa lahti kirjutada:

$$\frac{\sum f_i x_i}{\sum f_i} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_n x_n}{\sum f_i} = \frac{f_1}{\sum f_i} x_1 + \frac{f_2}{\sum f_i} x_2 + \ldots + \\ + \frac{f_n}{\sum f_i} x_n = p_1 x_1 + p_2 x_2 + \ldots + p_n x_n = \sum p_i x_i.$$

Tabelarvutusprogrammides on lihtsa aritmeetilise keskmise leidmiseks funktsioon **AVERAGE**. Kaalutud aritmeetilise keskmise leidmiseks funktsioon puudub ning järelikult tuleb kaalutud aritmeetiline keskmine arvutada vastavalt valemile (2.6), organiseerides arvutused nii nagu näites 2.3.

Eelmises alapeatükis nägime, et kui mingil intervallskaalas mõõdetud tunnusel on väga palju erinevaid väärtusi, siis sellise arvukogumi kirjeldamisel kasutatakse tihti intervallimist. Intervallitud rea aritmeetilise keskmise leidmisel tuleb kasutada kaalutud aritmeetilist keskmist.

N02Keskmised
N2.4

Näide 2.4. Keskmine elamispinna suurus elaniku kohta Tallinnas

2011. aastal toimunud rahvaloenduse käigus küsiti muuhulgas ka elamispinna suurust ühe elaniku kohta. Tabelis 2.1 on toodud Tallinnas elavate leibkondade jaotus elamispinna suuruse järgi[a]. Leiame leibkonna keskmise elamispinna suuruse, kasutades aritmeetilist keskmist.

Tabel 2.1. Elamispind ühe elaniku kohta

Elupind elaniku kohta, m^2	Leibkondade arv ehk sagedus
0–12	14 698
12–16	22 599
16–20	22 539
20–24	23 921
24–32	32 360
32–40	26 733
40–52	22 969
52–64	10 362
64–84	7 677

Kasutada tuleb kaalutud aritmeetilise keskmise valemit (2.6), kus sageduseks on vastavasse klassi kuuluvate leibkondade arv. Mis võtta aga intervallide korral väärtuseks x_i? Loogiline on võtta selleks intervalli keskpunkt. Valemi (2.6) kasutamiseks leitakse seejärel vastavad korrutised $f_i x_i$, summeeritakse sagedused ja korrutised ning leitakse nende jagatis. Arvutused on koondatud tabelisse 2.2. Paneme tähele, et sageduste summa on leibkondade koguarv.

Tabel 2.2. Arvutused keskmise elamispinna suuruse leidmiseks

Elupind elaniku kohta, m^2	Leibkondade arv arv ehk sagedus f_i	Intervalli keskmine x_i	Korrutised $f_i x_i$
0–12	14 698	6	88 188
12–16	22 599	14	316 386
16–20	22 539	18	405 702
20–24	23 921	22	526 262
24–32	32 360	28	906 080
32–40	26 733	36	962 388
40–52	22 969	46	1 056 574
52–64	10 362	58	600 996
64–84	7 677	74	568 098
KOKKU	183 858		5 430 674

$$\bar{x} = \frac{\sum_i f_i x_i}{\sum_i f_i} = \frac{5430674}{183858} \approx 29{,}5.$$

Vastus: 2011. aastal oli Tallinnas keskmine elamispind elaniku kohta 29,5 m^2.

[a] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel RL085: tavaeluruumides elavad tavaleibkonnad, 31.detsember 2011.

Mõnikord pole äärmistel klassidel määratud alumist või ülemist piiri. Sellised klassid on lahtised ehk avatud. Avatud klassi keskpunkti leidmisel võetakse klassi laiuseks **kõrvaloleva klassi laius**.

Näide 2.5. Annetamine heategevuseks

2013. aastal viis uuringufirma TNS Emor Eestis läbi uuringu heategevusalaste hoiakute kohta (Heategevusalaste hoiakute uuring 2013). Küsitlusele vastas 1140 Eesti elanikku vanuses 18–60 aastat. Üheks küsimuseks oli: „Kas olete püsiannetaja (toetate mingit valdkonda või ühendust teatud regulaarsusega, rohkem kui üks kord)?“ Püsiannetajatelt küsiti ka, millise summa ulatuses nad püsiannetusi teevad. Pakutud vastusevariandid olid: „kuni 5 €“, „kuni 10 €“, „11–25 €“, „26–50 €“, „50+ €“, „ei täpsusta“. Vastava variandi valinud püsiannetajate protsendid on toodud tabelis, 14% ei täpsustanud summat.

Püsiannetuse summa €	Klassi laius €	Klassi keskpunkt x_i	Protsent püsiannetajatest	Protsent summa märkinutest p_i
kuni 5	5	2,5	35%	40,7%
6–10	5	8	16%	18,6%
11–25	15	18	17%	19,8%
26–50	25	38	10%	11,6%
51 ja rohkem	25	62	8%	9,3%

Paneme tähele, et sagedusklassid on rangelt piiratud ning alumised piirid on kaasa arvatud. Seetõttu kuuluvad näiteks klassi „6–10“ väärtused 6, 7, 8, 9 ja 10 ning keskpunkt on 8. Keskmise püsiannetuse suuruse leidmiseks tuleb kasutada osakaalude abil esitatud kaalutud aritmeetilise keskmise valemit (2.7). Need, kes summat ei märkinud, jätame arvutusest välja ja leiame protsendi summa märkinutest. Viimane intervall „51 ja rohkem“ on avatud. Selle intervalli keskmise leidmisel võtame klassi laiuseks eelneva klassi laiuse 25.

Aritmeetiline keskmine valemist (2.7):

$\bar{x} = 0{,}407 \cdot 2{,}5 + 0{,}186 \cdot 8 + 0{,}198 \cdot 18 + 0{,}116 \cdot 38 + 0{,}093 \cdot 62 \approx 16{,}2.$

Vastus: keskmine püsiannetuse summa on 16,2 eurot.

Mitme kogumi üldkeskmine

Kaalutud aritmeetilist keskmist tuleb kasutada ka juhul, kui meil on antud mitme erineva kogumi aritmeetilised keskmised ning tuleb leida **üldkeskmine**. Olgu meil näiteks kolm arvukogumit A, B ja C keskmistega $\bar{x}_A$, $\bar{x}_B$ ja $\bar{x}_C$. Kogumite mahud on vastavalt n_A, n_B ja n_C. Leiame $n = n_A + n_B + n_C$ arvu aritmeetilise keskmise.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{n_A} x_{Ai} + \sum_{i=1}^{n_B} x_{Bi} + \sum_{i=1}^{n_C} x_{Ci}}{n_A + n_B + n_C}, \tag{2.8}$$

kus x_{Ai}, x_{Bi} ja x_{Ci} on vastavalt kogumitesse A, B ja C kuuluvad arvud. Valemi (2.3) põhjal võime lugejas olevad summad esitada vastavate aritmeetiliste keskmiste abil:

$$\sum_{i=1}^{n_A} x_{Ai} = n_A \bar{x}_A; \quad \sum_{i=1}^{n_B} x_{Bi} = n_B \bar{x}_B; \quad \sum_{i=1}^{n_C} x_{Ci} = n_C \bar{x}_C. \tag{2.9}$$

Pannes seosed (2.9) valemisse (2.8), saame mitme kogumi üldkeskmise arvutamiseks järgmise valemi:

$$\bar{x} = \frac{n_A \bar{x}_A + n_B \bar{x}_B + n_C \bar{x}_C}{n_A + n_B + n_C}, \tag{2.10}$$

mis vastab kaalutud aritmeetilise keskmise valemile (2.6), kus kaaludeks on üksikute osakogumite mahud n_A, n_B ja n_C.

Järgnevalt toome ära aritmeetilise keskmise **matemaatilised omadused**, mis tulenevad otseselt valemist (2.1).

Aritmeetilise keskmise matemaatilised omadused

1. Konstandi aritmeetiline keskmine on võrdne selle konstandiga. Kui $x_1 = x_2 = ... = x_n = a$, siis

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

2. Aditiivsus. Mis tahes suuruste summa (vahe) aritmeetiline keskmine on võrdne nende aritmeetiliste keskmiste summaga (vahega):

$$\overline{x \pm y} = \bar{x} \pm \bar{y}. \tag{2.11}$$

3. Tasakaaluomadus: individuaalväärtuste ja nende aritmeetilise keskmise vaheline hälvete summa on null:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0. \tag{2.12}$$

Seda omadust on lihtne tõestada. Selleks kirjutame summa lahti ja kasutame aritmeetilise keskmise arvutusvalemit (2.1):

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x}) &= x_1-\bar{x}+x_2-\bar{x}+\ldots+x_n-\bar{x}=\sum_{i=1}^{n}x_i-n\bar{x}= \\ &= \sum_{i=1}^{n}x_i-n\frac{\sum_{i=1}^{n}x_i}{n}=\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}x_i=0.\end{aligned}$$

4. Kui vähendada (suurendada) variantide arvväärtusi suvalise arvu a võrra, siis väheneb (suureneb) aritmeetiline keskmine sama arvu võrra:

$$\begin{aligned}\frac{\sum_{i=1}^{n}(x_i+a)}{n}=\frac{\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}a}{n}=\frac{\sum_{i=1}^{n}x_i+na}{n}= \\ =\frac{\sum_{i=1}^{n}x_i}{n}+\frac{na}{n}=\bar{x}+a.\end{aligned}$$

5. Kui kõiki väärtusi vähendada (suurendada) suvaline arv k korda, siis väheneb (suureneb) aritmeetiline keskmine sama arv korda:

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

6. Kui kaalutud aritmeetilise keskmise korral vähendada (suurendada) kõikide väärtuste kaalusid suvaline arv k korda, siis aritmeetiline keskmine ei muutu:

$$\frac{\sum(kf_i)x_i}{\sum kf_i}=\frac{k\sum f_ix_i}{k\sum f_i}=\frac{\sum f_ix_i}{\sum f_i}.$$

2.2. Mediaan

Kui me soovime pikka saia jagada kaheks võrdseks tükiks, siis me murame selle keskelt pooleks. Ladina keeles on keskkoht *medius*. Sealt tulebki termin mediaan, mis tähendab järjestatud väärtuste rea keskpunkti.

Näide 2.6. Keskmise hinnaga teler

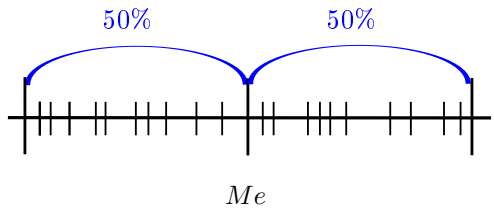
Elektroonikapoodi astub sisse ostja ja pöördub müüja poole: „Sooviksin osta keskmise hinnaga televiisorit.“ Poes on müügil telerid hinnaga 159, 179, 250, 544, 1099, 2677 ja 3999 eurot. Millise hinnaga telerit peaks müüja pakkuma?

Ilmselt soovib ostja telerit, mis poleks liiga odav ega ka liiga kallis, mille hind oleks pakutavate telerite hinnaskaala keskel. Selleks on teler hinnaga 544 eurot.

Järjestatud variatsioonrea keskmise liikme väärtus on mediaan. Mediaanist mõlemale poole jääb ühesugune arv väärtusi.

Mediaan jaotab järjestatud variatsioonrea kaheks võrdseks osaks, nii et mõlemas osas on 50% rea liikmetest.

Mediaani leidmisel ei ole oluline üksikute väärtuste suurus, oluline on vaid nende järjestus. Seepärast saab mediaani kasutada ka järjestusskaala korral. Intervallskaala korral kasutatakse mediaani siis, kui tahetakse kindlaks määrata mingi tunnuse väärtuste jaotuse keskpunkti. Mediaani tähistuseks on tavaliselt Me.



Joonis 2.1. Mediaanist mõlemale poole jääb 50% järjestatud variatsioonrea väärtustest

Mediaani leidmiseks intervallskaala korral tuleb arvud reastada kasvavas (või kahanevas) järjekorras.

- Paarituarvulise variatsioonrea korral on mediaan keskmise elemendi väärtus.
- Paarisarvulise variatsioonrea korral on mediaan kahe keskmise elemendi väärtuste aritmeetiline keskmine.

Kui järjestatud variatsioonreas on n väärtust, siis mediaani järjenumber

$$j = \frac{n+1}{2}. \tag{2.13}$$

Olgu meil näiteks viie mõõtmise tulemused $\{7, 1, 2, 1, 3\}$. Järjestame kasvavas järjekorras: $\{1, 1, \mathbf{2}, 3, 7\}$. Näeme, et mediaan on 2. Võime kasutada ka valemit (2.13). Järjenumber $j = (5+1)/2 = 3$, mis tähendab, et mediaan on kolmas arv.

Lisame arvu 6. Nüüd on kogumis kuus väärtust {1, 1, **2**, **3**, 6, 7}. Mediaan on $\frac{2+3}{2} = 2{,}5$. Valemi (2.13) kasutamisel saame järjenumbriks $j = (6+1)/2 = 3{,}5$, mis tähendab, et mediaan on kolmanda arvu 2 ja neljanda arvu 3 vahel, s.t saame sama tulemuse.

Olgu meil järgmine arvurida {4, 8, 12, 15, 20}. Mediaan on 12. Asendame arvu 20 arvuga 2000, tulemuseks saame arvurea {4, 8, 12, 15, 2000}. Mediaan on ikka 12. Järeldus: kui aritmeetilist keskmist võivad oluliselt mõjutada ekstreemsed väärtused, siis mediaani need ei mõjuta.

Mediaani **omadusi**:

Mediaani omadusi

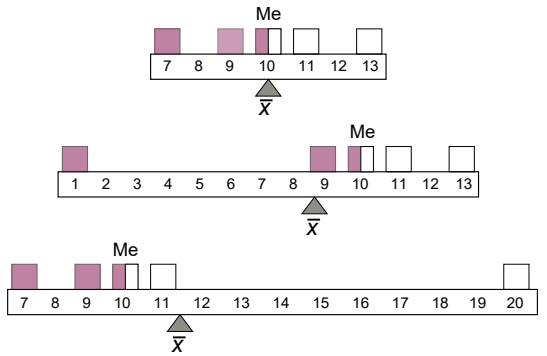
- mediaani võib kasutada järjestusskaala ja intervallskaala korral;
- mediaan ei ole tundlik ekstreemsetele väärtustele.

Milline on aritmeetilise keskmise ja mediaani asend teineteise suhtes? Selleks leiame kolme erineva arvukogumi mediaani ja aritmeetilise keskmise ning võrdleme neid (vt ka joonist 2.2).

A = {7, 9, 10, 11, 13}. Mediaan on 10, aritmeetiline keskmine samuti 10. Mediaan ja aritmeetiline keskmine langevad kokku, jaotus on sümmeetriline.

B = {1, 9, 10, 11, 13}. Mediaan on 10, aritmeetiline keskmine 8,8. Aritmeetiline keskmine on väiksem kui mediaan, sest esineb üks ekstreemselt väike väärtus.

C = {7, 9, 10, 11, 20}. Mediaan on 10, aritmeetiline keskmine 11,4. Aritmeetiline keskmine on suurem kui mediaan, sest esineb üks ekstreemselt suur väärtus.



Joonis 2.2. Mediaani Me ja aritmeetilise keskmise $\bar{x}$ asukohad sümmeetrilise jaotuse korral ning ekstreemselt väikest ja ekstreemselt suurt väärtust omavate jaotuste korral

Mediaani ja aritmeetilise keskmise võrdlus

Ekstreemselt suured või väikesed väärtused kallutavad aritmeetilist keskmist enda poole, mediaani need aga ei mõjuta. Järelikult saame mediaani ja aritmeetilise keskmise võrdlemisel öelda, kas

- jaotus on sümmeetriline: mediaan ja aritmeetiline keskmine langevad ligikaudu kokku;

- esineb ekstreemselt väikesi väärtusi: aritmeetiline keskmine on väiksem kui mediaan;
- esineb ekstreemselt suuri väärtusi: aritmeetiline keskmine on suurem kui mediaan.

Seepärast tuuakse tunnuse kirjeldamisel tihti ära mõlemad keskmised: aritmeetiline keskmine ja mediaan. See annab lugejale lisainformatsiooni: kas jaotus on sümmeetriline või esineb aritmeetilist keskmist nihutavaid ekstremaalseid väärtusi.

Näide 2.7. Kasutatud autode hinnad

Portaalis Forte[a] 10. novembril 2014 ilmunud artiklis „Loe, milline on seis kasutatud autode turul Eesti suurima automüügiportaali andmetel“ on järgmine lõik:
„Möödunud aastaga võrreldes on tänavu oktoobris hinnad püsinud stabiilsena, nii keskmine kui ka mediaanhind. Kui mullu oli portaali auto24 kaudu müüdud sõiduautode keskmine hind 8650 eurot, siis käesoleva aasta oktoobris oli see vaid veidi tõusnud, 9070 eurole. Sama trend toimus mediaanhinnaga — mullu oli see 5700 eurot, aga tänavu 5900 eurot.“
Mediaanhind 5900 eurot näitab, et pooled müüdud autodest olid odavamad kui 5900 eurot ja pooled kallimad kui 5900 eurot. Keskmise hinna all mõeldakse siin ilmselt hindade aritmeetilist keskmist. Kuna aritmeetiline keskmine on suurem kui mediaan, siis oli müüdud autode hulgas üksikuid ekstreemselt kõrge hinnaga autosid.

[a] `http://forte.delfi.ee/`

Vaatleme veel mõningaid näiteid aritmeetilise keskmise ja mediaani kasutamise kohta.

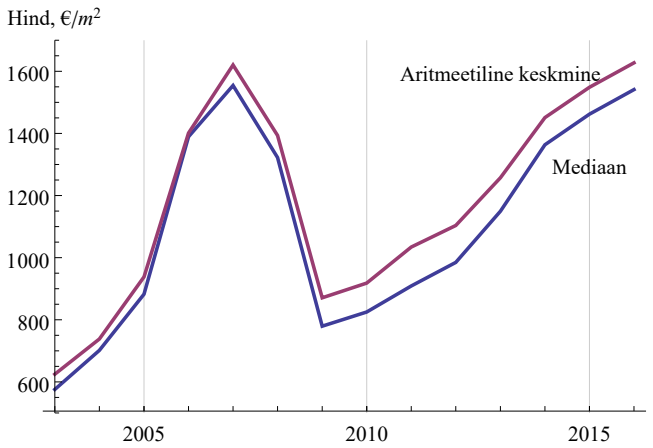
1. Peres on viis liiget: isa ja ema töötavad ning kolm last käivad koolis. Isa palk on 1100 eurot kuus ning emal 900 eurot kuus. Kui suur on pereliikmete keskmine sissetulek? Ilmselt on mõistlik kasutada aritmeetilist keskmist: $(1100 + 900)/5 = 400$ eurot kuus, mis on keskmine sissetulek ühe pereliikme kohta. Mediaani kasutamine ei ole siin sobiv, sest mediaan on 0 eurot kuus.
2. Riikide võrdlemisel elanike rikkuse järgi on sobivam kasutada mediaansissetulekut. Kui kasutada sissetulekute aritmeetilist keskmist, siis see võib olla suur nendes riikides, kus rikkus on koondunud väikese grupi inimeste kätte ja ülejäänud elavad vaesuses. Mediaansissetulek näitab aga seda sissetulekut, millest pooled inimesed saavad rohkem ja pooled vähem.
3. Kui inimene soovib määrata oma asendit palgaskaalal, siis ei sobi selleks võrdlus palkade aritmeetilise keskmisega. Asendi määra-

miseks tuleb kasutada mediaani. Sellega võrdlemine annab infot, kas konkreetse inimese palk on palgasaajate esimeses või teises pooles.

4. Kinnisvara hindade muutumist on sobiv kirjeldada mediaanhinna abil. Tehinguhindade aritmeetilist keskmist võib oluliselt mõjutada üks väga kalli kinnisvaraga tehtud tehing.

Näide 2.8. Korterite ruutmeetri hind Tallinnas

Joonisel on toodud Tallinna korterite ruutmeetri hinna aritmeetiline keskmine ja mediaan aastatel 2003–2016[a]. 2003–2009 olid aritmeetiline keskmine ja mediaan ligikaudu ühesuurused. Järelikult sel ajal oli hindade jaotus ligikaudu sümmeetriline. Kuni aastani 2007 kinnisvarahinnad tõusid („kinnisvaramull"), siis „mull lõhkes" ja hinnad hakkasid langema. Aastast 2009 on aga hinnad hakanud jälle tõusma ning sellest ajast on aritmeetiline keskmine mediaanist märgatavalt suurem. Järelikult on müüdud korterite hulgas olnud ekstreemselt kõrge ruutmeetri hinnaga kortereid, kallimate korterite hinnad kasvasid rohkem. 2016. aastaks olid hinnad saavutanud 2007. aasta taseme.



[a] Allikas: Maa-amet, tehingute andmebaas, `http://www.maaamet.ee`

Tabelarvutusprogrammides on mediaani leidmiseks funktsioon **MEDIAN**. Kui soovime seda kasutada järjestusskaalas mõõdetud tunnuse korral, siis tuleb arvestada sellega, et kasutatav tarkvara ei saa aru sõnade tähendusest. Näiteks kui meil on järjestusskaalas mõõdetud tunnuse väärtused „väga halb", „halb", „hea" ja „väga hea", siis funktsiooni MEDIAN kasutamiseks tuleb need väärtused eelnevalt kodeerida „1", „2", „3" ja „4". Ainult siis suudab tarkvara neid väärtusi mediaani leidmiseks sorteerida.

Mediaanklass

Kuidas leida mediaani siis, kui andmed on esitatud sagedusklassidena? Siis saab leida **mediaanklassi**, kuhu mediaan kuulub. Mediaanklassi määramiseks tuleb eelnevalt leida kumulatiivsed sagedused või kumulatiivsed suhtelised sagedused.

Kumulatiivne sagedus ja kumulatiivne suhteline sagedus

k-nda klassi **kumulatiivne sagedus** on kõigi eelnevate klasside sageduste f_i summa kuni käesoleva klassini (kaasaarvatud):

$$F_k = \sum_{i=1}^{k} f_i. \tag{2.14}$$

Kumulatiivne suhteline sagedus on kumulatiivne sagedus jagatud kogumi mahuga $n = \sum f_i$:

$$W_k = \frac{F_k}{n}. \tag{2.15}$$

Kumulatiivne suhteline sagedus esitatakse tavaliselt protsentides.

Kumulatiivse sageduse praktilisel arvutamisel lähtutakse valemist (2.14) tulenevast omadusest, et k-nda klassi kumulatiivne sagedus saadakse, kui eelmise klassi kumulatiivsele sagedusele liidetakse vaadeldava klassi sagedus:

$$F_k = F_{k-1} + f_k. \tag{2.16}$$

Olgu meil 14 inimese jagunemine viide vanuserühma, s.t on antud vastavate klasside sagedused (tabel 2.3).

Tabel 2.3. Vanuseline jaotus

Vanuserühm ehk klass	Sagedus f	Kumulatiivne sagedus F	Kumulatiivne suhteline sagedus W
10–20	2	2	14,3%
20–30	2	4	28,6%
30–40	5	$2+2+5=$ $=4+5=9$	$9/14 \approx 64{,}3\%$
40–50	3	12	85,7%
50–60	2	14	100%

Vanuserühmad on ligikaudu piiratud. Tuletame meelde, et sellisel juhul klassi ülemine piir on klassi kaasa arvatud ja alumine piir välja arvatud. Näiteks inimene, kelle vanus on täpselt 40 aastat, kuulub klassi „30–40“. Tabeli kahes viimases veerus on leitud kumulatiivsed sagedused ning kumulatiivsed suhtelised sagedused. Paneme tähele, et

viimase klassi kumulatiivne sagedus võrdub kogumi mahuga. Viimase klassi kumulatiivne suhteline sagedus on 100%, mis tähendab, et kõigi kogumisse kuuluvate inimeste vanused on allpool viimase klassi ülemist piiri või võrduvad sellega.

Kuna mediaan jagab järjestatud variatsioonrea kaheks võrdseks osaks, siis mediaanklass m on kumulatiivsete sageduste abil leitav järgmiselt. Mediaanklass on see klass, mille

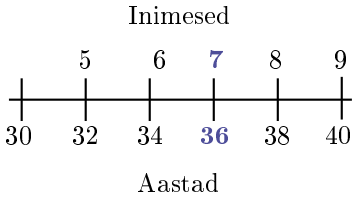
1) kumulatiivne sagedus $F_m \geq n/2$ ja
2) vahetult eelneva klassi kumulatiivne sagedus $F_{m-1} < n/2$.

Kasutada võib ka kumulatiivseid suhtelisi sagedusi. Siis on mediaanklass see, mille

1) kumulatiivne suhteline sagedus $W_m \geq 50\%$ ja
2) vahetult eelneva klassi kumulatiivne suhteline sagedus $W_{m-1} < 50\%$.

Emba-kumba lähenemist kasutades näeme, et tabelis 2.3 on mediaanklassiks vanuserühm piiridega 30–40 aastat.

Kuidas aga täpsustada mediaanvanust selle klassi piires? Mediaanklassi kuulub viis inimest, kuid me ei tea nende täpseid vanuseid. Me eeldame, et mediaanklassi kuuluvate inimeste vanused jaotuvad selle klassi piires ühtlaselt, s.t vanusevahed on võrdsed. Jagades klassi laiuse 10 aastat nende viie inimese vahel, saame vanusevaheks kaks aastat. Inimesed paigutatakse nüüd mediaanklassi piires kaheaastaste vahedega. Arvestatakse sellega, et 30-aastane inimene kuulub eelmisesse klassi (joonis 2.3).



Joonis 2.3. Eeldame, et mediaanklassi piires on vanuste jaotus ühtlane. Järjestikuste isikute vanusevahe on kaks aastat

Nüüd valitakse mediaanvanuseks selle inimese vanus, kes jagab kogumi kaheks võrdseks osaks. Kuna $n/2 = 14/2 = 7$, siis mediaanvanuseks on seitsmenda inimese vanus 36 aastat.

Selle arutluskäigu saab kokku võtta üheks **interpolatsioonivalemiks**. Interpoleerimine on funktsiooni väärtuse leidmine vahemikus, mille otspunktides on funktsiooni väärtus teada. Mediaanklassi korral on teada vanuse väärtused klassi otspunktides.

Mediaani leidmine intervallrea korral

Intervallitud variatsioonridade korral kasutatakse mediaani leidmiseks järgmist interpolatsioonivalemit:

$$Me = L + \frac{d}{f_m} \cdot \left(\frac{n}{2} - F_{m-1} \right), \tag{2.17}$$

kus on kasutatud järgmisi tähistusi:
$n = \sum f_i$ on kogumi maht;
L mediaanklassi alumine piir;
d mediaanklassi laius;
f_m mediaanklassi sagedus;
F_{m-1} mediaanklassile eelneva klassi kumulatiivne sagedus.

Analüüsime valemi (2.17) osasid eraldi:

$\frac{d}{f_m}$ on väärtuste vahemik mediaanklassi kuuluvate objektide vahel;

$\frac{n}{2} - F_{m-1}$ on mediaanile vastava objekti järjenumber mediaanklassis.

Nende kahe korrutis näitab, kui palju on mediaan suurem mediaanklassi alumisest piirist L.

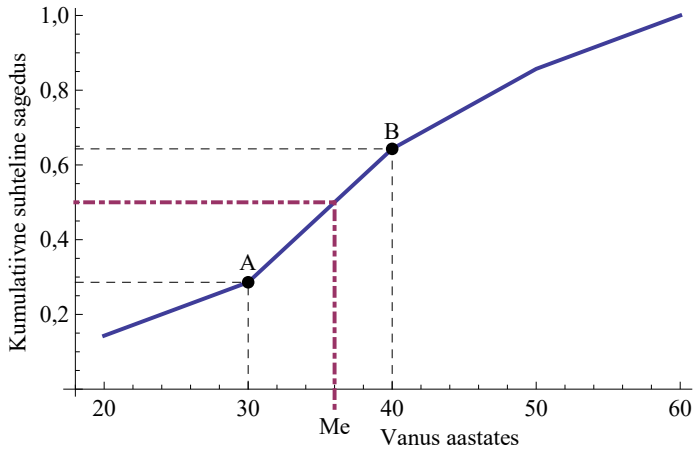
Kasutame valemit (2.17) tabelis 2.3 toodud andmete korral:

$$Me = 30 + \frac{10}{5} \left(\frac{14}{2} - 4 \right) = 36.$$

Selles arvutuses $14/2-4 = 3$ näitab, et mediaanvanus on mediaanklassi 3. inimesel. Jagatis $10/5 = 2$ tähendab, et iga mediaanklassi kuuluva isiku kohta tuleb kaks aastat. Järelikult mediaan on mediaanklassi alumisest piirist $2 \cdot 3 = 6$ aasta võrra suurem.

Mediaani ligikaudset väärtust saab hinnata ka kumulatiivse suhtelise sageduse diagrammilt. Joonisel 2.4 on tabelis 2.3 toodud kumulatiivsete suhteliste sageduste põhjal konstrueeritud diagramm. Punktid A ja B vastavad mediaanklassi alumisele ja ülemisele piirile. Sirglõik AB näitab, et kasutame lineaarset interpoleerimist: eeldame, et kumulatiivne suhteline sagedus muutub nende punktide vahel lineaarselt. Mediaan vastab kumulatiivsele suhtelisele sagedusele 0,5 (kriips-punktjoon). Diagrammilt mediaani leidmine on graafiline interpoleerimine, mis on ebatäpsem kui valemi järgi interpoleerimine.

Kui mediaanklassi saab leida nii järjestus- kui ka intervallskaalas mõõdetud tunnuse korral, siis mediaani leidmine selle klassi piires on võimalik vaid pideva intervallskaala korral. Järjestusskaala korral ei saa me arvuliselt määrata mediaanklassi laiust d ning järelikult ei saa ka kasutada interpolatsioonivalemit (2.17).



Joonis 2.4. Mediaani leidmine kumulatiivse suhtelise sageduse diagrammilt. Andmed on võetud tabelist 2.3, punktid A ja B vastavad mediaanklassi alumisele ja ülemisele piirile

Näide 2.9. Elaniku kohta tuleva elupinna mediaan Tallinnas


N02Keskmised
N2.9

Leiame näites 2.4 toodud andmete põhjal 2011. aastal Tallinnas ühe elaniku kohta tuleva elamispinna mediaani. Mediaanklassi määramiseks lisame tabelisse 2.1 veerud „Kumulatiivne sagedus“ ja „Kumulatiivne suhteline sagedus“.

Elupind elaniku kohta, m^2	Leibkondade arv ehk sagedus	Kumulatiivne sagedus	Kumulatiivne suhteline sagedus
0–12	14 698	14 698	8,0%
12–16	22 599	37 297	20,3%
16–20	22 539	59 836	32,5%
20–24	23 921	83 757	45,6%
24–32	32 360	116 117	63,2%
32–40	26 733	142 850	77,7%
40–52	22 969	165 819	90,2%
52–64	10 362	176 181	95,8%
64–84	7 677	183 858	100,0%

Kumulatiivse suhtelise sageduse põhjal leiame, et mediaanklass on 24–32 m^2 elaniku kohta. Interpolatsioonivalemi (2.17) kasutamiseks määrame vajalikud suurused:

kogumi maht $n = 183858$;
mediaanklassi alumine piir $L = 24$;
mediaanklassi laius $d = 8$;

mediaanklassi sagedus $f_m = 32360$;
mediaanklassile eelneva klassi kumulatiivne sagedus $F_{m-1} = 83757$.

Arvutus:

$$Me = 24 + \frac{8}{32360}\left(\frac{183858}{2} - 83757\right) \approx 26{,}0.$$

Vastus: 2011. aastal oli Tallinnas elaniku kohta tuleva elamispinna mediaan 26,0 m^2. Pooltes peredes oli elamispinda ühe pereliikme kohta vähem kui 26 m^2 ja pooltes peredes rohkem. Mediaan on mõnevõrra väiksem kui aritmeetiline keskmine 29,5 m^2, järelikult esineb ekstreemselt suuri väärtusi.

2.3. Kvantiilid

Mediaan jaotab järjestatud statistilise rea kaheks võrdseks osaks. Aga me võime rea jagada ka neljaks, kümneks või suuremaks arvuks võrdseks osaks.

Kvantiilid

Asendikeskmisi, mis jaotavad järjestatud statistilise rea võrdseteks osadeks, nimetatakse **kvantiilideks**.

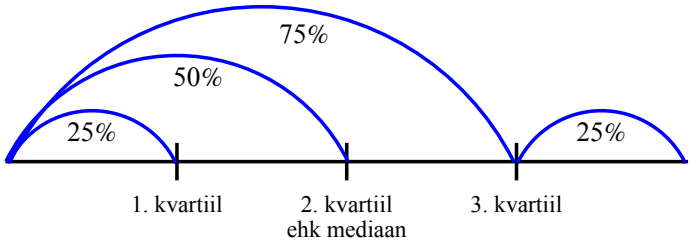
Tabel 2.4. Sagedamini kasutatavad kvantiilid

Kvantiili nimetus	Kvantiilide arv	Mitmeks osaks jaotavad	Märkused
mediaan	1	2	Mõlemale poole jääb 50% rea liikmetest.
kvartiilid	3	4	Igas neljandikus on 25% rea liikmetest.
detsiilid	9	10	Igas kümnendikus on 10% rea liikmetest.
protsentiilid	99	100	Igas sajandikus on 1% rea liikmetest

Sagedamini kasutatavad kvantiilid on toodud tabelis 2.4. Nagu näeme, on mediaan üks kvantiilidest.

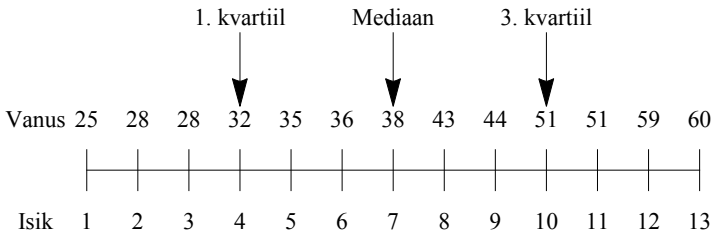
Kvartiilid

Kvartiilid jagavad järjestatud rea neljaks võrdseks osaks, igasse ossa jääb 25% rea liikmetest. Esimesest ehk alumisest kvartiilist on väiksemad 25% rea väärtustest, mediaanist väiksemad 50% ja 3. ehk ülemisest kvartiilist on väiksemad 75% rea väärtustest (vt joonis 2.5).



Joonis 2.5. Kvartiilid

Kui meil on 13 isiku vanused, siis kvartiilide leidmiseks tuleb need isikud vanuse järgi järjestada. Kuna $13/4 = 3{,}25$, siis esimene kvartiil on selle isiku vanus, kellest nooremaid on kolm. Kolmas kvartiil on selle isiku vanus, kellest vanemaid on kolm (vt joonis 2.6).



Joonis 2.6. Vanuse kvartiilid

Näide 2.10. Müügitulu kvartiilid

Oletame, et jaekaubandusettevõtte A müügitulu puhasrentaablus (puhaskasumi protsent müügitulust) oli 2012. aastal 5%. Ettevõtte juhtkond soovib teada, kas see on hea või halb. Selleks peab seda arvu millegagi võrdlema. Eesti Statistikaameti andmebaasis on tabelis EM024 „Ettevõtete asendikeskmised suhtarvud (kvartiilid, mediaan)“[a] toodud erinevate näitajate kvartiilid tegevusalade järgi. Tegevusalal „jaekaubandus, v.a mootorsõidukid ja mootorrattad“ oli müügitulu puhasrentaablus 2012. aastal järgmine (protsentides):

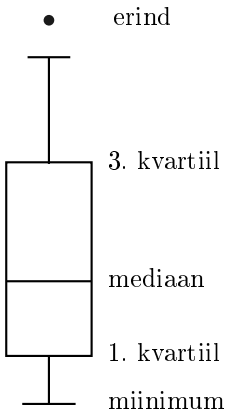
1. kvartiil	Mediaan	3. kvartiil
−0,47	1,46	8,02

On näha, et ettevõtte A müügitulu puhasrentaablus 5% on mediaani ja 3. kvartiili vahel, s.t ettevõte asub kõigi jaekaubandusettevõtete seas kolmandas neljandikus. Kõige halvem ei ole, kuid saab paremini.

[a] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel EM024: ettevõtete asendikeskmised suhtarvud (kvartiilid, mediaan).

Mediaani ja kvartiilide esitamiseks kasutatakse tihti **karpdiagrammi** (*boxplot*). Karbi alumine ja ülemine serv on määratud vastavalt esimese ja kolmanda kvartiiliga, seega karp sisaldab 50% kõikidest väärtustest (vt joonis 2.7).

Karpdiagramm



Joonis 2.7. Karpdiagramm

Karbi kõrgus on 3. ja 1. kvartiili vahe ehk kvartiilhaare. Karbile lisatakse vertikaaljooned, mida nimetatakse „vurrudeks“ ja seetõttu nimetatakse seda diagrammi mõnikord ka karp-vurrud diagrammiks (*box-whisker plot*). Vurrude otsad tähistavad

- miinimumi ja maksimumi või
- väärtust, mis ei ole 1. kvartiilist väiksem (3. kvartiilist suurem) kui 1,5-kordne kvartiilhaare.

Viimasel juhul esitatakse kaugemal asuvad väärtused eraldi punktidena ja neid nimetatakse **erinditeks** (*outliers*). Karpdiagrammi pakkus esmakordselt välja 1977. aastal USA matemaatik John Tukey.

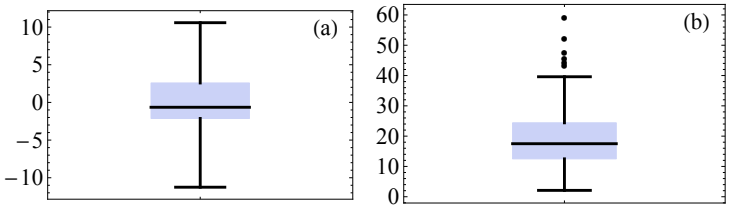
Näide 2.11. SKP kasvumäär ja imikute suremus

Joonisel 2.8 on toodud SKP kasvumäär aastas (protsentides) ja imikute suremus (imikute surmade arv 1000 sünni kohta) 50 Aasia, Ladina-Ameerika ja Vaikse ookeani riigis aastal 2009[a]. SKP kasvumäära korral erindid puuduvad ning vurrude otsad

näitavad miinimumi ja maksimumi. Imikute suremus on aga kuues riigis väga suur. Ülemine vunts ulatub väärtuseni 39,6. See on saadud järgmiselt:

- 3. kvartiili ja 1. kvartiili vahe on $24{,}28 - 12{,}55 = 11{,}73$;
- selle vahe 1,5-kordne väärtus $1{,}5 \cdot 11{,}73 \approx 17{,}59$;
- viimane väärtus liidetakse 3. kvartiilile ja saadakse $24{,}28 + 17{,}59 = 41{,}87$;
- vunts ulatub väärtuseni, mis pole kaugemal kui 41,87.

Väärtusest 41,87 on imikute suremus suurem kuues riigis: Haiiti (59,3), Birma (52,2), Paapua Uus-Guinea (47,7), Kambodža (45,6), Laos (44,3) ja Boliivia (43,4). Need on erindid.



Joonis 2.8. (a) SKP kasvumäär aastas (protsentides) ja (b) imikute suremus 50 Aasia, Ladina-Ameerika ja Vaikse ookeani riigis aastal 2009

[a] Allikas: *World DataBank* `http://databank.worldbank.org`

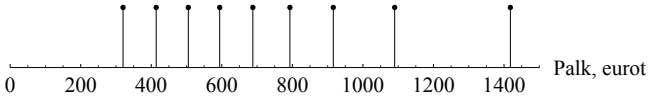
Detsiilid jagavad järjestatud variatsioonrea kümneks võrdseks osaks. Esimesest detsiilist väiksemad on 10% rea väärtustest, teisest detsiilist väiksemad 20% rea väärtustest jne. Viimasest, üheksandast detsiilist, on väiksemad 90% ja suuremad 10% kõikidest väärtustest. *Detsiilid*

Näide 2.12. Töötasu detsiilid

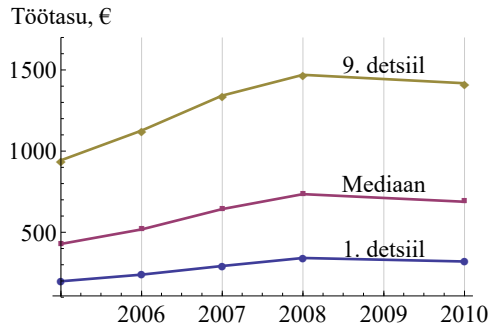
Tabelis on toodud täistööajaga töötajate brutokuutöötasu detsiilid Eestis 2010. aasta oktoobris[a]. Andmed on eurodes. Näeme, et 320 eurot oli palk, millest väiksemat said 10% töötajatest ja suuremat 90% töötajatest. Mediaanpalk vastab viiendale detsiilile ja oli 688 eurot. 10% kõikidest töötajatest said kuus rohkem kui 1418 eurot.

D1	D2	D3	D4	Mediaan D5	D6	D7	D8	D9
320	414	505	594	688	793	916	1090	1418

Tabeli põhjal on koostatud joonis, kus detsiilide asukohad palgateljel on kujutatud vertikaalsete lõikudena. Igasse vahemikku kuulub ühesugune arv töötajaid. Näeme, et suuremate palkade korral on ka erinevused palkades suuremad, sest sama arv palgasaajaid jaguneb laiali suuremas vahemikus.



Joonisel 2.9 on toodud töötasu esimese, viienda (mediaan) ja üheksanda detsiili muutumine aastatel 2005–2008. Näeme, et palkade suurenedes suurenevad ka palgaerinevused.



Joonis 2.9. Brutokuutöötasu detsiilid Eestis 2005–2010

[a] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel PA622: täistööajaga töötajate brutokuutöötasu, oktoober. Mehed ja naised, kõik tegevusalad kokku.

Neid statistilise rea osasid, mis on kvantiilide vahel, nimetatakse **kvantiilvahemikeks**. Kvartiilide korral tekib neli kvartiilvahemikku, detsiilide korral 10 detsiilvahemikku.

Klassifitseeritud tunnuse korral saab kvantiile leida samasuguse interpolatsioonivalemi abil nagu mediaani puhul. Tuleb leida otsitava kvantiili järjenumber ning määrata, millisesse klassi see kuulub. Seejärel rakendatakse interpolatsioonivalemit. Kuna sagedamini leitakse kvartiile, siis üldistame valemit (2.17) kvartiilide jaoks.

Kvartiili leidmine intervallitud rea korral

Intervallitud variatsioonridade korral kasutatakse kvartiilide leidmiseks järgmist valemit:

$$Q_r = L + \frac{d}{f_Q} \cdot (j_r - F_{Q-1}), \tag{2.18}$$

kus on kasutatud järgmisi tähistusi:
r kvartiili järk 1, 2 või 3;
j_r otsitava kvartiili järjenumber reas; $j_r = r\frac{n}{4}$, kus $n = \sum f_i$ on kogumi maht;
L kvartiilklassi alumine piir;
d kvartiilklassi laius;
f_Q kvartiilklassi sagedus;
F_{Q-1} kvartiilklassile eelneva klassi kumulatiivne sagedus.

Protsentiilid

Kui me soovime variatsioonrea jagada rohkem kui 10 osaks, kasutatakse **protsentiile**. k-protsentiil ehk k järku protsentiil on tunnuse selline väärtus, millest väiksemate väärtuste osakaal on $k\%$. Näiteks kümnes protsentiil on esimene detsiil, 25. protsentiil on esimene kvartiil, 75. protsentiil on kolmas kvartiil.

Protsentiili järgu võib anda ka kümnendmurruna, mis on 0 ja 1 vahel (0,1, 0,25, 0,75). Nii võib protsentiile kasutada variatsioonrea jagamiseks veel väiksemateks osadeks. Siis vastab protsentiili järgule 0,01 väärtus, millest väiksemad on 1% kõikidest väärtustest. Järgule 0,001 vastab väärtus, millest väiksemad on 0,1% kõikidest väärtustest.

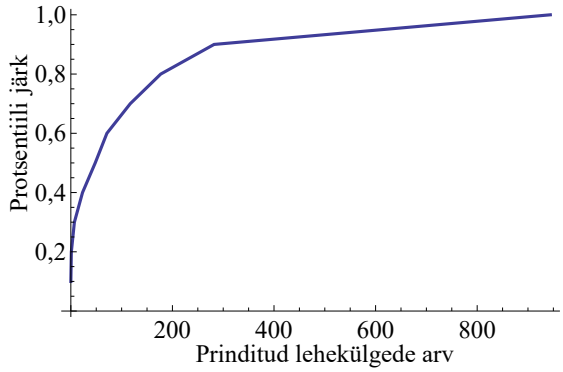
Näide 2.13. Printimise ressursi jaotus

N02Keskmised
N2.13

1999. aasta kevadsemestril juurutati Audentese Ülikoolis uus printimise tarkvara, mis võimaldas registreerida iga üliõpilase poolt prinditud lehekekülgede arvu ja panna peale limiite. Seni võisid üliõpilased arvutiklassis printida nii palju, kui soovisid. Esialgu limiite ei kehtestatud ja semestri jooksul koguti lihtsalt andmeid printimise mahtude kohta. Tabelis on toodud prinditud lehekülgede arvu protsentiilid. On näha, et 30% üliõpilastest printisid kuni 7 lk semestris, 5% üliõpilastest oli aga prinditud lehekülgede arv 461 ja 944 vahel.

Protsentiili järk	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%	100%
Protsentiil, lk	0	1	7	23	49	71	117	177	282	461	944

Protsentiilid võib esitada ka diagrammil.



Funktsioon **QUARTILE.INC** leiab kvartiilid tabelarvutuses. Lisaks kvartiilidele leiab QUARTILE.INC ka miinimum- ja maksimumväärtuse.

Funktsiooni argument *quart*	Mille funktsioon väljastab
0	miinimum
1	1. kvartiil
2	2. kvartiil ehk mediaan
3	3. kvartiil
4	maksimum

Funktsioon **QUARTILE.EXC** erineb selle poolest, et väljastatud arv ise ei kuulu vahemikku, mis on määratud 25, 50 või 75 protsendiga.

Protsentiilide leidmiseks on tabelarvutuses funktsioon **PERCENTILE.INC**. Funktsiooni argument k on protsentiili järk, mis peab olema 0 ja 1 vahel.

10% arvudest on $\leq x$, siis x =PERCENTILE.INC(arvud;0,1).
50% arvudest on $\leq x$, siis x =PERCENTILE.INC(arvud;0,5).
1% arvudest on $\leq x$, siis x =PERCENTILE.INC(arvud;0,01).

Seda funktsiooni kasutasime protsentiilide leidmiseks näites 2.13. On olemas ka funktsioon **PERCENTILE.EXC**, mille abil leitakse väärtus x nii, et k osa arvudest on sellest rangelt väiksemad.

10% arvudest on $< x$, siis x =PERCENTILE.EXC(arvud;0,1).
50% arvudest on $< x$, siis x =PERCENTILE.EXC(arvud;0,5).
1% arvudest on $< x$, siis x =PERCENTILE.EXC(arvud;0,01).

Tabelarvutuses on võimalik lahendada ka pöördülesanne: anda arvukogumist ette mõni arv ja leida, milline protsentiili järk sellele vastab. Näiteks kui soovime leida, kui suur osa väärtustest on väiksemvõrdne mingist arvust. Siis kasutatakse tabelarvutusfunktsiooni **PERCENTRANK.INC**.

Kui suur osa on ≤ 50 =PERCENTRANK.INC(arvud;50).
Kui suur osa on > 50 =1−PERCENTRANK.INC(arvud;50).

Kui tahame leida, kui suur osa väärtustest on väiksem mingist arvust, kasutame funktsiooni **PERCENTRANK.EXC**.

Kui suur osa on < 50 =PERCENTRANK.EXC(arvud;50).
Kui suur osa on ≥ 50 =1−PERCENTRANK.EXC(arvud;50).

Näide 2.14. Internetikasutajate arv 100 elaniku kohta erinevates riikides


N02Keskmised
N2.14

Kui suur on internetikasutajate arv 100 elaniku kohta erinevates riikides? Kasutame 2012. aasta andmeid, mis Maailmapanga veebilehel[a] olevas andmebaasis on 201 riigi kohta. Esitame siin kolm kõige suurema näitajaga riiki.

Riik	Internetikasutajaid 100 el kohta
Island	96,2
Norra	94,6
Rootsi	93,2
...	...

Leiame kvartiilid kõigi 201 riigi näitajate põhjal, kasutades funktsiooni QUARTILE.INC.

Kvartiili nr	Kvartiil
1	12,6
2	38,2
3	62,3

Näeme, et neljandikus maailma riikidest on internetikasutajaid 100 elaniku kohta vähem kui 12,6 ja pooltes riikides vähem kui 38,2. Eestis oli 2012. aastal 78,4 internetikasutajat 100 elaniku kohta, järelikult Eesti asub viimases neljandikus.
Eesti asukoha täpsemaks määramiseks kasutame tabelarvutuses funktsiooni PERCENTRANK.INC. See väljastab väärtuse 0,87, järelikult 13% riikidest on internetikasutajate arv 100 elaniku kohta suurem kui Eestis.

[a] The World Bank `http://www.worldbank.org/`

2.4. Mood

Sõna mood viib tavaliselt esimesena mõtte riietusele. See riietus, värv, stiil, mida võib kohata kõige sagedamini, on moes. Ka statistikas on mood seotud esinemissagedusega.

Mood

Mood on variatsioonreas kõige sagedamini esinev väärtus. Moodi kasutatakse siis, kui kogumit soovitakse iseloomustada kõige **tüüpilisema** väärtusega.

Mood on ainuke statistiline keskmine, mida saab kasutada nimiskaala korral. Jooniselt 1.8, kus on toodud vastajate jaotus sotsiaalmajandusliku seisundi järgi (nimiskaalas), on kerge määrata moodi: see on „töötav".

Ankeetküsitlustes on vastusevariandid sageli antud kas nimi- või järjestusskaalas. Küsitlustulemustest kokkuvõtete tegemisel kasutatakse sellisel juhul tihti moodi: milline vastusevariant valiti kõige sagedamini.

Näide 2.15. Sotsiaaluuringus osalenute haridustase ja ametiala

2013. aasta Eesti sotsiaaluuringus osales 15 503 isikut. Muuhulgas paluti küsitluses märkida ka oma haridustase (järjestusskaalas) ning praegune või viimane ametiala (nimiskaalas). Haridustaseme jättis märkimata 2723 isikut ning ametiala 4142 isikut. (*Eesti sotsiaaluuring* 2013)

Haridustase	Vastanuid
I taseme haridus: alghariduseta, alghariduseega, põhihariduseega, baashariduseta kutseharidus	2848
II taseme haridus: keskharidus, kutseõpe põhihariduse baasil	6319
III taseme haridus: kutseõpe keskhariduse baasil, kõrgharidus, magister, doktor	3163

Praegune või viimane ametiala	
Seadusandjad, kõrgemad ametnikud ja juhid, tippspetsialistid	2495
Keskastme spetsialistid ja tehnikud, ametnikud	1619
Teenindus- ja müügitöötajad	1518
Põllumajanduse ja kalanduse oskustöölised, oskus- ja käsitöölised	2073
Seadme- ja masinaoperaatorid	1717
Lihttöölised, relvajõud	1489

Tabelist näeme, et haridustaseme mood on „II taseme haridus: keskharidus, kutseõpe põhihariduse baasil" ning ametiala mood „Seadusandjad, kõrgemad ametnikud ja juhid, tippspetsialistid".

Intervallskaala korral on mood kõige kergemini määratav statistiline keskmine. Näites 2.3 toodud tubade arvu jaotusest saab ilma igasuguste arvutusteta kohe määrata kõige sagedamini esinenud tubade arvu 2.

Näide 2.16. Keskmine bensiinihind

Linna kuues tanklas oli bensiini 95 hind järgmine (€/l):

Tankla 1	Tankla 2	Tankla 3	Tankla 4	Tankla 5	Tankla 6
1,149	1,123	1,149	1,139	1,105	1,149

Milline on keskmine bensiini hind? Aritmeetiline keskmine 1,136 €/l ja mediaan 1,144 €/l antud juhul ei sobi, sest kumbagi hinda üheski tanklas ei küsitud. Antud kontekstis on keskmine hind 1,149 €/l, mis on kõige sagedamini esinenud hind.

Moodi kasutatakse keskmiste hindade leidmisel turustatistikas, rõiva- ja jalatsinumbrite keskmise suuruse kindlaksmääramisel. Ka nafta ja kulla hind maailmaturul leitakse tavaliselt moodina. Nimetatakse seda ka modaalhinnaks (*modal price*). Hinnamuutuste analüüsimisel võib baashinnaks võtta modaalhinna ja uurida, kui sageli on hinnad sellest kõrgemal või madalamal (Kehoe ja Midrigan, 2008).

Moodi **omadusi**:

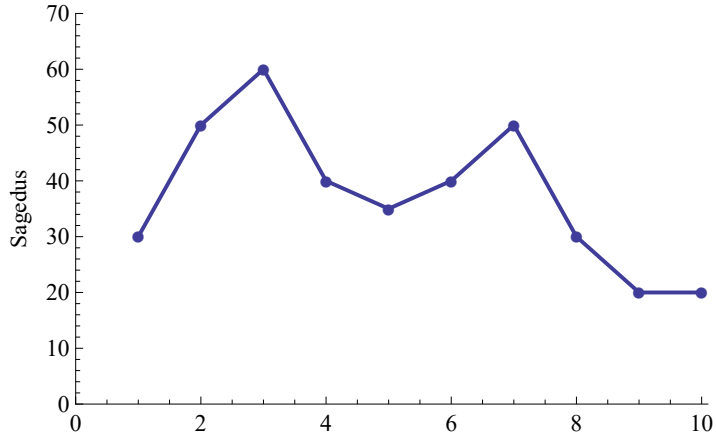
Moodi omadusi

1) moodi saab kasutada nimi-, järjestus- ja intervallskaala korral;
2) mood on ainuke statistiline keskmine, mille väärtus arvukogumis alati esineb;
3) mõnedel andmekogumitel võib mood puududa: kõik väärtused esinevad ühepalju kordi;
4) mõnedel andmekogumitel võib olla mitu moodi: sagedusdiagrammil esinevad üksteisest eraldatud tipud ehk lokaalsed maksimumid nagu joonisel 2.10. Sellisel juhul on tegemist **multimodaalse kogumiga**. Nende tippude kõrgused võivad olla erinevad. Sellisel juhul on tõenäoliselt tegemist mittehomogeense kogumiga, mille võib mingi tausttunnuse alusel jagada mitmeks unimodaalseks (ühe moodiga) osakogumiks.

Multimodaalse kogumi korral võib eristada peamoodi (*major mode*), mis on kõige sagedamini esinev väärtus, ja kõrvalmoode (*minor modes*). Joonisel 2.10 on peamood 3 ja üks kõrvalmood 7.

Sõltuvalt moodi asukohast pakkus Norra sotsioloog ja matemaatik Johan Galtung jaotuste klassifitseerimiseks välja süsteemi AJUS, mida hiljem on veidi modifitseeritud (Galtung, 1967):

A: unimodaalne jaotus, mood on keskel;

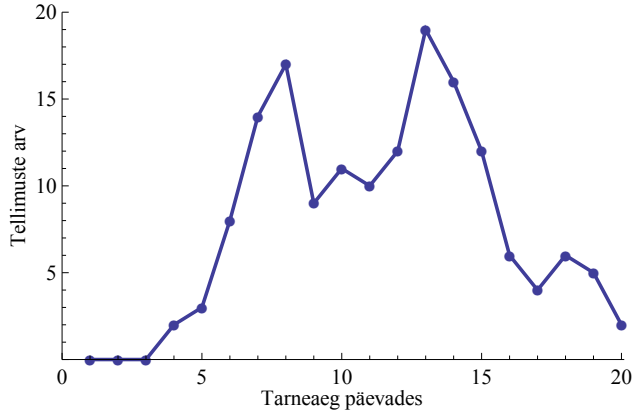


Joonis 2.10. Multimodaalne kogum, millel on kaks moodi 3 ja 7

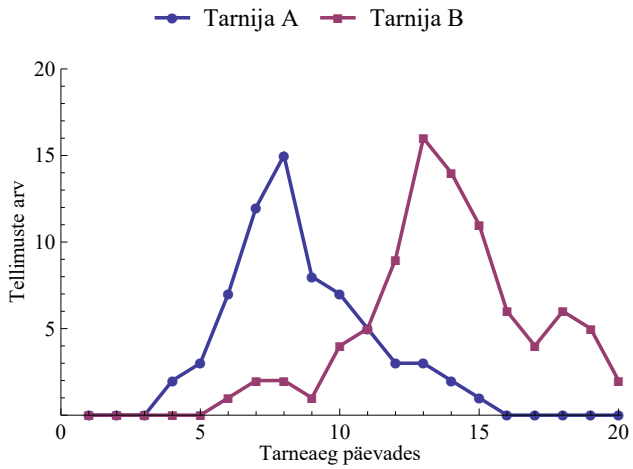
J: unimodaalne jaotus, mood on paremas ääres;
L: unimodaalne jaotus, mood on vasakus ääres;
U: bimodaalne jaotus, moodid on äärtes;
S: multimodaalne jaotus;
F: mood puudub.

Näide 2.17. Tarneaegade jaotus

Ettevõttes analüüsiti kolme kuu jooksul väljastatud 156 tellimuse tarneaega. Sagedustabeli põhjal konstrueeritud diagrammilt on näha, et esineb kaks tippu: 8 päeva (17 tellimust) ja 13 päeva (19 tellimust).

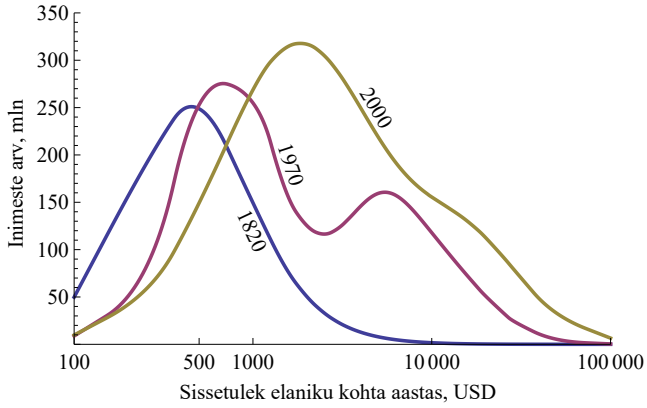


Ilmselt on tegemist kahe erineva tarnijaga. Ühe tarnija korral on tüüpilisem tarneaeg (mood) 8 päeva ja teisel tarnijal 13 päeva. Kui esitame tarneaegade jaotuse tarnijate kaupa, saame kaks unimodaalset kogumit.



Näide 2.18. Sissetulekute jaotus maailmas

OECD kirjastatud kogumikus „How Was Life?: Global Well-being since 1820“ (Zanden ja Baten, 2014) on vaatluse all maailmamajanduse areng alates aastast 1820. Muuhulgas analüüsiti sissetulekute jaotust ja selle muutust. 1820. aastal oli jaotus unimodaalne ning kõige rohkem oli maailmas inimesi sissetulekuga 500 USD aastas. Maailm oli praegusega võrreldes vaene. 1970-ndatel aastatel jagunes maailm kaheks: rikkamad ja vaesemad riigid (bimodaalne jaotus). Selle sajandi alguses on vaesemad riigid rikastele järele jõudnud ning jaotus on jälle unimodaalne.

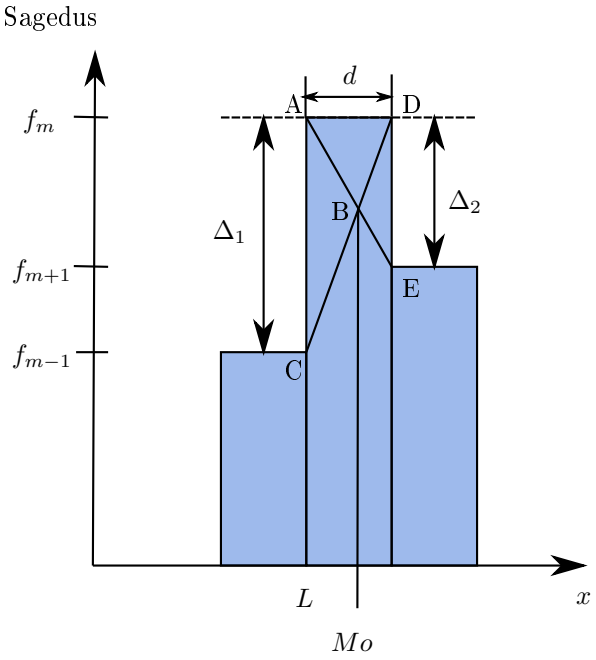


Tabelarvutuses on moodi leidmiseks funktsioon **MODE.SNGL**. See leiab moodi ainult arvudest. Kui tegemist on nimi- või järjestusskaalaga, tuleb väärtused numbriliselt kodeerida. Multimodaalse kogu-

mi korral leiab järjestikused moodid funktsioon **MODE.MULT**. Kuna see funktsioon väljastab mitu väärtust, siis on tegemist massiivifunktsiooniga ning selle sisestamine on analoogne funktsiooni FREQUENCY sisestamisega.

Moodklass

Suure arvu väärtuste korral grupeeritakse need intervallidesse ja siis leitakse moodintervall ehk **moodklass**, mis on kõige suurema sagedusega klass. Moodi täpsema väärtuse leidmiseks moodklassis kasutatakse interpolatsioonivalemit. Valemi saamiseks lähendatakse histogrammi pideva joonega ja leitakse selle joone maksimumkoht.



Joonis 2.11. Moodi leidmine moodklassis

Mood Mo jaotab moodklassi kaheks lõiguks pikkustega $Mo-L$ ja $d-(Mo-L)$ (vt joonis 2.11). Lõik $Mo-L$ on kolmnurga ABC kõrgus ja lõik $d-(Mo-L)$ on kolmnurga DBE kõrgus. Kuna kolmnurgad ABC ja DBE on sarnased, siis nende kõrguste suhe on võrdne kolmnurkade aluste ehk lõikude $AC=\Delta_1$ ja $DE=\Delta_2$ suhtega:

$$\frac{Mo-L}{d-(Mo-L)}=\frac{\Delta_1}{\Delta_2}.$$

Avaldame sellest võrdusest moodi Mo:

$$\begin{aligned}\Delta_2(Mo-L)&=\Delta_1(d-Mo+L),\\ \Delta_2 Mo-\Delta_2 L&=\Delta_1 d-\Delta_1 Mo+\Delta_1 L,\\ Mo(\Delta_1+\Delta_2)&=L(\Delta_1+\Delta_2)+d\Delta_1,\end{aligned}$$

$$Mo = L + d\frac{\Delta_1}{\Delta_1 + \Delta_2}.$$

Intervallitud variatsioonrea korral leitakse mood valemist

$$Mo = L + d\frac{\Delta_1}{\Delta_1 + \Delta_2}, \tag{2.19}$$

kus on kasutatud järgmisi tähistusi (vt ka joonis 2.11):
L moodklassi alumine piir;
d moodklassi laius;
$\Delta_1 = f_m - f_{m-1}$;
$\Delta_2 = f_m - f_{m+1}$;
f_m moodklassi sagedus;
f_{m-1} moodklassile eelneva klassi sagedus;
f_{m+1} moodklassile järgneva klassi sagedus.

Moodi leidmine intervallitud rea korral

Kui moodklassi kõrval olevate klasside sagedused on ühesugused $f_{m-1} = f_{m+1}$, siis ka $\Delta_1 = \Delta_2$ ning valemist (2.19)

$$Mo = L + d \cdot \frac{1}{2},$$

mis tähendab, et mood asub moodklassi keskel. Kui ühel pool asuva klassi sagedus on suurem, siis on mood nihkunud selle klassi poole. Kui moodklassiks on esimene klass, siis sellele eelnev klass puudub ja järelikult selle sagedus $f_{m-1} = 0$. Samamoodi, kui moodklassiks on viimane klass, siis $f_{m+1} = 0$.

Näide 2.19. Kõige tüüpilisem elamispinna suurus elaniku kohta Tallinnas

Näites 2.4 leiti sagedustabeli põhjal keskmise elamispinna suurus inimese kohta kaalutud aritmeetilise keskmisena ja näites 2.9 leiti mediaan. Nüüd leiame samade andmete põhjal (tabel 2.5) kõige tüüpilisema elamispinna suuruse, s.o moodi.
Moodklassiks on kõige suurema sagedusega klass ja selleks on klass, mille korral elamispinna suurus jääb vahemikku 24 kuni 32 m^2. Valemi (2.19) kasutamiseks paneme kirja vajalikud andmed:
moodklassi alumine piir $L = 24$:
moodklassi laius $d = 32 - 24 = 8$;
moodklassi sagedus $f_m = 32360$;

moodklassile eelneva klassi sagedus $f_{m-1} = 23921$;
moodklassile järgneva klassi sagedus $f_{m+1} = 26733$.

Tabel 2.5. Elamispind ühe elaniku kohta

Elamispinna suurus, m^2	Leibkondade arv ehk sagedus f_i
0–12	14 698
12–16	22 599
16–20	22 539
20–24	23 921
24–32	32 360
32–40	26 733
40–52	22 969
52–64	10 362
64–84	7 677

Moodi arvutus valemi (2.19) järgi:

$$Mo = 24 + 8 \cdot \frac{32360 - 23921}{(32360 - 23921) + (32360 - 26733)} \approx 28{,}8.$$

Vastus: Tallinnas oli kõige tüüpilisem elamispinna suurus elaniku kohta ehk mood 2011. aastal 28,8 m^2. Eelnevalt leitud aritmeetiline keskmine oli 29,5 m^2 ja mediaan 26,0 m^2.

2.5. Harmooniline keskmine

Harmoonia tuleb kreeka keelest ja tähendab kooskõla. Kitarrikeelel tekivad harmoonilised toonid siis, kui seisulainete poollainepikkused vastavad harmoonilisele reale $1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n}$. Harmooniline rida on järelikult arvude $1, 2, 3, \ldots$ pöördväärtuste rida[1]. Ka harmooniline keskmine on seotud pöördväärtustega.

Näide 2.20. Bensiiniliitri keskmine hind

Autojuht ostis kuu aja jooksul kolm korda bensiini, iga kord 50 euro eest. Esimene kord oli liitri hind 1,149 eurot, teine kord

[1] Matemaatikas on harmooniline rida $\sum_{i=1}^{\infty} \frac{1}{k^{\alpha}}$, kus α on reaalarv.

1,139 eurot ja viimasel korral 1,105 eurot. Milliseks kujunes bensiiniliitri keskmine hind?

$$\text{keskmine hind} = \frac{\text{raha kokku}}{\text{kogus kokku.}} \tag{2.20}$$

Kuna koguseid pole otseselt antud, tuleb need arvutada. Esimene kord oli koguseks 50/1,149 liitrit, teine kord 50/1,139 liitrit ja kolmas kord 50/1,105 liitrit.

$$\begin{aligned} &\frac{50+50+50}{\frac{50}{1{,}149}+\frac{50}{1{,}139}+\frac{50}{1{,}105}} = \\ &= \frac{3}{\frac{1}{1{,}149}+\frac{1}{1{,}139}+\frac{1}{1{,}105}} \approx 1{,}13. \end{aligned} \tag{2.21}$$

Vastus: bensiiniliitri keskmiseks hinnaks kujunes 1,13 €/l .

Valemis (2.21) on nimetajas hindade pöördväärtused. Sellist keskmist nimetatakse harmooniliseks keskmiseks. Paneme tähele, et kui võtame valemist (2.21) pöördväärtuse, siis saame pöördväärtuste aritmeetilise keskmise: lugejas on pöördväärtuste summa ja nimetajas pöördväärtuste arv:

$$\frac{\frac{1}{1{,}149}+\frac{1}{1{,}139}+\frac{1}{1{,}105}}{3} \approx \frac{1}{1{,}13}.$$

Harmooniline keskmine on pöördväärtuste aritmeetilise keskmise pöördväärtus:

Harmooniline keskmine

$$\bar{x}_{harm} = \frac{n}{\sum \frac{1}{x_i}}. \tag{2.22}$$

Millal tuleb kasutada harmoonilist keskmist? See sõltub sellest, millised andmed on antud.

Näide 2.21. Bensiiniliitri keskmine hind II

Autojuht ostis kuu aja jooksul kolm korda bensiini. Esimene kord 43,5 liitrit hinnaga 1,149 €/l, teine kord 45,2 liitrit hinnaga

1,139 €/l ja kolmas kord 43,9 liitrit hinnaga 1,105 €/l. Milliseks kujunes bensiiniliitri keskmine hind?
Keskmise hinna jaoks kasutame jälle seost (2.20). Seekord on kogused antud, arvutada tuleb bensiini ostmiseks kulutatud rahasumma.

$$\frac{43{,}5 \cdot 1{,}149 + 45{,}2 \cdot 1{,}139 + 43{,}9 \cdot 1{,}105}{43{,}5 + 45{,}2 + 43{,}9} \approx 1{,}13. \qquad (2.23)$$

Vastus: bensiiniliitri keskmiseks hinnaks kujunes 1,13 €/l. Nüüd tuli kasutada kaalutud aritmeetilist keskmist.

Näites 2.21 olid antud hinnad ja kogused ning kasutada tuli kaalutud aritmeetilist keskmist. Järelikult see, kumba keskmist keskmise hinna leidmisel kasutada, sõltub sellest, kas meil on antud kogused (aritmeetiline keskmine) või summad (harmooniline keskmine).

Näites 2.20 osteti bensiini iga kord ühe ja sama summa 50 euro eest. See taandus valemis (2.21) ära ja lõpptulemusena saime arvutada lihtsa harmoonilise keskmise valemi (2.22) põhjal. Tihti see nii ei ole ja kasutada tuleb kaalutud harmoonilist keskmist.

Kaalutud harmooniline keskmine

Kaalutud harmoonilise keskmise valem on

$$\bar{x}_{harm} = \frac{\sum f_i}{\sum \frac{1}{x_i} f_i}, \qquad (2.24)$$

kus f_i on väärtuse x_i kaal.

Näide 2.22. Keskmine tootlikkus

Tööjõu tootlikkus naturaalühikutes näitab, mitu ühikut keskmiselt üks inimene toodab mingi aja (tunni, päeva, kuu) jooksul. Olgu ettevõttes kolm tootmistsehhi, kus keskmine tootlikkus päevas on vastavalt 50 tk, 45 tk ja 40 tk ühe töötaja kohta. Esimeses tsehhis toodeti kokku 600 toodet, teises 675 ja kolmandas 400 toodet. Kui suur on ettevõttes keskmine tööjõu tootlikkus?

$$\text{keskmine tootlikkus} = \frac{\text{toodete arv kokku}}{\text{inimeste arv kokku}}. \qquad (2.25)$$

Teades tootlikkust igas tsehhis ja toodete arvu, saame leida inimeste arvu tsehhides. Esimeses tsehhis on 600/50 töötajat, teises 675/45 töötajat ja kolmandas 400/40 töötajat. Keskmine toot-

likkus

$$\frac{600+675+400}{\frac{600}{50}+\frac{675}{45}+\frac{400}{40}} \approx 45{,}27.$$

Vastus: keskmine tootlikkus on 45,27 toodet inimese kohta.

Näites 2.22 tuli kasutada kaalutud harmoonilise keskmise valemit, kus kaaluks oli toodete arv.

Üldistades toodud näiteid, võib anda reegli valiku tegemiseks aritmeetilise ja harmoonilise keskmise vahel. Olgu meil vaja leida mingi keskmine, mis üldiselt avaldub kahe näitaja summa suhtena.

- Kui on antud lugejas oleva näitaja väärtused, kasutame harmoonilist keskmist (näited 2.16 ja 2.22).
- Kui on antud nimetajas oleva näitaja väärtused, kasutame aritmeetilist keskmist (näide 2.21).

Näiteks keskmise hinna leidmisel võib lähtuda järgmisest skeemist:

- kui meil on info summade kohta:
  - üksikutele hindadele vastavad summad on ühesugused — lihtne harmooniline keskmine;
  - üksikutele hindadele vastavad summad on erinevad — kaalutud harmooniline keskmine;
- kui meil on info koguste kohta:
  - üksikutele hindadele vastavad kogused on ühesugused — lihtne aritmeetiline keskmine;
  - üksikutele hindadele vastavad kogused on erinevad — kaalutud aritmeetiline keskmine.

Millal harmooniline, millal aritmeetiline keskmine

Kahtluse korral on mõistlik panna kirja vastava näitaja arvutamise reegel, nagu näiteks valemid (2.20) ja (2.25) ning lähtuda sellest.

Tabelarvutuses on lihtsa harmoonilise keskmise leidmiseks funktsioon **HARMEAN**. Kaalutud harmoonilise keskmise leidmiseks funktsioon puudub.

2.6. Geomeetriline keskmine

Kui me tahame ristkülikule, mille küljed on a ja b, seada vastavusse sama suure pindalaga võrdsete külgedega ristkülikut ehk ruutu, siis selle ruudu külje pikkuse x saame seosest

$$x^2 = ab \Rightarrow x = \sqrt{ab}.$$

Suurus $\sqrt{ab}$ on arvude a ja b geomeetriline keskmine. Analoogselt vastab külgedega a, b ja c risttahukale sama suure ruumalaga kuup, nii et

kuubi külje pikkus x on risttahuka külgede geomeetriline keskmine:

$$x^3 = abc \Rightarrow x = \sqrt[3]{abc}.$$

Millal tuleb veel kasutada geomeetrilist keskmist?

Näide 2.23. Kütusehinna tõus

2015. aasta 13. veebruari Postimehe majandusrubriigis ilmus artikkel „Mootorikütuste hind tõusis rekordkiirusel“[a]. See algas järgmise lõiguga:
„Eesti suuremad kütusefirmad tõstsid esmaspäeval mootorikütuste jaehindu kahe sendi võrra liitri pealt, mis oli juba kolmas hinnatõus kaheksa päeva jooksul. See tähendab, et kaheksa päevaga kallines mootorikütus kaheksa sendi võrra ehk kaheksa protsenti, mis teeb keskmiseks tõusuks üks sent ehk üks protsent päevas.“
Analüüsime väidet, et kui kaheksa päevaga tõusis hind kaheksa protsenti, siis keskmine tõus oli 1% päevas. Olgu alghind p_0 ja eeldame, et iga päev tõuseb hind 1%. Hinnatõus esimesel päeval:

$$p_1 = p_0 + 0{,}01p_0 = 1{,}01p_0.$$

Hinnatõus teisel päeval:

$$p_2 = p_1 + 0{,}01p_1 = 1{,}01p_1 = 1{,}01 \cdot 1{,}01p_0 = 1{,}01^2p_0.$$

Hinnatõus kolmandal päeval:

$$p_3 = p_2 + 0{,}01p_2 = 1{,}01p_2 = 1{,}01 \cdot 1{,}01^2p_0 = 1{,}01^3p_0.$$

Niimoodi jätkates saame, et kaheksa päeva pärast

$$p_8 = 1{,}01^8p_0 \approx 1{,}0829p_0.$$

Näeme, et sellisel juhul, kui keskmine hinnatõus oleks 1% päevas, tõuseks hind kaheksa päevaga 8,29%, aga mitte 8%, nagu artiklis kirjas.
Kui suur aga oli keskmine hinnatõus päevas? Selleks paneme kirja võrrandi, kus lõpphind on alghinnast 8% võrra suurem, ja leiame tundmatu x, millega tuleb iga päeva kohta alghinda korrutada:

$$\begin{aligned} x^8p_0 &= 1{,}08p_0 \\ x^8 &= 1{,}08 \\ x &= \sqrt[8]{1{,}08} \approx 1{,}00967. \end{aligned}$$

Kontroll:

$$1{,}00967^8 \approx 1{,}0080.$$

Järelikult keskmine hinnatõus oli 0,967% päevas. Sellisel juhul tõuseb hind kaheksa päevaga 8%. Õige oleks olnud kirjutada ka ligikaudu 1% päevas.

[a] http://majandus24.postimees.ee/3089793/mootorikutuste-hind-tousis-rekordkiirusel

Näide 2.24. Tarbijahinnaindeksi muutus

2012. aastal tõusis Eesti tarbijahinnaindeks 3,9%, 2013. aastal 2,8% ja 2014. aastal langes 0,1%. Mitu protsenti muutus tarbijahinnaindeks keskmiselt aastas?
Protsentuaalne kasv $r\%$ tähendab korrutamist kordajaga $1 + 0{,}01r$. Seda nimetatakse kasvutempoks. Paneme tabelisse kirja protsentuaalsed kasvud ja vastavad kasvutempod.

Aasta	Protsentuaalne kasv	Kasvutempo
2012	3,9%	1,039
2013	2,8%	1,028
2014	−0,1%	0,999

Olgu tarbijahinnaindeksi algväärtus THI_0. Siis

2012. aasta lõpuks $THI_1 = 1{,}039\,THI_0$;
2013. aasta lõpuks $THI_2 = 1{,}039 \cdot 1{,}028\,THI_0$;
2014. aasta lõpuks $THI_3 = 1{,}039 \cdot 1{,}028 \cdot 0{,}999\,THI_0$.

Otsime keskmist kasvutempot aastas, mis olgu x. Kasutades keskmist kasvutempot kolm aastat järjest, peame saama väärtuse THI_3:

$$\begin{aligned} x \cdot x \cdot x \cdot THI_0 &= 1{,}039 \cdot 1{,}028 \cdot 0{,}999\,THI_0, \\ x^3 &= 1{,}039 \cdot 1{,}028 \cdot 0{,}999, \\ x &= \sqrt[3]{1{,}039 \cdot 1{,}028 \cdot 0{,}999} \approx 1{,}022. \end{aligned}$$

Vastus: keskmiselt kasvas tarbijahinnaindeks 2,2% aastas.

Vaadeldud näidetes jõudsime samuti geomeetrilise keskmise valemini. Tuletame selle valemi üldisemal juhul. Olgu meil mingi suurus a, mille kasvutempo i-ndal perioodil on x_i. Siis esimese kolme perioodi

jaoks

$$\begin{aligned} a_1 &= x_1 \cdot a_0; \\ a_2 &= x_2 \cdot a_1; \\ a_3 &= x_3 \cdot a_2. \end{aligned}$$

Avaldame viimase liikme a_3 esimese liikme a_0 kaudu:

$$a_3 = x_1 \cdot x_2 \cdot x_3 \cdot a_0.$$

Teiselt poolt võime me a_3 arvutamiseks kasutada kasvutempode geomeetrilist keskmist $\bar{x}_{geom}$:

$$a_3 = \bar{x}_{geom} \cdot \bar{x}_{geom} \cdot \bar{x}_{geom} \cdot a_0 = (\bar{x}_{geom})^3 \cdot a_0.$$

Võrreldes viimase kahe avaldise vasakuid ja paremaid pooli, saame avaldada $\bar{x}_{geom}$:

$$\begin{aligned} (\bar{x}_{geom})^3 &= x_1 \cdot x_2 \cdot x_3, \\ \bar{x}_{geom} &= \sqrt[3]{x_1 \cdot x_2 \cdot x_3}. \end{aligned}$$

Saime geomeetrilise keskmise valemi kolme väärtuse korral. Seda võime üldistada n väärtuse jaoks.

Geomeetriline keskmine

Geomeetriline keskmine on tunnuse väärtuste korrutis, mis on astendatud nende väärtuste arvu pöördväärtusega:

$$\bar{x}_{geom} = (x_1 \cdot x_2 \cdot \ldots \cdot x_n)^{\frac{1}{n}} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}. \qquad (2.26)$$

Kaalutud geomeetrilise keskmise leidmiseks kasutatakse valemit

$$\bar{x}_{geom} = (x_1^{f_1} \cdot x_2^{f_2} \cdot \ldots \cdot x_n^{f_n})^{\frac{1}{\Sigma f_i}} = \sqrt[\Sigma f_i]{x_1^{f_1} \cdot x_2^{f_2} \cdot \ldots \cdot x_n^{f_n}}, \qquad (2.27)$$

kus f_i on väärtuse x_i kaal ehk esinemissagedus.

Näide 2.25. Käibe keskmine kasvutempo

Ettevõtte käibe kasvutempo oli kolmel kuul 1,1, neljal kuul 1,05 ja viiel kuul 1,09. Kui suur oli käibe keskmine kasvutempo kuus? Kasutame kaalutud geomeetrilise keskmise valemit (2.27):

$$\bar{x}_{geom} = \sqrt[12]{1{,}1^3 \cdot 1{,}05^4 \cdot 1{,}09^5} = \sqrt[12]{2{,}4892} \approx 1{,}079.$$

Vastus: keskmine kasvutempo kuus oli 1,079.

Keskmist kasvutempot kasutatakse majandusstatistikas tihti hindade, palkade, SKP ja muude suuruste dünaamika iseloomustamisel. Sageli on muutus antud protsentuaalse muutusena, mida nimetatakse ka **juurdekasvutempoks**. Kui kasvutempo on 1,05, siis juurdekasvutempo on 0,05 ehk 5%: see näitab, kui suur osa juurde tuli.

Keskmise protsentuaalse muutuse leidmiseks EI TOHI kasutada aritmeetilist keskmist. See annab vale tulemuse, nagu nägime näites 2.23. Aritmeetilise keskmise leidmiseks tuleb üksikud väärtused summeerida, kuid ajas järgnevaid protsentuaalseid muutusi ei saa liita, sest need protsendid on võetud erinevatest arvudest.

Keskmine kasvutempo

Geomeetrilist keskmist kasutatakse **keskmise kasvutempo** arvutamisel.
Kui on antud protsentuaalsed muutused (juurdekasvutempod), siis **keskmise protsentuaalse muutuse** leidmiseks tuleb leida

1) kasvutempod;
2) kasvutempode geomeetriline keskmine $\bar{x}_{geom}$;
3) sellele vastav protsentuaalne muutus $\bar{x}_{geom} - 1$.

Tabelarvutuses leiab geomeetrilise keskmise **GEOMEAN**. Kaalutud geomeetrilise keskmise leidmiseks vastav funktsioon puudub.

2.7. Ruutkeskmine

Kui tuleb leida hälbeid või kõrvalekaldeid normidest, standarditest, planeeritavatest tasemetest või prognoosidest, tuleb hälvete keskmise suuruse arvutamiseks kasutada ruutkeskmist. Hälbed võivad olla nii positiivsed kui negatiivsed ja tavalise aritmeetilise keskmise leidmisel positiivsed ja negatiivsed hälbed kustutavad üksteist.

Näide 2.26. Laekumiste keskmine erinevus

Viiel kuul oli igakuiste laekumiste erinevus planeeritust järgmine (tuhandetes eurodes): 54; −22; −18; 32 ja −46. Leida keskmine kõrvalekalle planeeritust.
Aritmeetilise keskmise arvutamine annab tulemuseks 0. Selle asemel leiame ruutude aritmeetilise keskmise:

$$\frac{54^2 + (-22)^2 + (-18)^2 + 32^2 + (-46)^2}{5} = 1372{,}8.$$

Kuna me tõstsime summad ruutu, siis ühikuks on (tuhat eurot)^2. Loomulik on võtta sellest ruutjuur, siis saame ühikuks tuhat eurot:

$$\sqrt{1372{,}8} \approx 37{,}1.$$

Vastus: keskmine kõrvalekalle planeeritust oli 37,1 tuhat eurot.

Keskmise hälbe leidmisel leidsime hälvete ruutude aritmeetilise keskmise ja seejärel võtsime ruutjuure. Seda nimetatakse hälvete ruutkeskmiseks.

Ruutkeskmine

Ruutkeskmine on ruutjuur väärtuste ruutude aritmeetilisest keskmisest:

$$\bar{x}_{rk} = \sqrt{\frac{\sum x_i^2}{n}}. \tag{2.28}$$

Kaalutud ruutkeskmine:

$$\bar{x}_{rk} = \sqrt{\frac{\sum x_i^2 f_i}{\sum f_i}}, \tag{2.29}$$

kus f_i on väärtuse x_i kaal.

Ruutkeskmist kasutatakse tihti erinevate prognooside võrdlemisel: mida väiksem on prognoosi ruutkeskmine viga, seda parem prognoos. Selle järgi saab otsustada, milline prognoosimismeetod on parem.

Näide 2.27. Ehitustööde prognoosi ruutkeskmine viga

Tabelis on toodud Eestis teostatud ehitustööde mahtude kaks prognoosi 2014. aasta neljaks kvartaliks. Ühikuks on miljonit eurot. Tuleb otsustada, kumb prognoos on täpsem. Selleks on viimasesse veergu leitud mõlema prognoosi viga: tegelik väärtus miinus prognoos.

Kvartal	Prognoos 1	Prognoos 2	Tegelik	Prognoosi 1 viga	Prognoosi 2 viga
I	531,2	531,2	539,7	8,5	8,5
II	925,2	960,6	758,1	−167,1	−202,5
III	1147,2	1136,3	935,5	−211,7	−200,8
IV	969,9	959,6	876,5	−93,4	−83,1

Prognoosi 1 ruutkeskmine viga:

$$\bar{x}_{rk1} = \sqrt{\frac{8{,}5^2 + (-167{,}1)^2 + (-211{,}7)^2 + (-93{,}4)^2}{4}} \approx 142{,}8.$$

Prognoosi 2 ruutkeskmine viga:

$$\bar{x}_{rk2} = \sqrt{\frac{8{,}5^2 + (-202{,}5)^2 + (-200{,}8)^2 + (-83{,}1)^2}{4}} \approx 148{,}6.$$

Vastus: kuna 1. prognoosi ruutkeskmine viga on väiksem, on see prognoos täpsem.

Tabelarvutuses ruutkeskmise jaoks funktsioon puudub. Ruutude summa leidmiseks võib kasutada funktsiooni **SUMSQ**, tulemus jagada väärtuste arvuga ning võtta ruutjuur funktsiooniga **SQRT**.

2.8. Keskmiste liigitus ja järgnevus

Kõik vaadeldud statistilised keskmised saab jagada kaheks:

Mahu- ja asendikeskmised

- **mahukeskmised** on sellised keskmised, mille arvuline väärtus sõltub igast üksikust statistilise rea väärtusest;
- **asendikeskmised** ehk struktuurikeskmised reageerivad ainult niisugustele muutustele rea üksikliikmete väärtustes, millega kaasneb olulisi nihkeid ka rea struktuuris: muutub variatsioonrea liikmete asend üksteise suhtes või eri väärtusega liikmete osatähtsus.

Tabelis 2.6 on toodud keskmiste liigitus ja kasutusvõimalused: milliste skaalade korral saab üht või teist keskmist kasutada.

Tabel 2.6. Keskmiste liigitus ja kasutusvõimalused

Keskmine	Liik	Skaala, mille korral saab kasutada		
		nimi	järjestus	intervall
aritmeetiline keskmine	mahukeskmine			+
harmooniline keskmine	mahukeskmine			+
geomeetriline keskmine	mahukeskmine			+
ruutkeskmine	mahukeskmine			+
mediaan	asendikeskmine		+	+
kvantiilid	asendikeskmine		+	+
mood	asendikeskmine	+	+	+

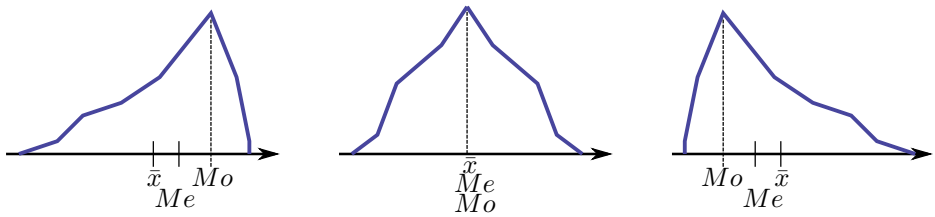
Samadest arvudest leitud erinevate keskmiste arvväärtused on erinevad. Mahukeskmiste korral kehtib järgmine suurusjärgnevus:

$$\bar{x}_{harm} \leq \bar{x}_{geom} \leq \bar{x}_{aritm} \leq \bar{x}_{rk}. \tag{2.30}$$

Aritmeetilise keskmise, mediaani ja moodi järjestus sõltub unimodaalse jaotuse korral aga jaotuse kujust ehk sellest, kas esineb ekstreemseid väärtusi ja kui esineb, siis kummal pool: miinimumi või maksimumi pool (vt joonis 2.12).

Suured ekstreemsed väärtused: $Mo < Me < \bar{x}$.
Väikesed ekstreemsed väärtused: $Mo > Me > \bar{x}$.

Multimodaalse jaotuse korral selline järgnevus ei pruugi kehtida.



Joonis 2.12. Aritmeetilise keskmise $\bar{x}$, mediaani Me ja moodi Mo suurusjärgnevus

Kvantitatiivse tunnuse (intervallskaalas) korral on soovitav esitada kõik kolm keskmist: aritmeetiline keskmine, mediaan ja mood (kui esineb), sest need annavad erinevat informatsiooni. Harmoonilisel, geomeetrilisel ja ruutkeskmisel on tunnetuslik väärtus vaid teatud tüüpi andmete ja probleemide korral.

Uurija ülesandeks on leida ja esitada sellised keskmised, mis iseloomustavad kogumit kõige paremini.

2.9. Ülesanded

Aritmeetiline keskmine

2.1. Erinevatel inimestel kulus ühe ülesande täitmiseks erinev aeg. Ajad (minutites) olid järgmised: 45, 47, 52, 36, 48, 47, 50, 51. Kasutades aritmeetilist keskmist, leida keskmine aeg, mis kulub antud ülesande täitmiseks. VASTUS lk 659.

2.2. Viie arvu summa on 90. Kui suur on nende arvude aritmeetiline keskmine? VASTUS lk 659.

2.3. Kolme arvu aritmeetiline keskmine on 4. Kui suur on nende arvude summa? VASTUS lk 659.

2.4. Tiiu, Anu ja Pille käisid koos lõunat söömas. Arve otsustasid nad jagada võrdselt ning igaüks maksis 3,4 eurot. Kui Tiiu praad mak-

sis 3,5 eurot ja Anu praad 3 eurot, siis kui palju maksis Pille praad? VASTUS lk 659.

2.5. 2011. aastal oli ettevõtte töötajate vanuste aritmeetiline keskmine 32 aastat. Kolme aasta jooksul jäi ettevõtte kollektiiv samaks: ühtegi töötajat ei tulnud juurde ega ei lahkunud. Kui suur oli selle ettevõtte töötajate vanuste aritmeetiline keskmine aastal 2014? VASTUS lk 659.

2.6. Ettevõtte tööjõukulu suurenes 8%. Kuidas muutus keskmine tööjõukulu ühe töötaja kohta, kui

a) töötajate arv jäi samaks;
b) töötajate arv suurenes 20%?

VASTUS lk 659.

2.7. Leida järgmiste kogumite mediaan: *Mediaan*

a) {8, 4, 30, 12, 15};
b) {70, 17, 14, 45, 28, 24};
c) {väga väike, suur, suur, väike, suur, väga suur}.

VASTUS lk 659.

2.8. Osakonnas töötava viie inimese vanused on järgmised: 26, 35, 37, 43, 65. Lahkus 65-aastane töötaja ja asemele tuli 41-aastane töötaja. Leida vanuste aritmeetiline keskmine ja mediaan enne ning pärast töötaja vahetumist. VASTUS lk 659.

2.9. Õppejõud analüüsis, kui palju aega kulutasid üliõpilased keskmiselt testi tegemiseks. Aritmeetiline keskmine oli 34 minutit ja mediaan 18 minutit. Kas see test oli üliõpilaste jaoks pigem kerge või pigem raske? VASTUS lk 659.

2.10. Olgu meil 5 erinevat positiivset täisarvu. Millised teisendused ei muuda selle arvukogumi mediaani?

a) Kõikide arvude korrutamine arvuga 2;
b) vähimast arvust arvu 1 lahutamine;
c) suurimast arvust arvu 1 lahutamine.

VASTUS lk 659.

2.11. Ehitusettevõttel oli 2013. aastal tööjõukulude osatähtsus müügitulus 25%. Milline oli selle ettevõtte asend teiste ehitusettevõtete seas, kui Eesti Statistikaameti andmetel oli ehitusvaldkonnas vastava näitaja 1. kvartiil 6,44%, mediaan 17,24% ja 3. kvartiil 30,71%? VASTUS lk 659. *Kvartiilid*

2.12. Uuringus osales 1500 inimest. Ankeedis oli ka küsimus vastaja sissetuleku kohta. Vastuste põhjal leitud sissetulekute 1. kvartiil oli 760 eurot ja mediaan 900 eurot. Mitme uuringus osaleja sissetulek oli

a) väiksem kui 760 eurot;

b) 760 ja 900 euro vahel;
c) suurem kui 900 eurot?

VASTUS lk 659.

2.13. Täiskasvanud rahvastiku tervisekäitumise uuringu raames küsitletakse inimesi ka toitumisharjumuste kohta. Tabelis on viimase 7 päeva jooksul keskmiselt söödud köögiviljade kogus 2014. aastal ja protsent vastanutest, mehed ja naised eraldi[2]. Leida söödud köögiviljade koguse mediaan meestel ja naistel. VASTUS lk 659.

Tarbimine, g	Mehed, %	Naised, %
0	11,2	5,5
0–100	12,5	6,3
100–200	33,8	27,4
200–300	20,0	26,8
üle 300	22,5	34,0

Mood

2.14. Linna turgudel läbiviidud vaatluse tulemusena selgus, et kõige sagedamini küsiti banaanikilo eest 1 euro 40 senti. Millise statistilise suurusega on tegemist? VASTUS lk 659.

2.15. 2013. aastal sai Eestis toimetulekutoetust 19 320 perekonda, nendest Harjumaal 5214 ja Võrumaal 744 perekonda[3]. Aastas võis toetust saada üks kuni 12 korda. Tabelis on toodud perede arv, kes said toetust üks kord, kaks korda jne. Milline oli toetuse saamise kordade arvu mood Harjumaal ja Võrumaal? VASTUS lk 659.

Kordade arv	1	2	3	4	5	6
Harjumaa	907	629	468	411	377	323
Võrumaa	100	76	55	65	46	38
Kordade arv	7	8	9	10	11	12
Harjumaa	317	260	245	226	294	757
Võrumaa	43	34	33	29	28	197

Aritmeetiline keskmine, mediaan, mood

2.16. 2014. aasta kevadel oli Statistika õppeaine eksamihinnete jaotus järgmine:

Hinne	„0“	„1“	„2“	„3“	„4“	„5“
Sagedus	20	22	21	13	9	4

[2] Allikas: Tervise Arengu Instituut `http://www.tai.ee/`. Tabel TKU11: viimase 7 päeva jooksul keskmiselt söödud köögiviljade kogus soo ja vanusrühma järgi.

[3] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel SK43: toimetulekutoetust saanud perekondadde arv piirkonna/haldusüksuse järgi.

Leida hinnete mood ja mediaan. VASTUS lk 659.

2.17. Klassi õpilaste kohta on teada järgmised andmed:

Nimi	Silmade värv	Käitumishinne	Pikkus (m)
Juku	sinine	mitterahuldav	1,56
Juhan	hall	rahuldav	1,62
Mall	pruun	hea	1,71
Liisi	pruun	eeskujulik	1,58
Ants	sinine	rahuldav	1,67
Kaarel	pruun	hea	1,66
Tõnu	hall	rahuldav	1,60
Mari	pruun	rahuldav	1,59
Tiina	sinine	hea	1,61

Milliste skaaladega on tegemist? Iga tunnuse jaoks määrata enda arvates seda tunnust kõige paremini iseloomustav statistiline keskmine ja leida selle väärtus. VASTUS lk 659.

2.18. Portaali Domus Kinnisvara ajaveebis 27. augustil 2014. a postitatud ülevaates „Pärnu korterituru trendid“[4] on kirjas:

„2014. aasta I ja II kvartalis on tehtud kokku 353 korteriomandi tehingut, eelmisel aastal samal ajal tehti kokku 322 korteriomandi tehingut. Keskmine müügihind 2013. aasta I poolaastal oli 751 €/m^2 (mediaanhind 706 €/m^2), käesoleval aastal on I poolaasta keskmine hind tõusnud tasemele 882 €/m^2 (mediaanhind 809 €/m^2).“

Kui suur oli Pärnu korterituru kogukäive 2013. aasta I poolaastal ja kui suur 2014. aasta I poolaastal? VASTUS lk 659.

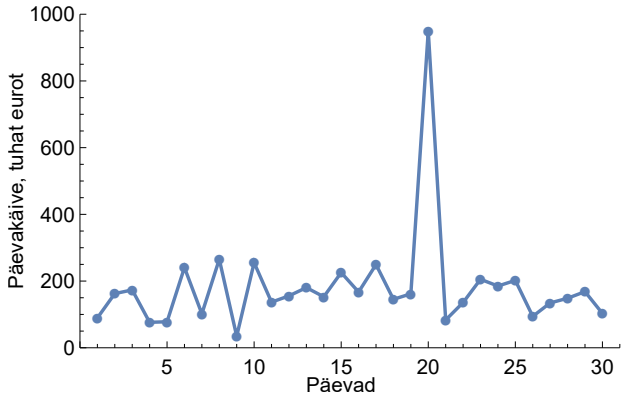
2.19. Viie inimese kuu sissetulekute aritmeetiline keskmine on 1280 eurot, mediaan 1100 eurot ja mood 800 eurot. Kõige väiksem sissetulek on 800 eurot ja kõige suurem 2200 eurot kuus. Leida kõigi viie inimese sissetulekud kuus. VASTUS lk 659.

2.20. Meediauuringu raames küsitleti 600 peret. Muuhulgas küsiti, mitu tundi päevas keskmiselt vaadatakse televiisorit. Tulemused on toodud tabelis. Leida mood. VASTUS lk 659.

Mitu tundi päevas	Sagedus
0–0,5	70
0,5–1	100
1–2	140
2–3	230
3 ja rohkem	60

[4] `http://www.domuskinnisvara.ee/blogi/2014/08/parnu-korterituru-trendid/`

2.21. Suure elektroonikapoe kõrval tehtavatel kaevamistöödel vigastas ekskavaatorijuht elektrikaablit, mille tõttu oli poes keset päeva kaks tundi voolukatkestus. Kuna sel ajal kaupa müüa ei saanud, siis nõuab pood kaevetööde teostajalt kahju hüvitamist. Kahjusumma arvutamisel arvestati, et saamata jäi 1/5 keskmisest päevakäibest, kuna pood on lahti 10 tundi päevas. Keskmine päevakäive leiti viimase kuu päevakäivete aritmeetilise keskmisena ja see oli 182,07 tuhat eurot. Kaevetöid teinud ettevõtte esindaja aga nõuab, et kahjusumma arvutamisel võetaks aluseks summa 158,7 tuhat eurot, mis on eelmise kuu päevakäivete mediaan. Viimase kuu päevakäibed on toodud joonisel 2.13. Kumba summat tuleks kahjutasu arvutamisel aluseks võtta? VASTUS lk 659.



Joonis 2.13. Elektroonikapoe viimase kuu käive päevas

Harmooniline keskmine

2.22. Esimesed 100 km läbis auto kiirusega 70 km/h, järgmised 100 km kiirusega 110 km/h. Milline oli auto keskmine kiirus? VASTUS lk 659.

2.23. Hulgimüüjalt A osteti kaupa X 6510 euro eest hinnaga 14 €/tk. Hulgimüüjalt B osteti seda kaupa 5075 euro eest hinnaga 14,50 €/tk ja hulgimüüjalt C osteti sama kaupa 4000 euro eest hinnaga 16 €/tk. Milliseks kujunes kauba keskmine hind? VASTUS lk 659.

2.24. Ühelt põllult koristas talumees 180 tonni teravilja ja seal oli saagikus 4,5 tonni hektari kohta. Teisel põllul oli saagikus 4 t/ha ning sealt koristas ta 80 tonni. Kolmandal põllul oli saagikus kõige väiksem, 3,8 t/ha ja sealt tuli saagiks 114 tonni. Kui suur oli keskmine saagikus? VASTUS lk 659.

Harmooniline ja aritmeetiline keskmine

2.25. Ema ostis kaks kilo banaane hinnaga 1,5 €/kg, isa ostis kolm kilo banaane hinnaga 1,9 €/kg ja poeg ostis kaks kilo banaane hinnaga

1,1 €/kg. Kui suureks kujunes banaanide keskmine kilohind? VASTUS lk 659.

2.26. 3 euro eest osteti banaane hinnaga 1,5 €/kg, 5,7 euro eest osteti banaane hinnaga 1,9 €/kg ja 2,2 euro eest osteti hinnaga 1,1 €/kg. Milliseks kujunes banaanide keskmine kilohind? VASTUS lk 659.

2.27. 5 töölist panid kokku igaüks 20 toodet, 7 töölist valmistasid 25 toodet töölise kohta ja 10 töölist 30 toodet töölise kohta. Leida keskmine tootlikkus. VASTUS lk 659.

2.28. 100 toodet pandi kokku tootlikkusega 20 toodet töölise kohta, 175 toote kokkupanekul oli tootlikkus 25 toodet töölise kohta ja 300 toote kokkupanekul 30 toodet töölise kohta. Leida keskmine tootlikkus. VASTUS lk 659.

Geomeetriline keskmine

2.29. Eesti ekspordi kasvutempo oli 2012. aastal 1,043, 2013. aastal 0,982 ja 2014. aastal 0,983. Kui suur oli keskmine kasvutempo aastas? VASTUS lk 659.

2.30. Kauba hind on kasvanud järgmiselt: 1. aastal 6%, 2. aastal 13%, 3. aastal 11% ja 4. aastal 15%. Mitu protsenti on kauba hind keskmiselt aastas kasvanud? VASTUS lk 659.

2.31. Viiel kuul tõusis tarbijahinnaindeks 1,5%, neljal kuul 2% ja kolmel kuul 3%. Milline oli keskmine tarbijahinnaindeksi kasv kuus? VASTUS lk 659.

2.32. Vallavalitsusel on sõlmitud leping ettevõttega, kes osutab antud valla piires prügiveoteenust. Leping on sõlmitud viie aasta peale ja vastav ettevõte on ainuke selle teenuse pakkuja antud vallas (monopolistlik ettevõte). Lepingus on seetõttu ka punkt, milles märgitakse, et viie aasta jooksul ei tohi keskmine hinnatõus aastas olla suurem kui 10%. Tegelik hinnatõus oli järgmine: 1. aastal tõsteti hinda 30%, teisel ja kolmandal aastal hind ei tõusnud, 4. aastal tõusis hind 15% ja viimasel aastal 7%. Vallavalitsus süüdistab ettevõtet, et viie aasta jooksul oli keskmine hinnatõus aastas 10,4%, mis on suurem kui lepingus lubatud 10%. Ettevõte aga väidab, et keskmine hinnatõus aastas jäi 10% piiridesse, täpsemalt oli 9,85%. Kummal on õigus? VASTUS lk 659.

2.33. Töötaja palk tõusis 1. aastal 5,8%, 2. aastal 8,5% ja 3. aastal 3,2%. Mitu protsenti peaks palk tõusma 4. aastal, et keskmine palgatõus oleks 6% aastas? VASTUS lk 659.

2.34. Näidata, et geomeetrilise keskmise leidmiseks võib valemi (2.26) asemel kasutada ka valemit $\bar{x}_{geom} = exp\left(\frac{1}{n}\sum_{i=1}^{n}\ln x_i\right)$. Näpunäide: lähtuda valemist (2.26) ja võtta mõlemast poolest naturaallogaritm.

ÜL02Keskmised

Järgmiste ülesannete andmed on failis ÜL02Keskmised

Erinevad keskmised

A.2.1. Tabelis on 2009. aastal loodud ettevõtete arv Jõgeva ja Põlva maakonna valdades[5]. Leida keskmine loodud ettevõtete arv ühe valla kohta kummaski maakonnas. VASTUS lk 659.

A.2.2. Kuu keskmine jääk pangakontol sõltub sellest, kui kaua mingi summa kontol on. Tabelis on toodud ühe pangakonto jäägi muutus 2015. a jaanuaris. Leida kuu keskmine jääk. VASTUS lk 659.

A.2.3. Tabelis on toodud omavalitsuste elanike arv kolmes Eesti maakonnas seisuga 1.07.2011. Igas maakonnas on välja jäetud maakonnakeskuse andmed.

1. Leida elanike arvu aritmeetiline keskmine ja mediaan kõigis kolmes maakonnas. Millises maakonnas on aritmeetiline keskmine kõige suurem?
2. Lisada tabelisse maakonnakeskuste andmed: Pärnu linn 4266, Põlva linn 6241, Tartu linn 98 365. Leida aritmeetiline keskmine ja mediaan koos maakonnakeskuste andmetega. Millises maakonnas on nüüd aritmeetiline keskmine kõige suurem?
3. Iga maakonna jaoks leida, mitu protsenti suurenes maakonnakeskuse lisamisel elanike arvu aritmeetiline keskmine ja mitu protsenti suurenes mediaan.

VASTUS lk 659.

A.2.4. Uuringukeskus Klaster viis 2006. aasta aprillis Eesti Üliõpilaskondade Liidu tellimusel läbi uuringu üliõpilaste sotsiaalmajanduslikust olukorrast (Üliõpilaste ...). Kokku osales uuringus 4532 üliõpilast. Üheks üliõpilastele esitatud küsimuseks oli, kui pikk peaks olema praktikaperiood tema erialal. Ette olid antud vastusevariandid, mis vahemikku võiks praktika pikkus jääda. Tabelis on toodud kolme valdkonna (ärindus ja haldus, sotsiaalteadused, tervis ja heaolu) üliõpilaste vastuste osakaalud. Leida iga valdkonna jaoks vastajate poolt pakutud keskmine praktika pikkus (aritmeetiline keskmine). VASTUS lk 659.

A.2.5. Tabelis on töötuse määr Eesti maakondades 2014. aastal[6]. Leida mediaan. Millistes maakondades oli töötuse määr mediaanist suurem? VASTUS lk 659.

A.2.6. Riigi arengutaseme üheks näitajaks on SKP elaniku kohta. Euroopa Liidus võrreldakse riigi näitajat tihti ELi keskmisega. Tabelis

[5] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel ER072: sündinud ettevõtted haldusüksuse järgi.

[6] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel TT4646 16 kuni pensionieealiste hõiveseisund makonna järgi.

on Euroopa riikide SKP elaniku kohta, võrrelduna ELi 28 riigi keskmisega (EL28 keskmine = 100) aastatel 2002 ja 2013.[7]. Mitmendasse neljandikku kuulus Eesti aastal 2002 ja mitmendasse neljandikku aastal 2013? VASTUS lk 659.

A.2.7. 1998. aastal oli Eestis 32,1 lauatelefoni 100 elaniku kohta. Oli see näitaja kõrge või madal? Tabelis on toodud lauatelefonide arv 100 elaniku kohta 188 maailma riigis[8]. Leida, mitmendas neljandikus asus Eesti maailma riikide seas. VASTUS lk 659.

A.2.8. Maksu- ja Tolliamet avaldab oma veebilehel juriidilistest isikutest maksuvõlglaste nimekirja. Nimekirjas on need, kelle maksuvõlg on vähemalt 1000 eurot[9]. Tabelis on maksuvõlglaste võla suurus seisuga 2.02.2015, kokku 4742 kirjet. Leida

a) kõige suurem maksuvõlg;
b) kõigi nimekirjas olevate võlglaste maksuvõlg kokku;
c) protsentiilid järguga 0,1, 0,5, 0,95, 0,99;
d) kui palju on võlglaste hulgas neid, kelle võlg on suurem kui protsentiil 0,99.

VASTUS lk 660.

A.2.9. Poes viidi läbi küsitlus leivatoodete eelistuse kohta. Küsitleti 40 ostjat ja küsimuseks oli „Millist leiba kõige sagedamini ostate?“ Vastusevariandid olid

Leivasort	Kood andmetabelis
Viru	1
Madise	2
Taluleib	3
Toolse	4
Muu	5

Millist leivasorti ostsid ostjad kõige sagedamini? VASTUS lk 660.

A.2.10. Hinnavaatlusel uuriti sealiha kilogrammi hinda linna erinevatel turgudel. Toodud andmete põhjal leida sealiha keskmine hind linnas. Millist keskmist tuleb kasutada? VASTUS lk 660.

A.2.11. Hindade muutumise analüüsimisel on üheks meetodiks uurida, kui sagedasti erineb mingi kauba hind selle baashinnast. Baashinnaks võetakse tavaliselt hindade mood ehk modaalhind mingi perioodi jooksul. Tabelis on 5 kg maisi hind (USD) kahel Zimbabwe turul jaanuar 2010 kuni detsember 2015[10].

[7] Allikas: Eurostat `http://ec.europa.eu/eurostat`
[8] Allikas: International Telecommunication Union, `http://www.itu.int/`
[9] Täpsemalt vt `http://www.emta.ee/index.php?id=35438`
[10] Allikas: The Humanitarian Data Exchange `https://data.hdx.rwlabs.org/` Global Food Prices Database.

1. Leida maisi modaalhind kummagi turu jaoks.
2. Leida kummagi turu jaoks, kui suurel osal vaadeldud kuudest oli hind modaalhinnast erinev.
3. Kummal turul on hinna erinemine modaalhinnast sagedasem?
4. Leida kummagi turu jaoks, kui suurel osal vaadeldud kuudest oli hind modaalhinnast kõrgem.

VASTUS lk 660.

A.2.12. Tabelis on toodud andmed Eestis registreeritud bensiinimootoriga sõiduautode kohta seisuga 31.12.2015[11]. Autod on jaotatud mootori võimsuse järgi sagedusklassidesse.

1. Leida registreeritud autode võimsuse aritmeetiline keskmine, mediaan ja mood.
2. Konstrueerida diagramm, kus horisontaalteljel on klassi keskpunkt ja vertikaalteljel sagedustihedus.

VASTUS lk 660.

A.2.13. Aastatel 1989–1994 alustati Chicago Ülikooli ärikooli Booth School of Business ja kaupluseketi Dominick's Finer Foods vahel koostööd kaubahindade uurimisel. Andmebaasis Dominick's Database[12] on rohkem kui 3500 kauba hinnad ja üheksa aasta läbimüük 100 poes. Sellest andmebaasist on võetud andmed poe River Forest toiduainete müügi päevakäivete kohta 1. jaanuar 1988 – 30. aprill 1997, kokku 3353 päeva. Toiduainete päevakäibed on rühmitatud sagedusklassidesse.

1. Leida päevakäivete aritmeetiline keskmine, mediaan ja mood.
2. Konstrueerida diagramm, kus horisontaalteljel on klassi keskpunkt ja vertikaalteljel sagedustihedus.

VASTUS lk 660.

A.2.14. Tabelis on toodud kasutusloa saanud eluruumide arvu kasvutempo aastatel 2001–2007[13]. Kui suur oli sellel perioodil keskmine kasvutempo aastas? VASTUS lk 660.

A.2.15. Tööstustoodangu tootjahinnaindeks iseloomustab Eestis valmistatud tööstustoodete hindade muutust. Indeks hõlmab nii kodumaisele turule kui ka välisturule valmistatud tooteid. Tabelis on tootjahinnaindeksi muutus protsentides võrreldes eelmise aastaga aastatel

[11] Allikas: Maanteeameti veebileht `http://www.mnt.ee/`. Sõidukite ja juhilubade statistika

[12] Kilts Center for Marketing `https://research.chicagobooth.edu/kilts/marketing-databases/dominicks`

[13] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel EH04: ehitusloa saanud ja kasutusse lubatud eluruumide uusehitus.

2003–2014[14]. Leida, mitu protsenti muutus tootjahinnaindeks keskmiselt aastas. VASTUS lk 660.

A.2.16. Päikeseprillide müügiga tegelev ettevõte tellis 14 erinevat mudelit prille. Igat mudelit telliti 20 tuhande euro eest. Tabelis on toodud nende mudelite sisseostuhind eurodes. Kui suur on keskmine päikseprillide sisseostuhind? VASTUS lk 660.

A.2.17. Tabelis on toodud 20 tooteartikli omahind ja kui suure summa eest on seda toodet kuu aja jooksul toodetud. Kui suur on toodete keskmine omahind? VASTUS lk 660.

A.2.18. Lossimine on kauba laevalt mahalaadimine sadamas. Tabelis on toodud kaupade lossimine Eesti sadamates, tuhat tonni[15].

1. Leida kuised kasvutempod aastatel 2009–2011. Kasvutempo on kahe järjestikuse väärtuse jagatis.
2. Millisel aastal oli kuu keskmine kasvutempo kõige väiksem?

VASTUS lk 660.

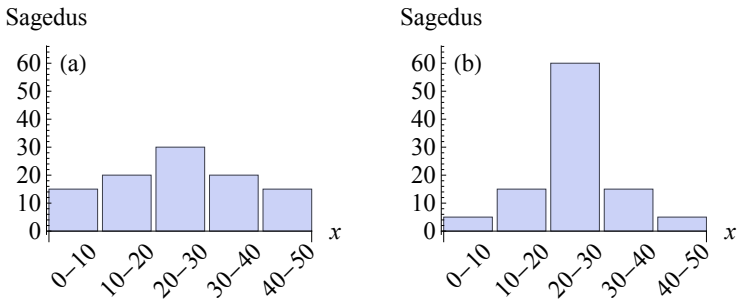
[14] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel IA029: tööstustoodangu tootjahinnaindeksi muutus võrreldes eelmise aastaga.

[15] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel TS185: kaupade lastimine ja lossimine Eesti sadamates.

Peatükk 3

Variatsioonnäitarvud ja jaotuse kuju näitarvud

Kui kogumi liikmetel on ühe ja sama tunnuse väärtused erinevad, siis selle tunnuse väärtus **varieerub** ehk **hajub**. Keskmised kajastavad tunnuse väärtuste üldist keskmist nivood, ei kajasta aga nende omavahelisi erinevusi, ei kajasta varieerumist.



Joonis 3.1. Kahe erineva arvukogumi varieerumine

Joonisel 3.1 on toodud kahe erineva tunnuse sagedusdiagrammid. Mõlema tunnuse väärtused jaotuvad sümmeetriliselt ning neil on ühesugune mood, mediaan ja aritmeetiline keskmine. Aga diagrammidelt on näha, et varieerumine on erinev. Diagrammil (b) on rohkem väärtusi kogunenud keskmise ümber, seal on varieerumine väiksem kui diagrammil (a).

Varieerumise iseloomustamiseks kasutatakse mitmesuguseid **variatsioonnäitarve** ehk **hajuvuskarakteristikuid**, mis väljendavad kogumi üksikliikmete vahelisi erinevusi.

3.1. Variatsioonamplituud

Kui me soovime teada, millistes piirides mingi tunnuse väärtused varieeruvad, siis me leiame väärtuste miinimumi ja maksimumi. Nende vahe on variatsioonamplituud.

Variatsioon-amplituud

Variatsioonamplituud ehk **haare** (*range*) on arvukogumi kõige suurema liikme x_{max} ja kõige väiksema liikme x_{min} vahe:

$$R = x_{\max} - x_{\min}. \tag{3.1}$$

Tabelarvutuses kasutatakse maksimaalse väärtuse leidmiseks funktsiooni **MAX** ja minimaalse väärtuse jaoks funktsiooni **MIN**. Variatsioonamplituud on siis nende abil leitud tulemuste vahe

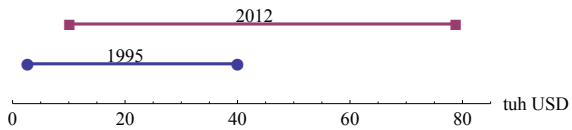
$$R = \text{MAX(arvud)} - \text{MIN(arvud)}.$$

N03Varieerumine N3.1

Näide 3.1. Netosissetulek elaniku kohta OECD riikides

Üheks riigi elatustaseme näitajaks on netosissetulek elaniku kohta. Tabelis ning joonisel 3.2 on toodud selle väärtuse minimaalne ja maksimaalne väärtus OECD riikide hulgas 1995. ja 2012. aastal[a]. Netosissetulek elaniku kohta on tuhandetes USA dollarites aastas. Näeme, et suurenenud on nii miinimum kui ka maksimum, kuid viimane on suurenenud rohkem. Seda näeb hästi variatsioonamplituudi suurenemisest.

	1995	2012
Miinimum, tuh USD	2,64	10,15
Maksimum, tuh USD	40,02	78,81
Variatsioonamplituud, tuh USD	37,38	68,66



Joonis 3.2. Netosissetulek elaniku kohta OECD riikides: miinimum, maksimum ja variatsioonamplituud

[a] Allikas: Maailmapank `http://databank.worldbank.org`

Variatsioonamplituud sõltub ainult kahest äärmisest väärtusest ja ei anna varieerumisest täielikku pilti. Joonisel 3.1 toodud jaotustel on

näiteks variatsioonamplituud ühesugune, varieerumine aga selgelt erinev. Lisaks sellele sõltub variatsioonamplituud enamasti kogumi mahust: mahu vähenemisel üldiselt minimaalne väärtus suureneb ja maksimaalne väärtus väheneb. Paremini iseloomustavad varieerumist need näitarvud, mis sõltuvad varieeruva tunnuse igast üksikust väärtusest x_i.

3.2. Keskmine absoluuthälve

Varieerumise iseloomustamiseks oleks sobiv kasutada üksikute väärtuste x_i hälbeid aritmeetilisest keskmisest $\bar{x}$. Aga vastavalt aritmeetilise keskmise tasakaaluomadusele (2.12) on hälvete summa

$$\sum(x_i - \bar{x}) = 0$$

ning järelikult hälvete aritmeetiline keskmine on alati 0. Hälvete absoluutväärtuste summa on aga nullist erinev ja seepärast saab varieerumise iseloomustamiseks kasutada hälvete absoluutväärtuste aritmeetilist keskmist.

Keskmine absoluuthälve

Keskmine absoluuthälve (*mean absolute deviation*) on väärtuste x_i ja nende aritmeetilise keskmise $\bar{x}$ vaheliste hälvete absoluutväärtuste aritmeetiline keskmine:

$$MAD = \frac{\sum |x_i - \bar{x}|}{n}. \tag{3.2}$$

Keskmist absoluuthälvet kasutatakse siiski harva, sest analüütiliste arvutuste tegemine (võrrandite lahendamine, diferentseerimine, integreerimine) on absoluutväärtust sisaldava suurusega komplitseeritud. Mõnikord nimetatakse suurust (3.2) ka **keskmiseks lineaarhälbeks** (näiteks õpikus (Aarma ja Vensel, 2005)).

Tabelarvutuses on keskmise absoluuthälbe leidmiseks funktsioon **AVEDEV**.

3.3. Dispersioon ja standardhälve

Et summeerimisel positiivsed ja negatiivsed hälbed üksteist ei kustutaks, on teine võimalus hälvete absoluutväärtuste kasutamise kõrval hälvete ruutude $(x_i - \bar{x})^2$ kasutamine. Nende aritmeetiline keskmine on dispersioon, millel on paremad matemaatilised omadused kui keskmisel absoluuthälbel.

Dispersioon

Dispersioon (*variance*) on hälvete ruutude aritmeetiline keskmine:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}. \tag{3.3}$$

Grupeeritud (sagedustabelina esitatud) andmete korral kasutatakse dispersiooni leidmiseks hälvete ruutude kaalutud aritmeetilist keskmist:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i}, \tag{3.4}$$

kus f_i on väärtuse x_i esinemissagedus.

Dispersiooni arvutamiseks võib kasutada ka valemit

$$\sigma^2 = \overline{x^2} - \bar{x}^2, \tag{3.5}$$

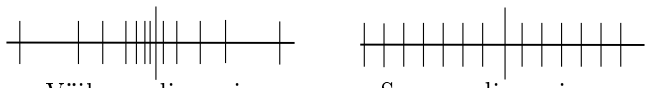
s.t tuleb leida ruutude aritmeetiline keskmine $\overline{x^2}$, aritmeetilise keskmise ruut $\bar{x}^2$ ja nende vahe. Valemite (3.3) ja (3.5) ekvivalentsust on näidatud lisas A.1.

Tabelarvutuses on kogumi dispersiooni leidmiseks funktsioon **VAR.P** (*VARiance of Population*).

Dispersiooni omadusi

Dispersiooni omadusi

1. Dispersioon on seda suurem, mida rohkem on tunnusel aritmeetilisest keskmisest tugevasti hälbivaid väärtusi ja mida suuremad on need hälbed (joonis 3.3).
2. Dispersioon on alati mittenegatiivne.
3. Konstantse tunnuse dispersioon on null.
4. Kui arvukogumi igale arvule liita (või lahutada) mingi konstant a, jääb dispersioon samaks.
5. Kui arvukogumi igat arvu korrutada mingi konstandiga c, siis dispersioon suureneb c^2 korda.



Joonis 3.3. Dispersioon on suurem, kui hälbed keskmisest on suuremad

Dispersiooni mõõtühikuks on vastava tunnuse mõõtühiku ruut, mida tihti on raske tõlgendada. Samuti ei saa dispersiooni võrrelda aritmeetilise keskmisega. Näiteks kui hinda mõõdetakse eurodes, siis hinna dispersiooni ühikuks on euro^2. Et tõlgendamine oleks lihtsam, võetakse dispersioonist ruutjuur ja saadakse standardhälve, mida praktikas kasutatakse rohkem.

Standardhälve (*standard deviation*) on ruutjuur dispersioonist:

$$\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}. \tag{3.6}$$

Standardhälve

Standardhälbe mõõtühikud on samad, mis aritmeetilisel keskmisel ja üksikutel väärtustel. Näiteks kui hind on eurodes, siis on ka hinna standardhälve eurodes.

Näide 3.2. Detailide läbimõõdu standardhälbe leidmine

N03Varieerumine
N3.2

Detailide töötlemise kvaliteedi määramiseks mõõdeti nende läbimõõdud ja leiti kogumi standardhälve, mille suurus iseloomustab töötlemise täpsust. Tulemused on toodud järgnevas tabelis.

i	Läbimõõt d_i, mm	$d_i - \bar{d}$	$(d_i - \bar{d})^2$
1	12,3	0,26	0,0676
2	11,9	−0,14	0,0196
3	12,0	−0,04	0,0016
4	11,8	−0,24	0,0576
5	12,2	0,16	0,0256

Läbimõõtude aritmeetiline keskmine $\bar{d} = 12{,}04$. Peale selle leidmist leiame vahed $d_i - \bar{d}$ ja vahede ruudud $(d_i - \bar{d})^2$. Vahede ruutude summa on

$$\sum(d_i - \bar{d})^2 = 0{,}172.$$

Standardhälve:

$$\sigma = \sqrt{\frac{\sum(d_i - \bar{d})^2}{n}} = \sqrt{\frac{0{,}172}{5}} \approx 0{,}185.$$

Vastus: detailide läbimõõdu standardhälve on 0,185 mm.

Kogumi standardhälbe leidmiseks kasutatakse tabelarvutuses funktsiooni **STDEV.P**, mis tuleb terminist *STandard DEViation of Population*.

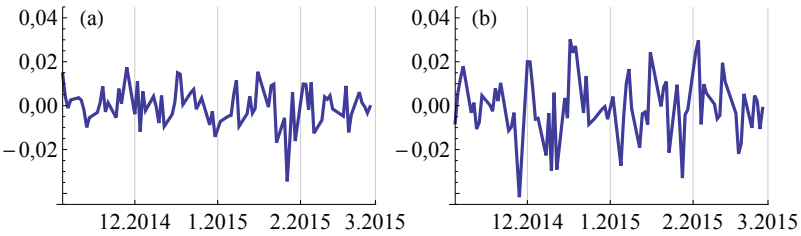
Järgnevalt toome mõningaid näiteid standardhälbe kasutamise kohta.

1. Erinevate kaupadega kauplemisel on oluline laovarude õigeaegne täiendamine õiges koguses. Kui kaup saab otsa, jääb müügitulu saamata. Kui aga laovarud on liiga suured, on palju raha varude all kinni

ning lisaks on vaja suurt laopinda. Laovarude juhtimise mudelites kasutatakse eelmiste perioodide müüginumbrite standardhälvet, et leida optimaalne tellitava partii suurus ja tellimisaeg (vt ka näidet 5.26 alapeatükis 5.10).

2. Tootmisettevõtte toodangu parameetrid peavad võimalikult vähe varieeruma. Kvaliteedikontrollis kasutatakse varieerumise hindamiseks standardhälvet. Statistiline protsessiohje (*statistical process control*, SPC) on metoodika, kus protsessi jälgimiseks kasutatakse spetsiaalseid diagramme, millele on kantud mõõdetava tunnuse mõõtmistulemused (näiteks toote kaal, mõõtmed, kliendi teenindamiseks kulunud aeg, arvete töötlemiseks kulunud aeg). Keskjoonele, mis vastab aritmeetilisele keskmisele, lisatakse ülemine ja alumine kontrolljoon, mille arvutamisel kasutatakse standardhälvet. Eesmärgiks on ennetada toodanguparameetrite väljumist lubatud piiridest[1].

3. Finantsanalüüsis kasutatakse dispersiooni ja standardhälvet investeeringu riski hindamisel. Dispersiooni kasutamise mõte riski mõõduna tugineb eeldusel, et mida hajuvamad on tulud, seda suurem on nende ebakindlus tulevikus. Finantsanalüüsis nimetatakse hajuvuse hindamist ka **volatiilsuse** (*volatility*) hindamiseks. (Sander, 1999)



Joonis 3.4. Kahe aktsia tulumäärad vahemikus 1. nov 2014 kuni 1. märts 2015. (a) *Procter & Gamble Company*, tulumäära standardhälve on 0,0087. (b) *ExxonMobil Corporation*, tulumäära standardhälve on 0,014. Viimase aktsia tulumäära volatiilsus on suurem

Eelnevad suurused olid kõik absoluutsed variatsioonnäitarvud. Nende abil ei saa võrrelda eri ühikutes mõõdetavate suuruste varieerumist. Näiteks kui töötajate palkade standardhälve on 400 eurot ja tööstaaži standardhälve 15 aastat, siis kumb suurus varieerub rohkem? Sellele küsimusele ei saa vastata, sest me ei saa omavahel võrrelda eurosid ja aastaid.

Ka samades ühikutes mõõdetud suuruste varieerumise võrdlemisel ei saa me alati otsustada vaid standardhälbe põhjal. Oletame, et kaks treialit treivad erinevaid detaile. Nende töö kvaliteedi hindamiseks

[1] Vt näiteks `http://qualityamerica.com/`

mõõdame detailid üle ja arvutame keskmise läbimõõdu $\bar{d}$ ning standardhälbe σ. Tulemuseks saame:

$$\begin{array}{lll} \text{treial A} & \bar{d} = 1 \text{ cm}, & \sigma = 0{,}1 \text{ cm}; \\ \text{treial B} & \bar{d} = 10 \text{ cm}, & \sigma = 0{,}1 \text{ cm}. \end{array}$$

Detailide standardhälbed on võrdsed, kuid kas me saame väita, et treialite töö kvaliteet on ühesugune, et hajumine detailide mõõtmetes on ühesugune? Ilmselt on treiali B töö kvaliteet parem.

3.4. Variatsioonikordaja

Eelmise alapeatüki lõpus toodud näites treialite ja nende töö kvaliteedi kohta pidime võrdlema standardhälvet ja aritmeetilist keskmist. Kui me tahame selle võrdlemise tulemust esitada ühe arvuna, siis sobib nende arvude suhe.

Variatsioonikordaja

Variatsioonikordaja (*coefficient of variation*) on standardhälbe σ ja aritmeetilise keskmise $\bar{x}$ suhe:

$$V = \frac{\sigma}{\bar{x}}. \tag{3.7}$$

Variatsioonikordaja on ilma ühikuteta **suhteline** variatsioonnäitarv. Variatsioonikordaja näitab, kui suure osa moodustab standardhälve aritmeetilisest keskmisest ning esitatatakse kas kümnendmurruna või protsendina.

Näide 3.3. Omavalitsuste tulud elaniku kohta ja nende varieerumine

2010. aastal oli kolmes maakonnas kohalike omavalitsuste tulu elaniku kohta järgmine[a]:

	Harju maakond	Tartu maakond	Pärnu maakond
Aritmeetiline keskmine, €	9272	5879	6193
Standardhälve, €	2089	1586	580

Millises maakonnas oli tulude varieerumine elaniku kohta kõige suurem? Leiame vastavad variatsioonikordajad.

$$\text{Harju} \quad V = \frac{2089}{9272} \approx 0{,}23,$$
$$\text{Tartu} \quad V = \frac{1586}{5879} \approx 0{,}27,$$
$$\text{Pärnu} \quad V = \frac{580}{6193} \approx 0{,}09.$$

Vastus: kõige suurem varieerumine oli Tartu maakonnas.

[a] Allikas: Rahandusministeeriumi veebileht `http://www.fin.ee/kov`

Tabelarvutuses variatsioonikordaja leidmiseks funktsiooni pole. Arvutamiseks kasutatakse valemit (3.7), kus standardhälve on leitud funktsiooniga STDEV.P ja aritmeetiline keskmine funktsiooniga AVERAGE.

3.5. Tšebõšovi teoreem

Varieerumise analüüsimisel on oluline teada, kui suur osa vaadeldava tunnuse väärtustest jääb aritmeetilisest keskmisest teatud kaugusele, täpsemalt mingisse vahemikku $(\bar{x} - a,\ \bar{x} + a)$. Tšebõšovi[2] teoreem võimaldab hinnata, kui suur osa väärtustest jääb standardhälbe kordsega määratud vahemikku.

Tšebõšovi teoreem

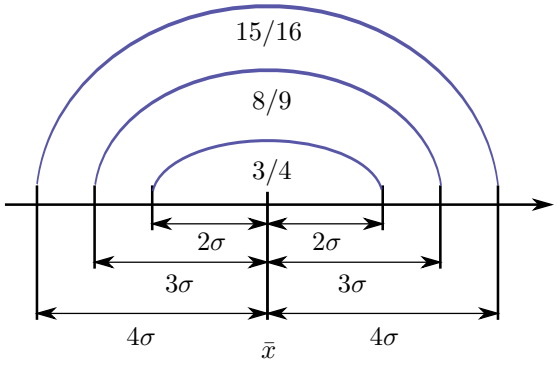
Tšebõšovi teoreem
Suvalise andmekogumi korral jääb vahemikku $\bar{x} \pm k\sigma$ vähemalt $1 - 1/k^2$ kõikidest väärtustest, kus k on ühest suurem positiivne arv.

Tšebõšovi teoreemi rakendamine $k = 2$, 3 ja 4 korral (vt ka joonis 3.5):

- vahemikku $\bar{x} \pm 2\sigma$ jääb vähemalt $1 - 1/2^2 = 3/4$, s.o 75% kõikidest väärtustest;
- vahemikku $\bar{x} \pm 3\sigma$ jääb vähemalt $1 - 1/3^2 = 8/9$, s.o ligikaudu 89% kõikidest väärtustest;
- vahemikku $\bar{x} \pm 4\sigma$ jääb vähemalt $1 - 1/4^2 = 15/16$, s.o ligikaudu 94% kõikidest väärtustest.

Rõhutada tuleb, et Tšebõšovi teoreem annab vastavasse vahemikku jäävate väärtuste osakaalu minimaalse väärtuse. Tegelikult võib selles

[2] P. L. Tšebõšov, XIX sajandi vene matemaatik.



Joonis 3.5. Tšebõšovi teoreemist tuleneb, et vähemalt 3/4 kõikidest väärtustest jääb vahemikku $\bar{x} \pm 2\sigma$, 8/9 jääb vahemikku $\bar{x} \pm 3\sigma$ ja 15/16 vahemikku $\bar{x} \pm 4\sigma$

vahemikus olla rohkem väärtusi. Osakaalu hinnangut saab täpsustada, kui me teame, millisele jaotusseadusele vaadeldav suurus allub. Näiteks normaaljaotusele alluva suuruse jaoks on erinevatele vahemikele vastavad osakaalud toodud alapeatükis 5.10 valem (5.82).

Näide 3.4. Poe päevakäivete jaotus ja Tšebõšovi teoreem

N03Varieerumine: N3.4

Alapeatükis 1 oli leheküljel 31 joonisel 1.6 toodud ühe poe käive päevas ajavahemikul 2. juuni kuni 30. detsember 1997, kokku 145 päeva. Leiame, kui suurel osal päevadest jäi käive vahemikku $\bar{x} \pm 2\sigma$, s.t võtame Tšebõšovi teoreemis oleva kordaja $k = 2$. Tšebõšovi teoreemi järgi peaks sellesse vahemikku jääma vähemalt 75% kõikidest väärtustest.
Päevakäivete aritmeetiline keskmine $\bar{x} \approx 15{,}82$ tuhat eurot ja standardhälve $\sigma \approx 10{,}30$ tuhat eurot. Vahemiku alumine piir on

$$5{,}83 - 2 \cdot 10{,}30 \approx -4{,}8$$

ja ülemine piir

$$15{,}83 + 2 \cdot 10{,}30 \approx 36{,}4$$

tuhat eurot. Kuna käive ei saa olla negatiivne, siis tuleb leida, mitmel päeval jäi käive vahemikku 0–36,4 tuhat eurot. Loendamine näitab, et selliseid päevi on 137, mis moodustab päevade koguarvust ligikaudu 94%. Tegelik osakaal on Tšebõšovi teoreemi järgi leitust suurem.

Mõnikord huvitab meid, kui suur osa väärtustest jääb vahemikust $\bar{x} \pm k\sigma$ välja. Ka selle hindamiseks võib kasutada Tšebõšovi teoreemi.

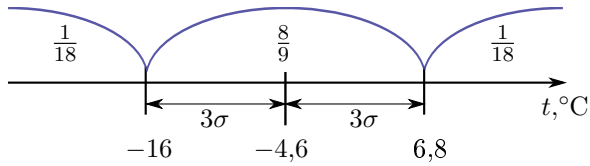
Näide 3.5. Veebruarikuu keskmine õhutemperatuur

Tartu Ülikooli füüsikahoone katusel asub e-ilmajaam, mis mõõdab pidevalt õhutemperatuuri, õhuniiskust, õhurõhku ja muid ilma parameetreid[a]. Mõõtmisi on tehtud alates 1999. aasta novembrist.
Selle ilmajaama andmetel on veebruarikuu keskmise õhutemperatuuri aritmeetiline keskmine aastatel 2000–2015 olnud $-4{,}6\,^\circ\mathrm{C}$ standardhälbega $3{,}8\,^\circ\mathrm{C}$. Mitmel aastal sajast võib veebruarikuu keskmine õhutemperatuur olla madalam kui $-16\,^\circ\mathrm{C}$?
Hindamiseks kasutame Tšebõšovi teoreemi. Algul leiame, mitme standardhälbe kaugusel on väärtus -16 aritmeetilisest keskmisest $-4{,}6$:

$$k = \frac{|t_i - \bar{t}|}{\sigma} = \frac{|-16 - (-4{,}6)|}{3{,}8} = \frac{11{,}4}{3{,}8} = 3. \tag{3.8}$$

Tšebõšovi teoreemi järgi jääb vähemalt 8/9 kõikidest väärtustest vahemikku $\bar{t} \pm 3\sigma$. Järelikult maksimaalselt 1/9 kõikidest väärtustest on aritmeetilisest keskmisest kaugemal kui 3 standardhälvet. Need jagunevad kaheks osaks: väiksemad kui $\bar{t} - 3\sigma$ ning suuremad kui $\bar{t} + 3\sigma$ (vt ka joonist):

1) $t_i < -16\,^\circ\mathrm{C}$;
2) $t_i > 6{,}8\,^\circ\mathrm{C}$.



Meid rahuldab tingimus 1). Eeldame, et veebruarikuu keskmiste temperatuuride jaotus on sümmeetriline. Siis jagunevad vahemikust $\bar{x} \pm 3\sigma$ välja jäävad väärtused t_i kaheks võrdseks osaks ning tingimust 1) rahuldavate väärtuste osakaal on pool 1/9-st:

$$\frac{1}{2} \cdot \frac{1}{9} = \frac{1}{18} \approx 0{,}056.$$

Järelikult saja aasta kohta maksimaalselt viiel-kuuel aastal võib veebruarikuu keskmine õhutemperatuur olla madalam kui $-16\,^\circ\mathrm{C}$. Seda muidugi eeldusel, et ei toimu globaalset kliima soojenemist või jahtumist.

[a] `http://meteo.physic.ut.ee/`

3.6. Standardiseeritud skaala

Näites 3.5 tuli leida, mitme standardhälbe kaugusel asus vaadeldav väärtus aritmeetilisest keskmisest (valem (3.8)). Tšebõšovi teoreemi kasutamisel ongi otstarbekas leida kauguse $|x_i - \bar{x}|$ suhe standardhälbesse. Ei pea aga kasutama absoluutväärtust, võib leida lihtsalt hälbe $x_i - \bar{x}$ ja standardhälbe σ suhte. See suhe annab lisaks kaugusele ka informatsiooni, kummal pool aritmeetilist keskmist vaadeldav väärtus asub.

Standardiseeritud väärtus

Standardiseeritud väärtus näitab, mitmekordse standardhälbe σ kaugusel aritmeetilisest keskmisest $\bar{x}$ asub vaadeldav väärtus x_i:

$$z_i = \frac{x_i - \bar{x}}{\sigma}. \tag{3.9}$$

Mõnikord nimetatakse standardiseeritud väärtust ka z-väärtuseks või z-skooriks (z-*score*). Standardiseeritud väärtus on ühikuta suurus. Seetõttu võimaldab see analüüsida mingi elemendi asukohta kogumis erinevates ühikutes mõõdetud tunnuste korral.

Näide 3.6. Eesti SKP elaniku kohta ja tööjõu tootlikkus võrreldes teiste riikidega

Tabelis on toodud näitajate SKP elaniku kohta (tuhat eurot) ning tööjõu tootlikkus (eurot töötunni kohta) aritmeetiline keskmine ja standardhälve Euroopa riikides aastal 2011[a]. Lisatud on Eesti vastavad näitajad.

	SKP elaniku kohta, tuh €	Tööjõu tootlikkus, € töötunni kohta
Aritmeetiline keskmine	22,1	29,2
Standardhälve	15,5	18,1
Eesti väärtused	9,1	10,8

Mõlema tunnuse korral on Eesti näitaja Euroopa keskmisest väiksem. SKP elaniku kohta on Eestis väiksem 13 tuhande euro võrra ja tootlikkus on väiksem 18,4 eurot töötunni kohta. Kui tahame analüüsida, kumb näitaja on Eestis halvem kui Euroopa keskmine, siis neid arve me võrrelda ei saa, sest ühikud on

erinevad. Võrdlemiseks leiame Eesti jaoks standardväärtused:

$$\text{SKP elaniku kohta} \quad z = \frac{9{,}1 - 22{,}1}{15{,}5} \approx -0{,}84,$$

$$\text{tööjõu tootlikkus} \quad z = \frac{10{,}8 - 29{,}2}{18{,}1} \approx -1{,}02.$$

Standardväärtusi saame omavahel võrrelda, sest need on ühikuta. Näeme, et tööjõu tootlikkusega oleme Euroopa keskmisest rohkem maas kui ühe elaniku kohta leitud SKP väärtusega.

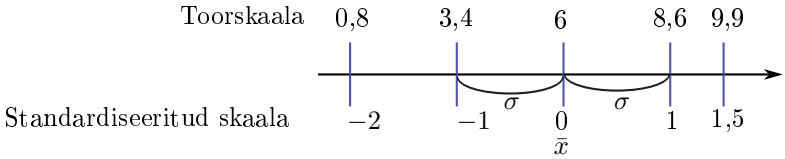
[a] Allikas: Eurostat.

Tabelarvutuses leiab standardiseeritud väärtuse funktsioon **STANDARDIZE**. Ette tuleb anda suuruse X väärtus, mille standardiseeritud väärtust soovitakse leida, aritmeetiline keskmine (*Mean*) ja standardhälve (*Standard_ dev*).

Tunnuse väärtuste esitamiseks võib kasutada z-skaalat ehk standardiseeritud skaalat (joonis 3.6). Väärtuste teisendamiseks töötlemata skaalalt ehk toorskaalalt (*raw scale*) standardiseeritud skaalale kasutatakse eespool toodud valemit (3.9).

Standardiseeritud skaala

Standardiseeritud skaalal on esitatud standardiseeritud väärtused (3.9). Kõigi standardiseeritud väärtuste aritmeetiline keskmine on 0 ja standardhälve 1.



Joonis 3.6. Toorskaala ($\bar{x} = 6$, $\sigma = 2{,}6$) ja standardiseeritud skaala

Kasutades standardiseeritud väärtust z, võib järeldused Tšebõšovi teoreemist formuleerida järgmiselt:

- vähemalt 3/4 ehk 75% väärtuste korral $|z| \leq 2$;
- vähemalt 8/9 ehk ligikaudu 89% väärtuste korral $|z| \leq 3$;
- vähemalt 15/16 ehk ligikaudu 94% väärtuste korral $|z| \leq 4$.

3.7. Jaotuse kuju iseloomustavad näitajad

Asümmeetria on jaotuskõvera maksimumi kõrvalekaldumine sümmeetriateljest. Kui jaotuskõvera maksimum (mood) on sümmeetriateljest

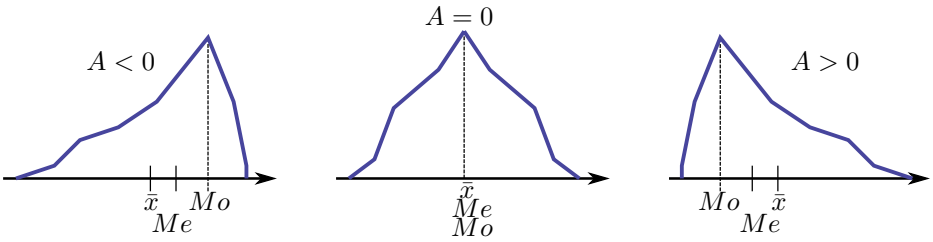
(mediaan) paremal, on tegemist negatiivse ehk vasakkaldelise asümmeetriaga (joonis 3.7 vasakul). Nimetus tuleneb sellest, et „saba“ on vasakul pool. Kui maksimum on sümmeetriateljest vasakul pool, jääb saba paremale ja tegemist on positiivse ehk paremkaldelise asümmeetriaga (joonis 3.7 paremal). Sümmeetrilise jaotuse korral langevad mediaan ja mood kokku. Kõrvalekalde suuruse iseloomustamiseks kasutatakse asümmeetriakordajat.

Asümmeetriakordaja

Arvukogumi asümmeetriat iseloomustab **asümmeetriakordaja** (*coefficient of skewness*)

$$A = \frac{1}{n\sigma^3}\sum(x_i - \bar{x})^3, \tag{3.10}$$

kus n on kogumi maht, σ standardhälve ja $\bar{x}$ aritmeetiline keskmine.



Joonis 3.7. Negatiivse asümmeetriaga, sümmeetriline ja positiivse asümmeetriaga jaotus

Valemist (3.10) on näha, et kui esineb ekstremaalselt väikesi väärtusi x_i, mis asuvad aritmeetilisest keskmisest kaugel vasakul, siis nende korral $x_i - \bar{x} << 0$ ja need muudavad asümmeetriakordaja negatiivseks. Ekstremaalselt suurte väärtuste x_i korral $x_i - \bar{x} >> 0$ ning asümmeetriakordaja A muutub positiivseks.

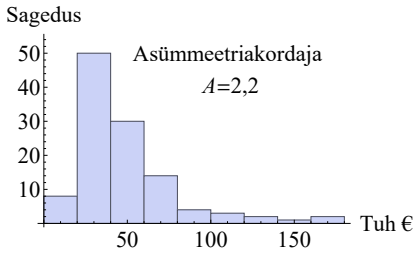
Tabelarvutuses leiab asümmeetriakordaja funktsioon **SKEW**.

Joonisel 3.8 on sõiduauto Audi 114 erinevate mudelite hindade jaotus auto24.ee portaalis 6. veebruari 2015. a seisuga[3]. Hindade jaotus on positiivse asümmeetriaga, mis tähendab, et üksikute mudelite hinnad on eriti kõrged. Audi hind ei saa olla väga madal, küll aga võib olla väga kõrge. Joonisel 3.9 on meeste surmad Eestis aastal 2013 rühmitatud vanuse järgi[4]. Jaotus on negatiivse asümmeetriaga.

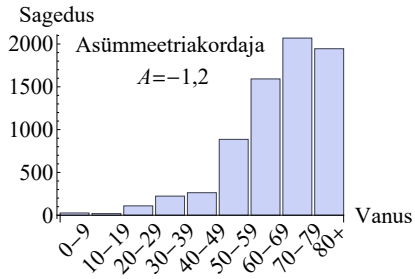
[3] `http://www.auto24.ee/`

[4] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel RV45: surnud soo ja vanuserühmade järgi.

N03Varieerumine
J3.8



Joonis 3.8. Audi mudelite hindade jaotus. Kokku 114 erinevat mudelit



Joonis 3.9. Meeste surmad Eestis 2013. aastal, rühmitatud vanuse järgi

Tugevalt asümmeetrilised jaotused esinevad tavaliselt siis, kui uuritaval suurusel on kas alumine või ülemine piir. Alumise piiri korral tekib positiivne asümmeetria ja ülemise piiri korral negatiivne asümmeetria.

Teine näitaja, mis iseloomustab jaotuse kuju, on püstakus.

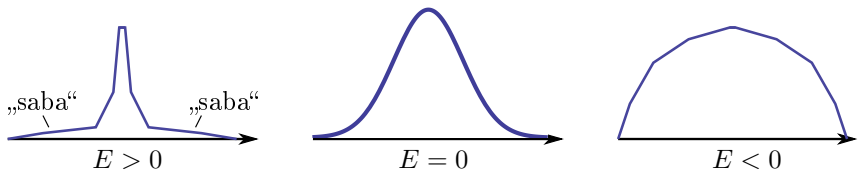
Püstakuse kordaja

Jaotuse püstakust iseloomustab **ekstsess** (*excess kurtosis*), mida võib nimetada ka **püstakuse kordajaks**:

$$E = \frac{1}{n\sigma^4}\sum(x_i - \bar{x})^4 - 3, \tag{3.11}$$

kus n on kogumi maht, σ standardhälve ja $\bar{x}$ aritmeetiline keskmine.

Püstakuse kordaja on null normaaljaotuse korral. Suurem püstakus tähendab, et enamik väärtusi on koondunud aritmeetilise keskmise ümber. Kaugemal asuvate väärtuste esinemissagedus on väike ning esinevad „sabad“ (joonis 3.10). Väikese püstakuse korral „sabad“ kaovad ja erinevate väärtuste esinemissagedus väga palju ei erine, jaotus on lamedam. Päris lamedaid jaotusi, kus kõikide väärtuste esinemissagedus on ligikaudu ühesugune, kohtab harva.



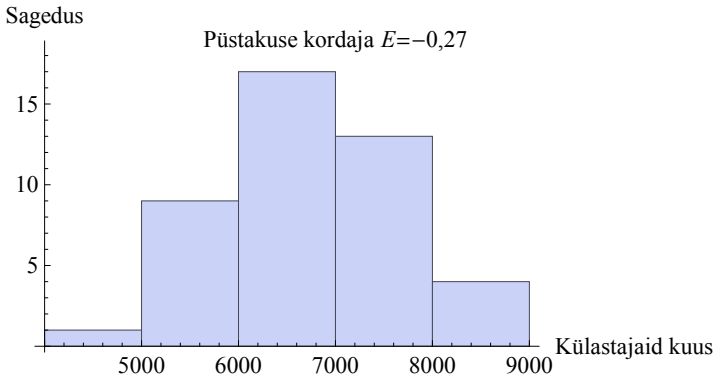
Joonis 3.10. Positiivse püstakusega ehk püstakas jaotus ($E > 0$), normaaljaotus ($E = 0$) ja negatiivse püstakusega ehk lame jaotus ($E < 0$).

Tabelarvutuses on püstakuse kordaja leidmiseks funktsioon **KURT**.

Valemiga (3.11) defineeritud püstakuse kordajat nimetatakse Fisheri[5] ekstsessikordajaks. Tihti kasutatakse ka Pearsoni[6] ekstsessikordajat, mille valem sarnaneb valemile (3.11), kuid arv 3 ei ole maha lahutatud:

$$E_P = \frac{1}{n\sigma^4}\sum(x_i - \bar{x})^4. \tag{3.12}$$

Selle väärtus on normaaljaotuse korral 3, püstaka jaotuse korral > 3 ning lameda jaotuse korral < 3. Inglise keeles on selle suuruse nimetus lihtsalt (*kurtosis*).



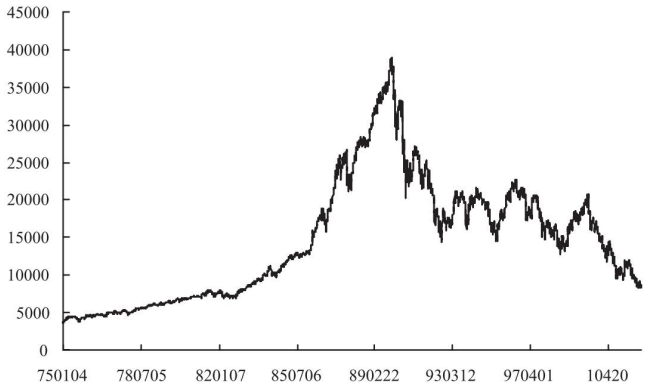
Joonis 3.11. Ühe Tallinna pangakontori külastajate arv ühes kuus jaanuar 1999 kuni august 2002

Joonistel 3.8 ja 3.9 toodud jaotused on mõlemad püstakad. Audi hindade korral on püstakuse kordaja $E = 5{,}6$ ja surmade vanuselise jaotuse korral $E = 1{,}47$, mis tähendab, et viimase püstakus on väiksem. Joonisel 3.11 on toodud ühe Tallinnas asuva pangakontori külastajate arv kuus ajavahemikus jaanuar 1999 kuni august 2002 (Antin, 2002). Selle jaotuse püstakuse kordaja on $E = -0{,}27$, mis tähendab, et jaotus on normaaljaotusest lamedam.

Näide 3.7. Börsiindeksi Nikkei tulumäära jaotus

Taisei Kaizoi analüüsis Tokio börsi indeksi Nikkei 225 kõikumisi aastatel 1975–2002 (Kaizoji, 2004). Selle perioodi võis jagada kaheks: aastatel 1975–1989 indeks kasvas ja 1990–2002 indeksi väärtus kahanes (joonis 3.12).

[5]Sir Ronald Aymler Fisher (1890–1962), inglise matemaatik ja statistik.
[6]Karl Pearson (1857–1936), inglise matemaatik.



Joonis 3.12. Indeksi Nikkei 225 muutumine 1975–2002. Horisontaalteljel olevad kuupäevad on formaadis *aakkpp*.

Indeksi tulumäära iseloomustavad kirjeldava statistika suurused kummalgi perioodil on toodud järgmises tabelis.

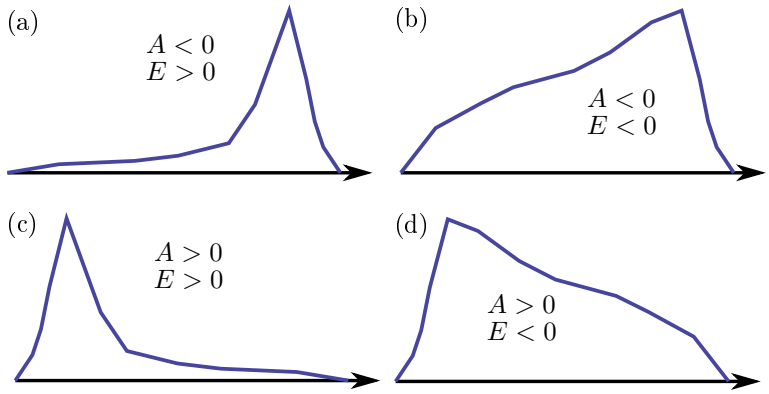
Periood	Aritmeetiline keskmine	Standardhälve	Asümmeetria kordaja	Püstakuse kordaja
Kasvuperiood (1975–1989)	0,0006	0,007	−0,3	6,5
Kahanemisperiood (1990–2002)	−0,0005	0,015	0,3	3,1

On näha, et kasvuperioodil oli indeksi tulumääral negatiivne asümmeetria, mis tähendab, et oli üksikuid päevi, kus tulumäär oli väga väike. Kahanemisperioodil oli tulumääral positiivne asümmeetria, mis tähendab, et esines üksikuid päevi, mil tulumäär oli väga suur. Püstakus oli väiksem kahanemisperioodil, järelikult sel ajal esines ekstreemseid väärtusi väiksema tõenäosusega kui kasvuperioodil, sest jaotus on lamedam.

Asümmeetrilised jaotused võib kalde ja püstakuse alusel jagada nelja gruppi:

a) vasakpoolse asümmeetriaga ja püstakas, saba on vasakul ning õhuke: ekstreemsed väärtused on vasakul ja neid on vähe (joonis 3.13 (a));

b) vasakpoolse asümmeetriaga ja lame, saba on vasakul ning paks: ekstreemsed väärtused on vasakul ja neid on palju (joonis 3.13 (b));

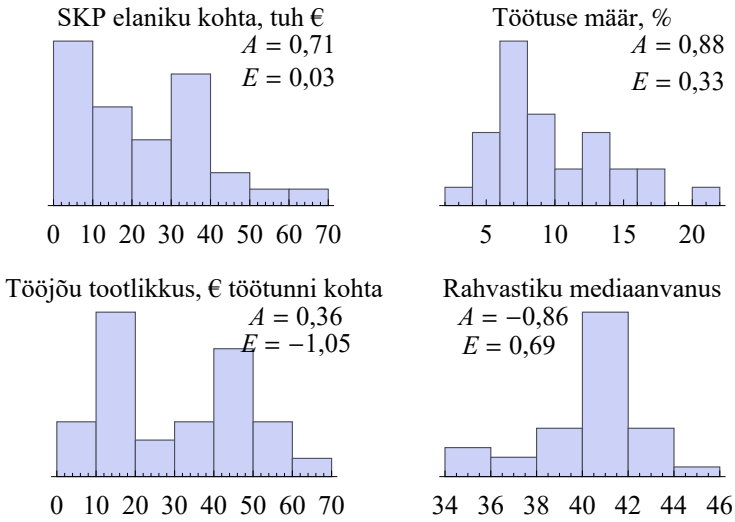


Joonis 3.13. Asümmeetria ja püstakuse kombinatsioonid

c) parempoolse asümmeetriaga ja püstakas, saba on paremal ning õhuke: ekstreemsed väärtused on vasakul ja neid on vähe (joonis 3.13 (c));
d) parempoolse asümmeetriaga ja lame, saba on paremal ning paks: ekstreemsed väärtused on paremal ja neid on palju (joonis 3.13 (d)).

Näide 3.8. Euroopa riikide näitajad

Järgnevalt on esitatud nelja tunnuse jaotus Euroopa riikides aastal 2011[a].



SKP elaniku kohta ja töötuse määr on mõlemad positiivse asümmeetriaga, kuid SKP elaniku kohta on lamedam, mis tähendab, et sellel näitajal on erinevate väärtuste esinemine ühtlasem. Nõrk

positiivne asümmeetria on ka tööjõu tootlikkusel ja see on nende näitajate hulgas kõige lamedam. Negatiivse asümmeetriaga on rahvastiku mediaanvanus. Järelikult esineb riike, kus rahvastik on oluliselt noorem, võrreldes keskmisega, kuid neid riike pole palju (püstakuse kordaja on positiivne).
Tööjõu tootlikkust võib nimetada bimodaalseks, sest selgelt eristuvad kaks tippu. See tähendab, et riigid võib jagada kaheks rühmaks, kus ühes rühmas on keskmine tootlikkus suurem ja teises väiksem. Selline multimodaalsus vähendabki püstakuse kordajat E ja muudab jaotuse lamedamaks. Ka näitajal „SKP elaniku kohta“ on kaks tippu, kuid sellel on kaks alarühma vähem eristatavad: tippudevahelised tulbad on kõrgemad.

[a] Allikas: Eurostat.

Kui standardhälvet võib leida ka kolme väärtuse korral, siis asümmeetria- ja püstakuse kordaja väärtusi on mõtet leida vaid suurte kogumite korral. Suureks võib lugeda kogumit, mille maht $n > 50$.

Uuritavat jaotust kirjeldavate statistiliste näitajate leidmiseks võib programmis Excel kasutada andmeanalüüsi vahendit *Descriptive Statistics* komplektist *Data Analysis*. Selle abil saab korraga leida kogumi mahu, aritmeetilise keskmise, moodi, mediaani, maksimumi, miinimumi, standardhälbe, asümmeetriakordaja ning püstakuse kordaja (vt ka lisa C.3).

3.8. Statistilised momendid

Eespool vaadeldud statistilised keskmised, hajuvuse ja kuju karakteristikud võib võtta ühe nimetuse alla: statistilised momendid.

Moment

Tunnuse k-ndat järku **moment** väärtuse a suhtes on väärtuste x_i ja arvu a vaheliste hälvete k-ndat järku astmete aritmeetiline keskmine:

$$M_k = \frac{\sum (x_i - a)^k}{n}, \tag{3.13}$$

kus n on väärtuste arv.

Tavaliselt vaadeldakse positiivse täisarvulise astendajaga momente ning piirdutakse esimese nelja momendiga. Momendid jagunevad alg-, kesk- ja tingmomentideks:

1) **algmomendid**, kui $a = 0$:

$$m_k = \frac{\sum x_i^k}{n}; \tag{3.14}$$

2) **keskmomendid**, kui $a = \bar{x}$:

$$\mu_k = \frac{\sum (x_i - \bar{x})^k}{n}; \tag{3.15}$$

3) **tingmomendid**, kui a on suvaline arv.

Näiteks aritmeetiline keskmine on 1. järku algmoment, dispersioon on 2. järku keskmoment. Kolmandat järku keskmoment on valemi (3.15) põhjal

$$\mu_3 = \frac{\sum (x_i - \bar{x})^3}{n}. \tag{3.16}$$

Kui me võrdleme seda avaldist asümmeetriakordaja valemiga (3.10), siis näeme, et asümmeetriakordaja on avaldatav kolmandat järku keskmomendi μ_3 kaudu:

$$A = \frac{\mu_3}{\sigma^3}. \tag{3.17}$$

Ekstsess ehk püstakuse kordaja on avaldatav neljandat järku keskmomendi μ_4 abil:

$$E = \frac{\mu_4}{\sigma^4} - 3. \tag{3.18}$$

Mingit järku keskmomendi leidmiseks saab kasutada algmomente. Näiteks dispersiooni arvutamiseks kasutatakse tihti dispersiooni definitsioonvalemist tuletatud mugavamat arvutusvalemit:

$$\sigma^2 = \mu_2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2 = m_2 - m_1^2,$$

kus $m_1 = \sum x_i/n$ on esimest järku algmoment ja $m_2 = \sum x_i^2/n$ teist järku algmoment.

Vaatame 2. järku tingmomenti

$$M(a) = \frac{\sum (x_i - a)^2}{n}. \tag{3.19}$$

See iseloomustab varieerumist ümber väärtuse a. Leiame, millise a korral on $M(a)$ minimaalne. Funktsiooni $M(a)$ minimeerimiseks leiame tuletise suuruse a järgi ja paneme selle võrduma nulliga:

$$\frac{dM(a)}{da} = \frac{d}{da}\frac{1}{n}\sum (x_i - a)^2 = \frac{1}{n}\sum \frac{d}{da}(x_i - a)^2 =$$
$$= \frac{1}{n}\sum 2(x_i - a)(-1) = -2\frac{1}{n}\sum (x_i - a).$$

Et tuletis oleks null, peab paremal pool olev summa võrduma nulliga:

$$\begin{aligned}\sum(x_i - a) &= 0\\ \sum x_i - na &= 0\\ \sum x_i &= na\\ \frac{1}{n}\sum x_i &= a.\end{aligned}$$

Vasakul pool võrdusmärki on aritmeetilise keskmise arvutusvalem.

Aritmeetilise keskmise tähendus

Aritmeetiline keskmine on tunnuse selline väärtus, mille suhtes on varieerumine (ruutkeskmise hälbe mõttes) kõige väiksem.

3.9. Kaheväärtuselise tunnuse standardhälve

Kaheväärtuseliseks tunnuseks (ka binaarne või dihhotoomne) nimetatakse tunnust, millel võivad olla vaid väärtused „0“ ja „1“. Nimiskaalas mõõdetud kaheväärtuselised tunnused kodeeritakse nullide ja ühtedega. Näiteks „sugu“ on kaheväärtuseline tunnus. Kodeerimisel võib võtta väärtuse mees koodiks „0“ ja väärtuse „naine“ koodiks „1“ või vastupidi. See, kumb väärtus kodeeritakse nulliga ja kumb ühega, ei oma tähtsust. Kaheväärtuseliseks tunnuseks võib olla ka vastus küsimusele, millele saab vastata kas „jaa“ või „ei“. Näiteks „Kas te pooldate astmelist tulumaksu?“, „Kas teil on eramu?“, „Kas te tarbite toodet X?“.

Kaheväärtuselise tunnuse korral on summa üle kõikide väärtuste x_i võrdne ühtede arvuga selles kogumis. Kui m on ühtede arv kogumis, siis $m = \sum x_i$ ja aritmeetiline keskmine on võrdne väärtuse „1“ esinemise suhtelise sageduse ehk osakaaluga p.

Kaheväärtuselise tunnuse aritmeetiline keskmine

Kaheväärtuselise tunnuse $\{0, 1\}$ aritmeetiline keskmine on

$$\bar{x} = p = \frac{m}{n}, \tag{3.20}$$

kus n on kogumi maht, m ühtede arv ja p ühtede osakaal kogumis.

Leiame kaheväärtuselise tunnuse dispersiooni:

$$\sigma^2 = \frac{\sum\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)}{n} = \frac{\sum x_i^2 - \sum 2x_i\bar{x} + \sum \bar{x}^2}{n}. \tag{3.21}$$

Vaatleme lugejas olevaid liikmeid eraldi. Kuna $1^2 = 1$ ja $0^2 = 0$, siis

$$\sum x_i^2 = m. \tag{3.22}$$

Valemi (3.21) lugejas oleva teise liikme jaoks saame seoseid (3.22) ja (3.20) kasutades

$$\sum 2x_i\bar{x} = 2\bar{x}\sum x_i = 2pm. \tag{3.23}$$

Valemi (3.21) lugejas oleva kolmanda liikme korral arvestame, et summeerimisel tuleb üksteisele liita n suurust $\bar{x}^2$:

$$\sum \bar{x}^2 = n\bar{x}^2 = np^2. \tag{3.24}$$

Kasutades seoseid (3.22)–(3.24), saame dispersiooni (3.21) jaoks

$$\sigma^2 = \frac{\sum x_i^2 - \sum 2x_i\bar{x} + \sum \bar{x}^2}{n} = \frac{m - 2pm + np^2}{n} = \frac{m}{n} - 2p\frac{m}{n} + \frac{np^2}{n} =$$
$$= p - 2p^2 + p^2 = p - p^2 = p(1-p).$$

Kui p on ühtede osakaal kogumis, siis $1-p$ on nullide osakaal. Kui me kodeerime ümber, s.t vahetame nullid ja ühed, siis ühtede osakaal on $1-p$ ja nullide osakaal p ning korrutiseks saame $(1-p)p$, mis võrdub korrutisega $p(1-p)$. Dispersioon ei sõltu sellest, kummale väärtusele me omistame koodi „0“ ja kummale koodi „1“.

Kaheväärtuselise tunnuse dispersioon ja standardhälve

Kaheväärtuselise tunnuse **dispersioon**

$$\sigma^2 = p(1-p) \tag{3.25}$$

ja **standardhälve**

$$\sigma = \sqrt{p(1-p)}, \tag{3.26}$$

kus p on üht väärtust omavate objektide osakaal kogumis.

Näide 3.9. Töötuse määra standardhälve

Töötuse määr ehk tööpuuduse määr on töötute osatähtsus tööjõus. Eestis oli 2014. aastal töötuse määr 15–74-aastaste meeste hulgas 7,9% ja naiste hulgas 6,8%[a]. Leiame töötuse määra standardhälbed.
Meestel

$$\sigma = \sqrt{0{,}079 \cdot (1 - 0{,}079)} \approx 0{,}270.$$

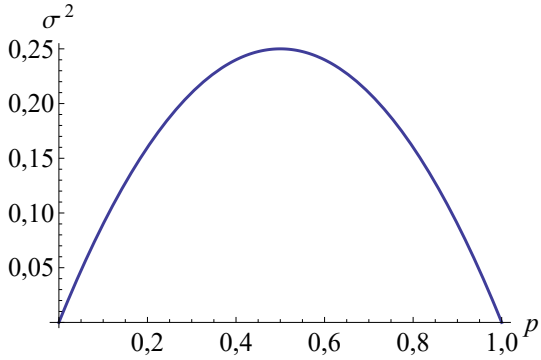
Naistel

$$\sigma = \sqrt{0{,}068 \cdot (1 - 0{,}068)} \approx 0{,}252.$$

Naistel on standardhälve ja seega varieerumine väiksem kui meestel, sest naiste kogum on homogeensem, seal on töötute osakaal väiksem.

[a] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel TT35: töötuse määr soo ja vanuserühma järgi.

Kaheväärtuselise tunnuse dispersiooni ja standardhälbe minimaalne väärtus on 0, kui $p = 0$ või kui $p = 1$. Sellisel juhul on kõigil kogumi objektidel ühesugune väärtus ning kogum on homogeenne. Milline on dispersiooni maksimaalne väärtus? Ilmselt siis, kui erinevaid väärtusi on täpselt ühepalju. Seda saab kontrollida arvuliselt (joonis 3.14) ning näidata matemaatiliselt.



Joonis 3.14. Kaheväärtuselise tunnuse dispersiooni σ^2 sõltuvus ühe väärtuse osakaalust p

Funktsiooni maksimumkoha leidmiseks tuleb võtta funktsioonist tuletis ja panna see võrduma nulliga. Vaatleme dispersiooni (3.25) funktsioonina osakaalust p ning leiame tuletise p järgi:

$$\big(\sigma^2(p)\big)' = \big(p - p^2\big)' = 1 - 2p.$$

Dispersiooni tuletis võrdub nulliga siis, kui

$$p = 0{,}5.$$

See on dispersiooni maksimumkoht. Dispersiooni maksimum on

$$\sigma^2_{\max} = 0{,}5 \cdot 0{,}5 = 0{,}25$$

ja standardhälbe maksimum

$$\sigma_{\max} = \sqrt{0{,}25} = 0{,}5.$$

3.10. Varieeruvuse hindamine asendikeskmiste abil

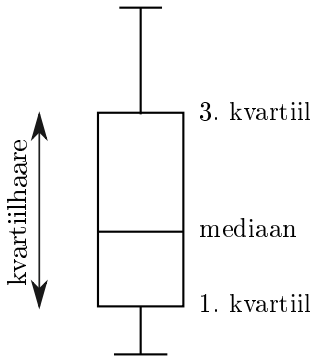
Nii nagu aritmeetiline keskmine, nii ka standardhälve sõltub igast üksikust väärtusest. Kui mõni väärtus on aritmeetilisest keskmisest väga kaugel, mõjutab see kohe ka standardhälvet. Kui me ei soovi, et üksikud ekstremaalsed väärtused mõjutaksid varieerumise hinnangut, kasutatakse varieerumise hindamiseks asendikeskmisi: kvartiile ja detsiile.

Kvartiilhaare

Kvartiilhaare (*interquartile range*) on kolmanda kvartiili Q_3 ja esimese kvartiili Q_1 vahe:

$$IQR = Q_3 - Q_1. \tag{3.27}$$

Kvartiilhaarde sisse jääb alati 50% variatsioonrea väärtustest.



Joonis 3.15. Karpdiagrammil on kvartiilhaare võrdne karbi kõrgusega

Näide 3.10. Keskmise kuupalga kvartiilhaare

Tabelis on keskmine kuupalk (eurot) Eesti ettevõtetes aastatel 2010–2013[a]. Tabeli viimases reas on leitud kvartiilhaare.

	2010	2011	2012	2013
3. kvartiil	630	670	750	770
Mediaan	350	380	410	430
1. kvartiil	230	240	260	280
Kvartiilhaare	400	430	490	490

On näha, et aastatel 2010–2012 keskmise palga kvartiilhaare suurenes, seega suurenesid ka erinevused palkades. Aastal 2013

aga tõusid esimene ja kolmas kvartiil ühepalju ning kvartiilhaare jäi samaks.

[a] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel EM024: ettevõtete asendikeskmised suhtarvud (kvartiilid, mediaan) tegevusala järgi.

Detsiilhaare, suhteline detsiilhaare

Lisaks kvartiilhaardele kasutatakse mõnikord ka **detsiilhaaret** (*interdecile range*), mis on üheksanda detsiili ja esimese detsiili vahe, ning suhtelise variatsioonnäitarvuna **suhtelist detsiilhaaret** (*interdecile ratio*), mis on detsiilhaare jagatud viienda detsiiliga ehk mediaaniga.

Näide 3.11. Sissetulekute võrdlus

Võrdleme sissetulekute jaotust neljas Euroopa riigis, kasutades Eurostati andmeid aastast 2001[a]. Tabelis on toodud ühe inimese kohta tuleva sissetuleku 1., 5. ja 9. detsiil (eurot), mille põhjal on leitud detsiilhaare ja suhteline detsiilhaare. Millises riigis on ebavõrdsus kõige suurem?

	Belgia	Kreeka	Hispaania	Portugal
1. detsiil, €	4840	1498	1767	1260
5. detsiil, €	9436	4292	5185	3460
9. detsiil, €	28 618	13 234	16 783	14 175
Detsiilhaare, €	23 778	11 736	15 016	12 915
Suhteline detsiilhaare	2,5	2,7	2,9	3,7

Kui kasutada otsustamiseks detsiilhaaret, siis see on kõige suurem Belgias. Kuid suhteline detsiilhaare on kõige suurem Portugalis ning selle põhjal võib väita, et kõige suurem ebavõrdsus sissetulekute jaotuses esineb seal.

[a] Allikas: Eurostat.

3.11. Sagedusklasside arvu sõltuvus tunnuse varieerumisest

Alapeatükis 1.6 vaatasime tunnuse väärtuste intervallimist ehk klassidesse jaotamist. Üheks probleemiks oli sobiva klasside arvu leidmine ning seal oli soovitatud selleks kasutada Sturgesi valemit, kus klasside arv sõltub kogumi mahust n.

Kuid mida enam vaadeldav tunnus varieerub, seda rohkem on sagedusklasse vaja, et saadav sagedustabel kajastaks kirjeldatava kogumi reaalset struktuuri. Seepärast on välja pakutud mitmed alternatiivsed valemid klassi laiuse määramiseks.

Scotti valem optimaalse klassi laiuse määramiseks (D. W. Scott, 1979):

$$d = \frac{3{,}5\sigma}{\sqrt[3]{n}}, \tag{3.28}$$

kus σ on standardhälve ja n kogumi maht. Seda kasutatakse näiteks histogrammi konstrueerimisel programmi Excel andmeanalüüsivahendiga *Histogram*, mis asub komplektis *Data Analysis*.

Freedman ja Diaconis (1981) soovitasid hajumise arvesse võtmiseks kasutada kvartiilhaaret:

$$d = \frac{2IQR}{\sqrt[3]{n}}, \tag{3.29}$$

kus IQR on kvartiilhaare ja n kogumi maht.

3.12. Ülesanded

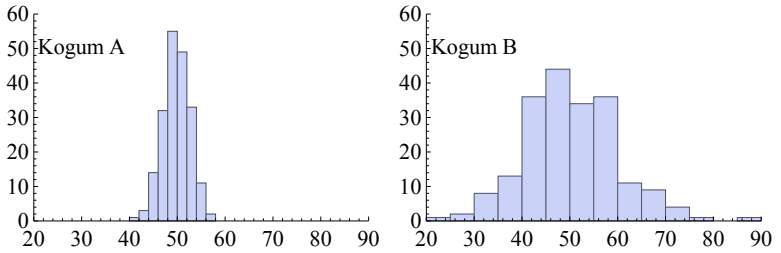
3.1. Olgu meil viis erinevat arvu: 1, 2, 5, 7 ja 10. Leida *Standardhälve*

a) keskmine hälve $\frac{1}{n}\sum(x_i - \bar{x})$;
b) keskmine absoluuthälve;
c) standardhälve.

VASTUS lk 661.

3.2. Arvukogumi dispersioon on 25. Kui suur on nende arvude standardhälve? VASTUS lk 661.

3.3. Joonisel 3.16 on toodud kahe arvukogumi histogramm. Kumma kogumi standardhälve on suurem? VASTUS lk 661.



Joonis 3.16. Kahe kogumi histogramm, ülesanne 3.3

3.4. Olgu meil kolm arvukogumit:
A={8, 9, 10, 11, 12};
B={10, 10, 10, 10, 10};

C={5, 7, 10, 13, 15}.
Hinnata ilma arvutamata, millise arvukogumi standardhälve on kõige suurem ja millisel kõige väiksem. VASTUS lk 661.

3.5. Arvukogumi aritmeetiline keskmine on $\bar{x}$ ja standardhälve σ. Kuidas aritmeetiline keskmine ja standardhälve muutuvad, kui
a) igale väärtusele liidetakse üks ja sama arv a;
b) igat väärtust korrutatakse ühe ja sama arvuga b?
VASTUS lk 661.

Variatsioonikordaja

3.6. Aastatel 2004–2009 oli autotootja BMW keskmine toodang 1,36 miljonit autot aastas standardhälbega 0,10 miljonit autot (BMW, 2013). Kulud aastas olid keskmiselt 37,4 miljardit eurot standardhälbega 2,38 miljardit eurot. Kumb suurus varieerus rohkem? VASTUS lk 661.

3.7. 2015. aastal tehti Tallinnas 8770 korteri ostu-müügi tehingut keskmise hinnaga 1548,33 eurot ruutmeetri kohta[7]. Korteritel pindalaga 10–29,99 m^2 oli keskmine pinnaühiku hind 1481,70 €/m^2 standardhälbega 599,41 €/m^2 ja korteritel pindalaga üle 70 m^2 oli pinnaühiku keskmine hind 1789,50 €/m^2 standardhälbega 703,22 €/m^2. Kas pinnaühiku hind varieerus rohkem suurematel või väiksematel korteritel? VASTUS lk 661.

3.8. 2010. aastal oli ettevõtte kasumi standardhälve 0,40 mln krooni ja variatsioonikordaja 0,19. Kui suur on standardhälve eurodes? Kas variatsioonikordaja muutub, kui minna kroonidelt üle eurodele? 1 EUR = 15,6466 EEK. VASTUS lk 661.

Tšebõšovi teoreem

3.9. Leida, kui suur osa tunnuse väärtustest on
a) aritmeetilisest keskmisest kaugemal kui 1,5 standardhälvet;
b) aritmeetilisele keskmisele lähemal kui 1,5 standardhälvet;
c) aritmeetilisest keskmisest kaugemal kui 5 standardhälvet;
d) aritmeetilisele keskmisele lähemal kui 5 standardhälvet.
VASTUS lk 661.

Standardiseeritud väärtus

3.10. 2014. aasta riigieksamil oli kitsa matemaatika tulemuste aritmeetiline keskmine 34,2 punkti standardhälbega 21,5 punkti. Eesti keele eksami tulemuste aritmeetiline keskmine oli 64,8 punkti standardhälbega 16,2 punkti[8]. Pille sai kitsas matemaatikas 50 punkti ja eesti keeles 80 punkti. Kummas aines oli tema tulemus silmapaistvam? VASTUS lk 661.

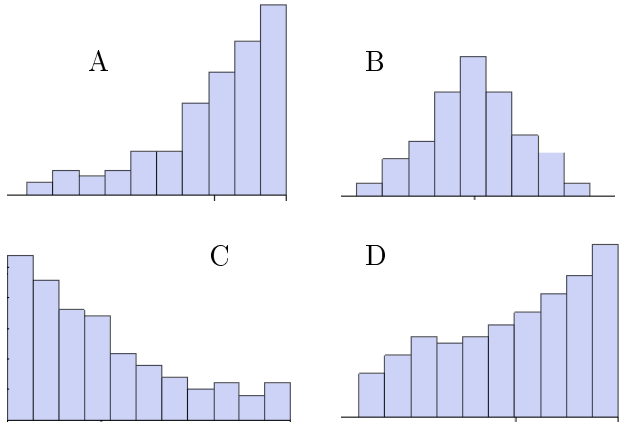
Asümmeetria ja püstakus

3.11. Joonisel 3.17 on nelja erineva jaotuse histogrammid. Määrata, milline asümmeetriakordaja ja püstakuse kordaja paar millisele histogrammile vastab.

[7] Allikas: Maa-amet, tehingute andmebaas, `http://www.maaamet.ee`
[8] Allikas: Innove, riigieksamite statistika 2014, `http://www.innove.ee`

1. $A = -0{,}02$, $E = -0{,}35$;
2. $A = -1{,}1$, $E = 0{,}61$;
3. $A = -0{,}42$, $E = -1{,}0$;
4. $A = -0{,}85$, $E = -0{,}28$.

VASTUS lk 661.



Joonis 3.17. Nelja erineva jaotuse histogrammid ülesande 3.11 juurde

Momendid

3.12. Panna kirja valemid järgmiste momentide jaoks:

a) 1. järku algmoment;
b) 3. järku algmoment;
c) 4. järku keskmoment.

VASTUS lk 661.

Kaheväärtuseline tunnus

3.13. 2013. aasta Eesti sotsiaaluuringus küsitleti 15 053 isikut. Vastajatel paluti märkida ka oma sugu (mees, naine) ning rahvus (eestlane, muu rahvus). Vastajate seas oli mehi 47% ning eestlasi 79% (*Eesti sotsiaaluuring* 2013). Leida tunnuste „sugu“ ning „rahvus“ standardhälve. VASTUS lk 661.

3.14. Tuletada valem variatsioonikordaja arvutamiseks kaheväärtuselise tunnuse korral. VASTUS lk 661.

Kvartiilhaare

3.15. Tabelis on keskmise kuupalga kvartiilid kolmel tegevusalal aastatel 2008 ja 2013[9]. Kõik suurused on tuhandetes eurodes. Leida tegevusala keskmise kuupalga kvartiilhaare aastatel 2008 ja 2013. Kas hajumine palkades on suurenenud või vähenenud? VASTUS lk 661.

[9] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel EM024: ettevõtete asendikeskmised suhtarvud (kvartiilid, mediaan).

Tegevusala	2008			2013		
	Q_1	Me	Q_3	Q_1	Me	Q_3
Töötlev tööstus	0,30	0,49	0,74	0,34	0,55	0,81
Hulgi- ja jaekaubandus	0,23	0,40	0,75	0,28	0,44	0,80
Veondus ja laondus	0,29	0,44	0,69	0,32	0,46	0,71

ÜL03Varieerumine

Järgmiste ülesannete andmed on failis ÜL03Varieerumine

Standardhälve ja variatsioonikordaja

A.3.1. USA-s on üle 7000 elektrijaama. Tabelis on 25 elektrijaama tootmiskulud, elektrienergia toodang ja töötajate keskmine tunnitasu 1955. aastal (Nerlove, 1963). Leida aritmeetiline keskmine, standardhälve ja variatsioonikordaja. Milline tunnus varieerub kõige rohkem? VASTUS lk 661.

A.3.2. Tabelis on andmed 1997. veebruaris Eesti internetikaubamajas müügil olnud kohvimasinate hindade kohta. Hinnad on teisendatud eurodesse. Analüüsida kohvimasinate hindade varieerumist erinevate tootjatel korral. Selleks leida aritmeetiline keskmine, standardhälve ja variatsioonikordaja terve kogumi ning eraldi nelja firma jaoks: AEG, Moulinex, Philips ja Severin. Millise tootja hinnad varieeruvad kõige rohkem? Millisel tootjal varieeruvad hinnad kõige vähem? VASTUS lk 661.

Standardhälve ja sagedustabel

A.3.3. Tehases, kus toodetakse kuullaagreid, võeti ühe tootmisliini toodangu hulgast välja 50 laagrit (valim 1) ning teise liini toodangu hulgast 100 laagrit (valim 2). Kummaski valimis mõõdeti ära kuullaagrite diameetrid. Mõõtmistulemused on toodud tabelites „Valim 1“ ja „Valim 2“.

1. Leida diameetrite aritmeetiline keskmine ja standardhälve kummaski valimis.
2. Võrrelda mõlema valimi variatsioonikordajaid. Mida võib järeldada?
3. Leida summaarse kogumi (150 tk) diameetrite aritmeetiline keskmine (üldkeskmine).
4. Seejärel avastati, et teise valimi mõõtmisel oli mõõteriista null paigast ära ja mõõtmistulemused olid tegelikest väärtustest 0,003 cm võrra väiksemad. Milline on teise valimi diameetrite tegelik aritmeetiline keskmine, standardhälve ja variatsioonikordaja?
5. Kas parandus mõjutas 2. osas tehtud järeldust?

Vastus lk 661.

A.3.4. Tabelis on juunikuu keskmine õhutemperatuur Tartus aastatel 2000–2014[10]. Leida, kui suurel osal kõikidest juunikuudest võib kuu keskmine õhutemperatuur olla

Tšebõšovi teoreem

a) madalam kui 11 °C;
b) kõrgem kui 19 °C.

Eeldada, et kuu keskmiste õhutemperatuuride jaotus on sümmeetriline. Vastus lk 662.

A.3.5. 33 Euroopa riigi kohta on toodud töötuse määr (%) ja elanikkonna mediaanvanus aastal 2011[11]. Leida mõlema tunnuse standardiseeritud väärtus Eesti jaoks. Kumb näitaja on Eestis Euroopa keskmisele lähemal? Vastus lk 662.

Standardiseeritud väärtus

A.3.6. Ajakirjas Chance ilmunud artiklis „Bordeaux Wine Vintage Quality and the Weather“ analüüsisid autorid, kuidas veini hind sõltub selle aasta ilmast, mil viinamarjad korjati (Ashenfelter, Ashmore ja Lalonde, 1995). Tabelis on kuue veinisordi tosina pudeli hinnad USA dollarites Londoni veinioksjonil aastatel 1961 ja 1962. Kummal aastal erineb veini Montrose hind teiste veinisortide hindadest rohkem? Vastus lk 662.

A.3.7. 50 Aasia, Ladina-Ameerika ja Vaikse ookeani riikide kohta on toodud järgmised 2009. aasta andmed[12]:

Asümmeetria

- SKP kasvumäär aastas, %;
- põllumajanduses toodetud lisandväärtuse kasvumäär aastas, %;
- tööstuses toodetud lisandväärtuse kasvumäär aastas, %;
- imikute suremus (surmade arv 1000 elussünni kohta).

Leida, milline tunnus on

a) kõige sümmeetrilisema jaotusega;
b) kõige suurema positiivse asümmeetriaga;
c) kõige suurema negatiivse asümmeetriaga.

Vastus lk 662.

A.3.8. Nord Pool Spot on üks maailma suurimaid elektribörse ja see tegutseb Põhjamaades, Saksamaal, Eestis, Leedus ja ka Suurbritannias. Tabelis on toodud elektrienergia keskmine börsihind päevas ja selle muutus börsi Nord Pool Spot neljas hinnapiirkonnas ajavahemikus aprill kuni september 2011[13]. Hinnapiirkonnad on: F1 Soome, DK1

Asümmeetria ja püstakus

[10] Allikas: Tartu Ülikooli füüsikaoskonna e-ilmajaam `http://meteo.physic.ut.ee/`

[11] Allikas: Eurostat `http://ec.europa.eu/eurostat`

[12] Allikas: World dataBank `http://databank.worldbank.org`

[13] Allikas: Nord Pool Spot `http://www.nordpoolspot.com/`

Taani 1. hinnapiirkond, DK2 Taani 2. hinnapiirkond, EE Eesti hinnapiirkond. Leida

a) hinnamuutus protsentides ajavahemikus 1.04.2011 kuni 30.09.2011 kõigis hinnapiirkondades;
b) protsentuaalse hinnamuutuse aritmeetiline keskmine, miinimum, maksimum, asümmeetriakordaja ja püstakuse kordaja. Mida võib kahe viimase näitaja põhjal järeldada?

VASTUS lk 662.

Kirjeldav statistika

A.3.9. 38 riigi kohta on toodud järgmised andmed[14]:

- keskmine eluiga;
- inimesi ühe televiisori kohta;
- inimesi ühe arsti kohta.

Leida kirjeldava statistika suurused kõigi tunnuste jaoks, kasutades Exceli programmis analüüsivahendit *Descriptive Statistics* (vt lisa C.3). Leida vastused järgmistele küsimustele, vajadusel teha lisaarvutusi:

1. Mida võib järeldada tunnuste jaotuste sümmeetria kohta?
2. Kummal tunnusel on jaotuse „saba“ lamedam ja kaugemale ulatuv, kas „inimesi ühe televiisori kohta“ või „inimesi ühe arsti kohta“?
3. Milline suurus varieerub kõige rohkem?

VASTUS lk 662.

A.3.10. Tabelis on toodud 1993. aastal väljalastud uute automudelite parameetrid (*The 1993 Cars — Annual Auto Issue* 1993).

1. Milline on kõige sagedamini esinev silindrite arv?
2. Leida hind, millest odavam oli neljandik mudelitest.
3. Kumb suurus varieerub rohkem, kas hind või võimsus?
4. Millise suuruse jaotus on kõige lamedam?
5. Kummad autod on keskmiselt võimsamad, USA-s või mujal toodetud?

VASTUS lk 662.

A.3.11. Toodud on 21 triikraua tootja, mudel, võimsus ja hind veebruaris 1999. Hinnad on teisendatud eurodesse.

1. Leida kõige sagedamini esinev triikraua võimsus.
2. Millisesse hinnavahemikku jäävad 25% kõige kallima triikraua hinnad?
3. Kui suur on hindade kvartiilhaare?

VASTUS lk 662.

Detsiilhaare

A.3.12. 2012. aasta leibkonna eelarve uuringus osalenud perede seas oli 1917 kolmeliikmelist peret (*Leibkonna eelarve uuring* 2012). Tabelis

[14] Allikas: World dataBank `http://databank.worldbank.org`

on nende perede kulutused leibkonnaliikme kohta aastas kahes hüviste grupis: toit ja mittealkohoolsed joogid ning side. Leida kummagi grupi jaoks kulutuste detsiilhaare ja suhteline detsiilhaare. Kumma grupi korral on kulutuste hajuvus suurem? VASTUS lk 662.

Erinevad variatsioonnäitarvud

A.3.13. Investeeringu objektiivne ehk statistiline risk on investeeringu tulemuse määramatus. Seda riski võib iseloomustada erinevate hajuvust kajastavate näitajatega nagu näiteks (Kaarma ja Paas, 2000):

- variatsioonamplituud;
- kvartiilhälve;
- standardhälve;
- variatsioonikordaja.

Tabelis on toodud nelja pensionifondi (LHV Pensionifond S, Nordea Pensionifond A, SEB Optimaalne Pensionifond ja Swedbank Pensionifond K2) osaku puhasväärtuse muutus protsentides perioodil 1. jaanuar 2013–23. oktoober 2015[15].

1. Leida nende pensionifondide jaoks kõik neli nimetatud variatsioonnäitarvu.
2. Iga näitarvu korral reastada pensionifondid, alustades kõige riskantsemast, ja leida fondide järjekorranumbrid ehk astakud. Veenduda, et erinevate näitarvude kasutamine riski hindamiseks annab erinevad järjestused.
3. Kokkuvõtliku tulemuse saamiseks leida iga pensionifondi jaoks nelja astaku summa.
4. Milline fond on nelja näitaja alusel kõige riskantsem ja milline kõige vähem riskantsem?

VASTUS lk 662.

Intervallimine

A.3.14. Tabelis on toodud mõningate Eesti ettevõtete müügitulu aastal 2006[16]. Andmed on teisendatud eurodesse. Koostada sagedustabel ja histogramm kahel juhul:

a) klasside arv on arvutatud Sturgesi valemi (1.1) abil;
b) klasside arv on arvutatud Freedmani ja Diaconise pakutud valemi (3.29) abil, kus kasutatakse kvartiilhaaret.

Kumb valem sobib antud juhul paremini? VASTUS lk 662.

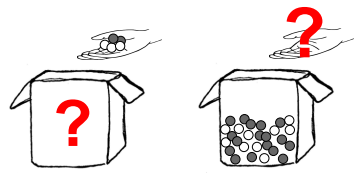
[15] Allikas: Pensionikeskus `http://www.pensionikeskus.ee/`

[16] Allikas: Eesti Ettevõtete Konkurentsivõime Edetabel `http://www.konkurents.ee`

Peatükk 4

Tõenäosusteooria elemente

Eelmises kahes alapeatükis vaatlesime vaatlusandmete kirjeldamiseks ning nendest kokkuvõtete tegemiseks kasutatavaid suurusi ja meetodeid. Kuid statistika ei piirdu ainult kokkuvõtete tegemisega. Järeldavas statistikas leitakse hinnanguid ja prognoose (veel) vaatlemata objektide ja situatsioonide jaoks. Need hinnangud ja prognoosid ei kehti kunagi sajaprotsendilise kindlusega. Üheks kesksekks mõisteks kujuneb sel juhul tõenäosus.



Statistika probleem

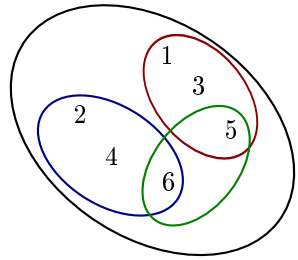
Tõenäosusteooria probleeem

Tõenäosusteooria on matemaatika haru, mis uurib juhuslike nähtuste üldisi seaduspärasusi. Statistika ja tõenäosusteooria vahekorda illustreerib kõrvalolev pilt. Järeldava statistika põhiprobleemiks on hinnata „suurest kastist“ võetud väikese valimi põhjal, mis seal „suures kastis“ on. Enne seda tuleks osata lahendada pöördülesannet ehk tõenäosusteooria ülesannet: me teame, mis on „suures kastis“ ja selle järgi oskame määrata, millise tõenäosusega millise peotäie saame.

4.1. Katse ja sündmus

Tõenäosusteoorias on **katse** protsess, mille käigus toimub teatud juhuslik valik etteantud katsetulemuste hulgast. Katse on suvaline arv kordi korratav ning katse teostamisel esineb üks ja ainult üks katsetulemus. Katse ei nõua inimese sekkumist, ka protsesside vaatlemist nimetatakse tõenäosusteoorias katseks. Näiteks ilmavaatluse tulemus võib olla pilvisus või päikesepaiste. Seega on „katse“ ja „katsetulemus“ tõenäosusteoorias laiemad mõisted kui tavaelus.

Sündmus on teatud katsetulemuste hulk. Sündmuses sisalduvad katsetulemused on vastava sündmuse jaoks soodsad. Täringu viskamisel on võimalikeks katsetulemusteks 1, 2, 3, 4, 5 ja 6 (joonis 4.1).



Joonis 4.1. Katsetulemused täringu viskamisel ja kolm sündmust: „paarisarvuline tulemus“, „paarituarvuline tulemus“, „5 või 6“

Sündmuseks võib olla paarisarvulise tulemuse saamine. Teine sündmus on paarituarvulise tulemuse saamine. Sündmuse võime defineerida ka nii: „tuleb 5 või 6“. Antud katsetulemuste hulgal võime defineerida mitmeid erinevaid sündmusi. Sündmus toimub siis, kui katse tulemuseks on mõni sündmuse jaoks soodne katsetulemus. Mõningaid näiteid katsetest ja nende katsete käigus toimuda võivatest sündmustest on toodud tabelis 4.1. Sündmusi tähistatakse tavaliselt suurte tähtedega: A, B, C, ...

Tabel 4.1. Näiteid katsete ja sündmuste kohta

Katse	Sündmus
Mündivise	A Tuleb „kull“ B Tuleb „kiri“
Kaardi võtmine pakist	A Tuleb risti B Tuleb ruutu C Tuleb punane mast D Tuleb kümme
Aktsia hinna muutus	A Aktsia hind tõuseb B Aktsia hind langeb C Aktsia hind ei muutu
Juhuslikult väljavalitud 20 kliendi küsitlus teenindamisega rahulolu kohta	A 17 klienti on rahul B 3 klienti ei ole rahul C 15 klienti on rahul D Üle 15 kliendi on rahul
Kuu keskmise käibe leidmine	A Kuu keskmine käive on vahemikus 10–15 tuh €

Kindel, võimatu ja juhuslik sündmus

Kindel sündmus on sündmus, mis antud katse korral toimub kindlasti. Kindla sündmuse hulka kuuluvad kõik antud katse võimalikud tulemused. Tähiseks on tavaliselt Ω.

Võimatu sündmus on sündmus, mis antud katse korral ei saa kunagi toimuda. Võimatu sündmuse katsetulemuste hulk on tühi hulk ja tähiseks $\emptyset$.

Juhuslik sündmus on sündmus, mis antud katse korral võib toimuda või mitte toimuda.

Juhuslikud sündmused on üksteist **välistavad**, kui need ei saa korraga toimuda. Juhuslikud sündmused on üksteist **mittevälistavad**, kui need saavad toimuda korraga. Kaardi võtmisel pakist on sündmused „tuleb risti“ ja „tuleb ruutu“ üksteist välistavad, kuid sündmused „tuleb risti“ ja „tuleb kümme“ mittevälistavad.

Sündmused moodustavad **täieliku süsteemi**, kui nad on üksteist välistavad ja katse tulemusena toimub vähemalt üks neist. Mõned näited täieliku süsteemi moodustavatest sündmustest:

- kaardi võtmisel pakist tuleb risti, ruutu, ärtu või poti;
- aktsia hinna muutus: hind tõuseb, jääb samaks või langeb;
- tänaval vastu tuleva inimese vanus on väiksem kui 30 aastat, 30–50 aastat, suurem kui 50 aastat.

Vastandsündmus

Sündmuse A **vastandsündmuseks** $\bar{A}$ nimetatakse sündmust, mis seisneb sündmuse A mittetoimumises. Vastandsündmused on üksteist välistavad ja moodustavad täieliku süsteemi.

Sündmustega võib teha tehteid, mille abil saadakse uusi sündmusi. Need tehted langevad sisuliselt kokku tehetega, mis tehakse vastavate katsetulemuste hulkadega. Seepärast kasutatakse tehtemärkidena hulgateoorias kasutatavaid märke „ühend“ $\cup$ (sündmuste summa) ja „ühisosa“ $\cap$ (sündmuste korrutis).

Tähistame ostja poolt leiva ostmist tähega L ja saia ostmist tähega S. Ostja võib osta kas ainult leiba, ainult saia või mõlemat. Need sündmused ei välista teineteist. Sündmus „ostja ostab kas leiba või saia või mõlemat“ on nende kahe sündmuse summa. Joonisel 4.2 on see kogu erinevates suundades viirutatud ala. Sündmus „ostja ostab leiba ja saia“ on aga nende sündmuste korrutis, joonisel 4.2 on see piirkond, kus kaks erineva nurga all viirutatud hulka kattuvad.

Sündmuste summa ja korrutis

Kahe sündmuse **summaks** nimetatakse sündmust C, milles toimub kas sündmus A või sündmus B või mõlemad koos:

$$C = A \cup B. \tag{4.1}$$

Kahe sündmuse A ja B **korrutiseks** nimetatakse sündmust C, mille korral esineb nii sündmus A kui ka sündmus B:

$$C = A \cap B. \tag{4.2}$$

Joonis 4.2. Sündmuste summa ja korrutis

Sündmuste summa ja korrutisega puutume tihti kokku päringute tegemisel andmebaasidest või otsingumootoritest. Kui andmebaasis on näiteks inimeste vanus täisaastates ja me soovime välja filtreerida isikuid, kelle vanus on 18 või 19, siis tuleb filtri defineerimisel kirjutada „vanus=18 OR vanus=19“. Inglisekeelne OR tähendab „või“ ja see tähistab sündmuste summat. Mõne tarkvara korral kasutatakse selleks püstkriipsu |. Erinevad tähistused sündmuste summa tähistamiseks on:

$$A \cup B, \quad A \text{ OR } B, \quad A \mid B.$$

Kui tahame andmebaasist välja filtreerida 18-aastased mehed, siis paneme tingimuseks „vanus=18 AND sugu=mees“. Ingliskeelne AND tähendab „ja“ ning see tähistab sündmuste korrutist. Teine võimalus selle tähistamiseks on märk &. Erinevad tähistused sündmuste korrutise tähistamiseks on:

$$A \cap B, \quad A \text{ AND } B, \quad A \,\&\, B.$$

Kui paneme andmebaasist inimeste vanuse järgi filtreerides tingimuseks „vanus=18 AND vanus=19“, siis on selge, et saame tulemuseks tühja hulga. Inimese vanus ei saa olla korraga 18 aastat ja 19 aastat, need on teineteist välistavad sündmused.

Kui sündmused A ja B on üksteist **välistavad**, siis nad ei saa korraga toimuda ja nende korrutis on võimatu sündmus:

$$A \cap B = \emptyset. \tag{4.3}$$

4.2. Tõenäosus

Ühe ja sama katse tulemuste põhjal määratletud sündmustel võib olla erinev võimalikkus. Sündmusel, mille soodsate katsetulemuste hulk on suurem, on rohkem võimalusi toimuda. Seda võimalikkust väljendab sündmuse tõenäosus. Juhuslikud sündmused on **võrdvõimalikud**, kui ühel neist pole rohkem võimalusi esiletulekuks kui teisel. Võrdvõimalikel sündmustel on võrdne arv soodsaid katsetulemusi.

Tõenäosuse arvväärtuse leidmiseks on kolm meetodit:

- teoreetiline;
- statistiline;
- subjektiivne.

Teoreetiline tõenäosus

Klassikaline ehk **teoreetiline tõenäosus**
Olgu katsel n katsetulemust. Eeldame, et

1) arvesse on võetud kõikvõimalikud katsetulemused, s.t mitte ühtegi ei ole välja jäetud;
2) kõik katsetulemused on üksteist paarikaupa välistavad;
3) kõik katsetulemused on võrdvõimalikud.

Sündmuse A tõenäosus $P(A)$ on siis sündmuse jaoks soodsate katsetulemuste arvu m ja kõigi katsetulemuste arvu n suhe

$$P(A) = \frac{m}{n}. \tag{4.4}$$

Katsetulemused on võrdvõimalikud, kui neil kõigil on ühesugune võimalus vastava katse tulemusena esineda. Võrdvõimalikkuse mõiste on tihti raskesti kontrollitav. Sageli kasutatakse selle tuvastamiseks sümmeetriat (münt, täring).

Tabelis 4.2 on toodud mõningaid näiteid klassikalise tõenäosuse leidmise kohta. Kõikidel juhtudel kehtivad mainitud kolm eeldust. Kui aga näiteks täringuga on manipuleeritud ning selle raskuskeset nihutatud, siis ei ole enam silmade 1–6 esinemine võrdvõimalik ning valemit (4.4) kasutada ei saa.

Tõenäosuse omadused

Tõenäosuse definitsioonist tulenevad järgmised tõenäosuse **omadused**.

- Kuna valemis (4.4) alati $0 \leq m \leq n$, siis juhusliku sündmuse A toimumise tõenäosus võib olla vahemikus nullist üheni: $0 \leq P(A) \leq 1$.
- Kindla sündmuse Ω korral $m_\Omega = n$ ning kindla sündmuse tõenäosus $P(\Omega) = 1$.
- Võimatu sündmuse $\emptyset$ korral $m_\emptyset = 0$ ning võimatu sündmuse tõenäosus $P(\emptyset) = 0$.

Tabel 4.2. Näiteid klassikalise tõenäosuse leidmisest

Katse	Sündmus	Soodsate tulemuste arv m	Kõigi tulemuste arv n	Tõenäosus $\frac{m}{n}$
Mündivise	Tuleb "kull"	1	2	$\frac{1}{2}$
Täringuvise	Tulemus on suurem kui 4	2	6	$\frac{2}{6}=\frac{1}{3}$
Kaardi võtmine 52 lehega pakist	Mast on ruutu	13	52	$\frac{13}{52}=\frac{1}{4}$
Laps sünnib	... nädalavahetusel	2	7	$\frac{2}{7}$

Kui sündmuse A jaoks soodsate katsetulemuste arv on m_A, siis vastandsündmuse $\bar{A}$ jaoks soodsate katsetulemuste arv

$$m_{\bar{A}} = n - m_A.$$

Jagame selle võrduse mõlemad pooled läbi katsetulemuste arvuga n:

$$\frac{m_{\bar{A}}}{n} = \frac{n}{n} - \frac{m_A}{n}.$$

Viimasest seosest saame valemi vastandsündmuse $\bar{A}$ tõenäosuse leidmiseks.

Vastandsündmuse tõenäosus

Kui sündmuse A tõenäosus on $P(A)$, siis selle vastandsündmuse $\bar{A}$ tõenäosus

$$P(\bar{A}) = 1 - P(A). \tag{4.5}$$

Kui sündmused $A_1, A_2, \ldots, A_k$ moodustavad sündmuste täieliku süsteemi, siis nende sündmuste jaoks soodsate katsetulemuste arvud peavad rahuldama seost

$$m_{A_1} + m_{A_2} + \ldots + m_{A_k} = n.$$

Jagame selle võrduse mõlemad pooled läbi katsetulemuste arvuga n:

$$\frac{m_{A_1}}{n} + \frac{m_{A_2}}{n} + \ldots + \frac{m_{A_k}}{n} = \frac{n}{n}.$$

Saime seose tõenäosuste jaoks.

Täieliku sündmuste süsteemi moodustavate sündmuste tõenäosuste summa on üks:

$$\sum_{i=1}^{k} P(A_i) = 1. \tag{4.6}$$

Täieliku sündmuste süsteemi tõenäosuste summa

Paneme tähele, et kui sündmuste täieliku süsteemi moodustavad ainult kaks sündmust, siis need on sündmus ja selle vastandsündmus. Sellisel juhul saame valemist (4.6), et

$$P(A) + P(\bar{A}) = 1,$$

mis on kooskõlas valemiga (4.5).

Kui pole võimalik leida klassikalist tõenäosust, kasutatakse statistilist (empiirilist) tõenäosust. Näiteks pole teoreetiliselt võimalik leida nisuterade idanemise tõenäosust. Selle määramiseks pannakse saagist võetud terade hulgast idanema 1000 tera. Kui neist idaneb 940, loetakse põllusaagi idanevuseks

$$P(\text{„idaneb“}) = \frac{940}{1000} = 0{,}94 = 94\%.$$

Kui me soovime leida tõenäosust, et juhuslikult väljavalitud inimene on töötu, siis me võime küsitleda 1000 inimest. Kui nende tuhande isiku hulgas on 200 töötut, siis tõenäosus

$$P(\text{„töötu“}) = \frac{200}{1000} = 0{,}2 = 20\%.$$

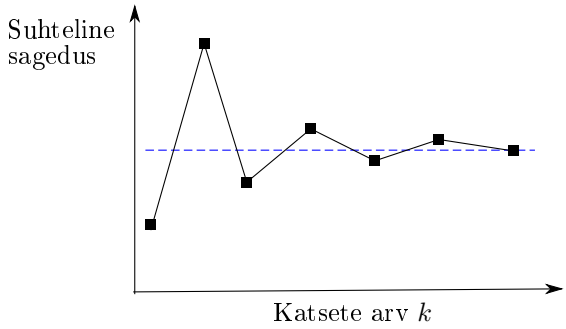
Seda nimetatakse ka töötuse määraks.

Kui k on katsete arv ja f sündmuse A esinemissagedus nendes katsetes, siis sündmuse A toimumise suhteline sagedus ehk **statistiline tõenäosus**

$$P^*(A) = \frac{f}{k}. \tag{4.7}$$

Statistiline tõenäosus

Kui katsete arv k ei ole eriti suur, on sündmuse toimumise sagedus (ja ka suhteline sagedus ehk statistiline tõenäosus) juhuslik suurus. Katsete arvu suurenedes aga statistiline tõenäosus enamasti koondub, s.t läheneb mingile konstantsele väärtusele (joonis 4.3).



Joonis 4.3. Katsete arvu suurenemisel suhteline sagedus koondub

Suurte arvude seadus

Suurte arvude seadus
Katsete arvu k suurenedes läheneb sündmuse esinemise statistiline tõenäosus sündmuse esinemise tõenäosusele üksikul katsel:

$$P^*(A) \xrightarrow{k\to\infty} P(A).$$

Kui lisaks statistilisele tõenäosusele on võimalik leida ka teoreetilist tõenäosust (näiteks täringuvise, kaardi tõmbamine pakist jpm), siis statistiline tõenäosus koondub teoreetiliseks tõenäosuseks.

Näide 4.1. Täringu viskamine

Täringu viskamisel on silmade arvu „6“ saamise teoreetiline tõenäosus $\frac{1}{6} \approx 0{,}1667$. Kui teostada erineva pikkusega katseseeriad, siis „6“ esinemise suhteline sagedus läheneb sellele arvule. Tabelis on toodud silmade arvu „6“ esinemise suhtelised sagedused erineva pikkusega katseseeriate korral, kusjuures iga k korral korrati viskeseeriaid kolm korda. Näeme, et visete arvu suurenedes läheneb suhteline sagedus teoreetilisele tõenäosusele.

Visete arv k	Seeria 1	Seeria 2	Seeria 3
50	0,120	0,180	0,160
100	0,180	0,170	0,200
1000	0,169	0,170	0,164
5000	0,165	0,165	0,160

Tõenäosuse **subjektiivset** hindamist kasutatakse, kui pole võimalik leida teoreetilist tõenäosust ega saa läbi viia ka katseid statistilise tõenäosuse leidmiseks. Subjektiivsel hindamisel kasutatakse olemasolevat informatsiooni ja intuitsiooni. Erinevad inimesed hindavad ühe ja

sama sündmuse tõenäosust erinevalt. Näiteks kui suure tõenäosusega võidab järgmise mängu korvpallimeeskond „Kärmed“? Objektiivselt pole seda tõenäosust võimalik leida ei teoreetiliselt ega ka katseliselt, igaüks võib anda aga oma subjektiivse hinnangu. Sageli kasutatakse **võrdsete võimaluste printsiipi**: kui võimalike alternatiivide tõenäosused pole teada, pole põhjust eeldada, et nad on erinevad. Kui on kaks võimalikku katsetulemust, võetakse nende mõlema tõenäosuseks 0,5. Kolme võimaliku tulemuse korral võetakse iga tulemuse tõenäosuseks 1/3.

4.3. Tehted tõenäosustega

Kui me teeme sündmustega mingi tehte, siis teades lähtesündmuste tõenäosusi, võime leida ka tehte tulemusena saadud sündmuse tõenäosuse.

Näide 4.2. Arvestuse saamise tõenäosus

200 loengukursust kuulanud üliõpilase seast sooritas esimese kontrolltöö positiivselt 160 üliõpilast ja teise kontrolltöö 140 üliõpilast. 124 üliõpilast sooritas mõlemad kontrolltööd positiivselt. Õppejõud otsustas, et arvestuse saavad need üliõpilased, kes sooritasid positiivselt vähemalt ühe kontrolltöö. Kui suur on tõenäosus, et selle kursuse kuulaja saab arvestuse?
Olgu sündmus A esimese kontrolltöö positiivne sooritamine ja sündmus B teise kontrolltöö positiivne sooritamine. Vastavad tõenäosused on siis:

$$P(A) = \frac{160}{200} = 0{,}8;$$
$$P(B) = \frac{140}{200} = 0{,}7.$$

Tõenäosus, et mõlemad kontrolltööd sooritatakse positiivselt, on

$$P(A \cap B) = \frac{124}{200} = 0{,}62.$$

Arvestuse saavad need, kes sooritasid positiivselt ainult esimese kontrolltöö, need, kes sooritasid positiivselt ainult teise kontrolltöö, pluss need, kes sooritasid positiivselt mõlemad kontrolltööd. Esimese kontrolltöö positiivselt sooritanud 160 üliõpilase seas on ka need, kes tegid ära mõlemad kontrolltööd. Ka 140 üliõpilase seas on mõlema kontrolltöö positiivselt sooritanud. Et me mõlema kontrolltöö sooritanuid kahekordselt ei arvestaks, tuleb arvestuse saajate arv leida järgmiselt: $160+140-124$. Arvestuse

saamise tõenäosus on siis:

$$\frac{160+140-124}{200} = \frac{160}{200} + \frac{140}{200} - \frac{124}{200} = 0{,}8 + 0{,}7 - 0{,}62 = 0{,}88.$$

See tähendab, et tõenäosuste summast tuleb lahutada korrutise tõenäosus.
Vastus: tõenäosus, et selle kursuse läbimisel õnnestub saada arvestus, on 0,88.

Sündmuste summa tõenäosus

Kahe sündmuse **summa tõenäosus** võrdub nende sündmuste tõenäosuste summaga, millest on lahutatud nende sündmuste korrutise tõenäosus:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{4.8}$$

Kui sündmused on teineteist välistavad, siis nende sündmuste korrutise tõenäosus on null (valem (4.3) ja valemis (4.8) jääb järele vaid tõenäosuste $P(A)$ ja $P(B)$ summa.

Kahe teineteist **välistava** sündmuse summa tõenäosus võrdub nende sündmuste tõenäosuste summaga:

$$P(A \cup B) = P(A) + P(B). \tag{4.9}$$

Näide 4.3. Ostjate vanusegrupid

Kui kuni 17-aastaseid ostjaid on 20% ostjate üldarvust (tõenäosus $P(A) = 0{,}20$) ja 18–25-aastaseid 15% (tõenäosus $P(B) = 0{,}15$), siis tõenäosus, et ostja on kuni 25-aastane, on

$$P(A \cup B) = P(A) + P(B) = 0{,}20 + 0{,}15 = 0{,}35.$$

Siin ei ole vaja korrutise tõenäosust maha lahutada, sest ostja ei saa olla korraga kuni 17-aastane ja 18–25-aastane, need on teineteist välistavad sündmused.

Tihti on sündmuse A tõenäosuse leidmisel teada lisainformatsiooni mingi teise sündmuse B toimumise kohta: toimus või ei toimunud. Näiteks tõenäosus, et ostja ostab poest beebitoitu, ei ole eriti suur. Kui aga

ostja ostab pabermähkmeid, siis tõenäosus, et ta ostab ka beebitoitu, on oluliselt suurem. Sellisel juhul on tegemist **tingliku tõenäosusega** (*conditional probability*). Seda kasutatakse näiteks sihtturunduse juures. Tõenäosus, et mingi toode rahuldab kõiki tarbijaid võrdväärselt, on üpris väike. Aga tõenäosus, et see rahuldab teatud tarbijate rühma, kes on ostnud näiteks mingit muud sellega seotud toodet, on oluliselt suurem. Seepärast on mõistlik kujundada turundustegevus valitud sihtgrupi jaoks.

Sündmuse A toimumise **tinglik tõenäosus** $P(A|B)$ on sündmuse A toimumise tõenäosus, kui on toimunud sündmus B.

Tinglik tõenäosus

Näide 4.4. Meeste ja naiste edutamise tõenäosus

Ettevõttes on 1200 töötajat: 960 meest ja 240 naist. Kahe aasta jooksul on edutatud 324 töötajat, nende seas oli 288 meest ja 36 naist. Ettevõtte juhtkonda süüdistatakse naiste diskrimineerimises, kuna naisi on vähe edutatud.

	Edutati	Ei edutatud	Kokku
Mees	288	672	960
Naine	36	204	240
Kokku	324	876	1200

Vastavad tõenäosused:

	Edutati	Ei edutatud	Kokku
Mees	0,24	0,56	0,80
Naine	0,03	0,17	0,20
Kokku	0,27	0,73	1,00

Tõenäosuste tabelist on näha, et suvalisest soost töötaja edutamise tõenäosus on 0,27. Kui suur on tõenäosus, et edutatud töötaja on mees või naine? Võtame kasutusele järgmised sündmuste tähistused: M — töötaja on mees; N — töötaja on naine; E — töötajat edutati.
Meil tuleb leida tinglikud tõenäosused $P(E|M)$ ja $P(E|N)$ ning võrrelda neid.
Mehe tõenäosuse edutamiseks leiame, kui edutatud meeste arvu jagame meeste koguarvuga: $\frac{288}{960} = 0{,}3$. Kasutades aga tõenäosu-

si, siis $\frac{0{,}24}{0{,}8} = 0{,}3$ ehk

$$P(E|M) = \frac{P(M \cap E)}{P(M)}.$$

Naise edutamise tõenäosuse leiame, kui jagame edutatud naiste arvu naiste koguarvuga: $\frac{36}{240} = 0{,}15$. Kasutades aga tõenäosusi, siis $\frac{0{,}03}{0{,}20} = 0{,}15$ ehk

$$P(E|N) = \frac{P(N \cap E)}{P(N)}.$$

Kuna naistel on edutamise tõenäosus kaks korda väiksem kui meestel, võib tõesti tegemist olla soolise diskrimineerimisega.

Tingliku tõenäosuse leidmine

Sündmuse A **tinglik tõenäosus** sündmuse B suhtes on leitav järgmiselt:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{4.10}$$

Kasutades tõenäosuse leidmiseks soodsate katsetulemuste arvu m ja kõikide katsetulemuste arvu n jagatist (valem (4.4) või (4.7)), võib tinglikku tõenäosust arvutada järgmiselt: olgu kõikide katsetulemuste arv n, sündmuse B jaoks soodsate katsetulemuste arv m_B ning sündmuste A ja B korrutise $A \cap B$ (esinevad mõlemad sündmused) jaoks soodsate tulemuse arv m_{AB}. Siis

$$P(B) = \frac{m_B}{n}, \quad P(A \cap B) = \frac{m_{AB}}{n}, \quad P(A|B) = \frac{m_{AB}}{m_B}.$$

Näide 4.5. Vaesusrisk paljulapselistes peredes

Suhteliseks vaesuseks loetakse seda, kui sissetulek leibkonna liikme kohta on väiksem kui 60% mediaansissetulekust. Nimetame olukorda, kus leibkond elab suhtelises vaesuses, vaesusriskiks, ja tähistame seda olukorda tähega V.
2013. aasta Eesti sotsiaaluuringu valimis oli vastajaid 15 053. Neist elas suhtelisest vaesuspiirist allpool 2770. Neid leibkondi, kus alla 25-aastaseid lapsi oli neli või rohkem, oli 791. Tähistame sündmust „laste arv leibkonnas neli või rohkem“ tähega L. Nendest elas suhtelises vaesuses 303. (*Eesti sotsiaaluuring* 2013)

Siis üldine vaesusriski tõenäosus on

$$P(V) = \frac{2770}{15053} = 0{,}184.$$

Vaesusriski tõenäosus nende perede korral, kus alla 25-aastasi lapsi on neli või rohkem, on

$$P(V|L) = \frac{303}{791} = 0{,}383.$$

Kuna sündmuse A toimumise tõenäosust, kui on toimunud sündmus B, nimetatakse tinglikuks tõenäosuseks $P(A|B)$, siis sündmuse A toimumise tõenäosust ilma informatsioonita sündmuse B kohta $P(A)$ võib nimetada tingimusteta tõenäosuseks (*unconditional probability*). On selge, et kui tinglik ja tingimusteta tõenäosused on võrdsed: $P(A|B) = P(A)$, siis sündmuse B toimumine ei mõjuta sündmuse A toimumise tõenäosust. Sellisel juhul öeldakse, et sündmused A ja B on sõltumatud. Kui aga $P(A|B) \neq P(A)$, siis sündmuse B toimumine mõjutab sündmuse A toimumise tõenäosust (suurendab või vähendab) ning A ja B on sõltuvad sündmused.

Sõltuvad ja sõltumatud sündmused

Kui kahe sündmuse A ja B korral kehtib võrdus

$$P(A|B) = P(A), \tag{4.11}$$

siis öeldakse, et A ja B on **sõltumatud** sündmused.
Kui

$$P(A|B) \neq P(A), \tag{4.12}$$

siis sündmuse B toimumine mõjutab sündmuse A toimumise tõenäosust ning A ja B on **sõltuvad** sündmused.

Näide 4.6. Kas koolitusest on kasu?

Kokkuvõtete tegemisel müügikoolitusprogrammist selgus, et 50-st eelmisel aastal preemia saanud müügiesindajast oli 20 läbinud koolituse. Kokku on ettevõttes 200 müügiesindajat ja koolituse läbis 40% kõigist müügiesindajatest.

1. Kui suur on tõenäosus, et müügiesindaja saab preemia?
2. Kui suur on tõenäosus, et preemia saanud müügiesindaja on läbinud koolituse?

3. Kui suur on tõenäosus, et müügiesindaja on saanud nii koolituse kui ka preemia?
4. Kui suur on tõenäosus, et koolituse läbija saab preemia?
5. Kas koolitusest on kasu?

Tähistame preemia saamist tähega P ja koolitusel käimist tähega K.

1. Tõenäosus, et müügiesindaja on saanud preemia:
$$P(P) = \frac{50}{200} = 0{,}25.$$
2. Tõenäosus, et preemia saanud müügiesindaja on läbinud koolituse:
$$P(K|P) = \frac{20}{50} = 0{,}4.$$
3. Tõenäosus, et müügiesindaja on saanud nii koolitust kui ka preemiat:
$$P(K \cap P) = \frac{20}{200} = 0{,}1.$$
4. Tõenäosus, et koolituse läbija saab preemia
$$P(P|K) = ?$$
Selle leidmiseks kasutame tingliku tõenäosuse leidmise valemit (4.10) ja arvestame, et koolituse läbimise tõenäosus $P(K) = 0{,}4$:
$$P(P|K) = \frac{P(K \cap P)}{P(K)} = \frac{0{,}1}{0{,}4} = 0{,}25.$$
5. Kas koolitusest on kasu? Kuna $P(P) = P(P|K)$, on preemia saamine ja koolituse läbimine sõltumatud sündmused. Järelikult müügikoolitusest ei ole kasu.

Kui sündmuse A tinglik tõenäosus B suhtes on teada ja tuleb leida nende sündmuste korrutise tõenäosus, siis seda saab leida seosest (4.10).

Sündmuste korrutise tõenäosus

Kahe sündmuse A ja B **korrutise tõenäosus** leitakse tingliku tõenäosuse abil:
$$P(A \cap B) = P(A|B) \cdot P(B), \tag{4.13}$$
$$P(A \cap B) = P(B|A) \cdot P(A). \tag{4.14}$$

Sündmuste korrutise tõenäosust nimetatakse mõnikord ka sündmuste **ühistõenäosuseks**.

Näide 4.7. Lehe tellijad

Päevalehel Meie Päev ilmub ka laupäevane lisaleht Meie Laupäev, mida peab eraldi tellima. Toimetusele on teada, et kõikidest ajalehtede tellijatest tellib 84% just nende päevalehte. Lisaks on teada, et kõigist lehe Meie Päev tellijatest tellib 75% ka laupäevast lisalehte. Kui suur on tõenäosus, et lugeja tellib nii päevalehe Meie Päev kui ka selle laupäevase lisalehe?
Olgu päevalehe tellimine sündmus M ja lisalehe tellimine sündmus L. Siis $P(M) = 0{,}84$ ja $P(L|M) = 0{,}75$. Mõlema lehe tellimise tõenäosus on

$$P(M \cap L) = P(L|M) \cdot P(M) = 0{,}75 \cdot 0{,}84 = 0{,}63.$$

Kuna sõltumatute sündmuste korral kehtib seos $P(A|B) = P(A)$, siis korrutise tõenäosuse valemis (4.13) võime teha vastava asenduse ning korrutise tõenäosuse leidmine lihtsustub.

Kahe teineteisest **sõltumatu** sündmuse korrutise tõenäosus võrdub nende sündmuste tõenäosuste korrutisega:

$$P(A \cap B) = P(A) \cdot P(B). \tag{4.15}$$

Sõltumatute sündmuste korrutise tõenäosus

Mündi ja täringu viskamisel kehtib alati see eeldus, et järjestikuste visete tulemused on sõltumatud ning siis saame alati kasutada valemit (4.15). Samuti võib paljudel muudel juhtudel eeldada, et sündmused on üksteisest sõltumatud.

Näide 4.8. Statistika eksami päev ja ilm

Kui päikesepaistelisi ilmasid on 40% päevadest, siis tõenäosus $P(\text{„päike“}) = 0{,}4$. Eksamigraafiku tegemisel tõenäosus, et statistika eksam satub esmaspäevale, on $P(\text{„E“}) = 1/5 = 0{,}2$ (eksamid toimuvad tööpäeviti). Tõenäosus, et statistika eksam satub päikesepaistelisele esmaspäevale, on

$$P(\text{„päike“} \cap \text{„E“}) = P(\text{„päike“}) \cdot P(\text{„E“}) = 0{,}4 \cdot 0{,}2 = 0{,}08.$$

Siin kasutame valemit (4.15), sest eksamigraafiku tegemisel ei arvestata ilmaga.

Kui sõltumatuse eeldus ei kehti, peame korrutise tõenäosuse leidmiseks valemist (4.13) (või (4.14)) teadma, kuidas ühe sündmuse toimumine mõjutab teise sündmuse toimumise tõenäosust, s.t peame teadma tinglikku tõenäosust.

Näide 4.9. Sally Clarki kohtuasi

9. novembril 1999. aastal mõisteti Suurbritannias oma kahe poja mõrvas süüdi 35-aastane Sally Clark. Tema esimene poeg suri 1996. aastal 3 kuu vanuselt ja teine poeg 1998. aastal 8-nädalaselt. Mõlemal korral oli ema Sally kodus oma lapsega. Sally Clark oma süüd ei tunnistanud ning kaitse väitis, et mõlemal juhul oli tegemist imiku äkksurmaga. Kohus tegi aga süüdimõistva otsuse, kasutades statistilist tõendusmaterjali. (2003 EWCA Crim 1020)

Imiku äkksurma sündroom ehk hällisurm on ühe kuu kuni ühe aasta vanuse terve imiku äkiline ja ootamatu surm, millele ei eelnenud mingeid haigusnähte ning mille põhjus jääb seletamatuks ka lahangul. Kohtus esinenud eksperdi pediaatriaprofessor Sir Roy Meadow tunnistuse järgi oli Clarki peredele sarnastes peredes hällisurma tõenäosus 1/8543. Kahe imiku hällisurma tõenäosus on järelikult

$$\frac{1}{8543} \cdot \frac{1}{8543} \approx \frac{1}{73 \text{ mln}}.$$

Selle põhjal oli kahe imiku hällisurm Clarki peres eksperdi sõnul väga ebatõenäoline. Kohus mõistis ema vangi.

Prof Meadow viga oli see, et ta pidas kummagi imiku surma sõltumatuteks sündmusteks ja kasutas valemit (4.15). Sellele juhtis tähelepanu Inglise Kuningliku Statistika Seltsi president prof P. Green oma kirjas lordkantslerile, kes on Briti kohtusüsteemi kõrgeim ametnik (Green, 2002). Hällisurma põhjus võib olla geneetiline ning sel juhul, kui peres esines üks hällisurm, on teise imiku hällisurma tõenäosus oluliselt suurem. Seega ei tohi kasutada sõltumatute sündmuste tõenäosuste korrutamist. Prof Greeni kirjas rõhutati ka, et kui kohtus kasutatakse statistilist tõendusmaterjali, peab eksperdiks olema kindlasti kvalifitseeritud statistik.

Sally Clark mõisteti õigeks peale teistkordset edasikaebamist 2003. aasta jaanuaris ja vabastati vanglast.

Tihti on vaja leida tõenäosust, et korraga toimub rohkem kui kaks sõltumatut sündmust. Näiteks viskame münti viis korda järjest ja soo-

vime leida tõenäosust, et kõikide visete tulemuseks on „kull“. See tõenäosus on

$$0{,}5 \cdot 0{,}5 \cdot 0{,}5 \cdot 0{,}5 \cdot 0{,}5 = 0{,}5^5 = 0{,}03125.$$

Analoogselt saame leida sõltumatute sündmuste ühistõenäosuse juhul, kui üksikute sündmuste tõenäosused ei ole võrdsed.

Kui meil on n sõltumatut sündmust $\{S_1, S_2, \ldots, S_n\}$ tõenäosustega $p_1, p_2, \ldots, p_n$, siis nende sündmuste **ühistõenäosus**

$$P(S_1 \cap S_2 \cap \ldots \cap S_n) = p_1 \cdot p_2 \cdot \ldots \cdot p_n. \qquad (4.16)$$

4.4. Ülesanded

4.1. Olgu katseks juhusliku kaardi tõmbamine 52-lehelisest kaardipakist.

1. Mitu võimalikku katsetulemust on?
2. Milline meetod (teoreetiline, statistiline või subjektiivne) on antud katse juures sobiv sündmuse tõenäosuse leidmiseks?
3. Kui suure tõenäosusega tõmmatakse pakist ruutu äss?
4. Kui suure tõenäosusega tõmmatakse pakist äss?
5. Kui suure tõenäosusega tõmmatakse pakist kuningas või äss?
6. Kui suure tõenäosusega tõmmatakse pakist ruutu üksteist?

Vastus lk 664.

4.2. Telemängus kasutatakse täringut, millel on kuus tahku. Kahel tahul on silmade arv 1, kahel tahul 2 ja kahel tahul 3. Kui suur on tõenäosus, et täringu viskamisel saadakse silmade arv 3? Vastus lk 664.

4.3. Olgu katseks kahe kaardi võtmine pakist, milles on neli kaarti täisarvudega 1, 2, 3 ja 4.

1. Panna kirja kõikvõimalike katsetulemuste hulk, kui enne teise kaardi võtmist pannakse esimene kaart tagasi. Mitu katsetulemust on?
2. Panna kirja kõikvõimalike katsetulemuste hulk, kui esimest kaarti tagasi ei panda. Mitu katsetulemust on?

Vastus lk 664.

4.4. Olgu meil 52-leheline kaardipakk. Pakist võetakse juhuslikult kaks kaarti. Leida tõenäosus, et teine juhuslikult tõmmatud kaart on äss, juhul kui

a) esimene kaart pandi enne teise tõmbamist pakki tagasi;

b) esimest kaarti tagasi ei pandud ja see ei olnud äss;
c) esimest kaarti tagasi ei pandud ja see oli äss.

VASTUS lk 664.

4.5. Loteriipiletite arv on 1000. Nende hulgas on üks 100-eurone võit, viis 50-eurost võitu, kakskümmend 25-eurost võitu ja viiskümmend 10-eurost võitu. Leida tõenäosus, et ühe loteriipileti ostmisel võib
a) võita vähemalt 25 eurot;
b) saada mingi võit.

VASTUS lk 664.

4.6. Vaasis on 12 valget, 15 punast ja 9 kollast roosi. Leida tõenäosus, et pimedas ruumis juhuslikult valitud roos pole kollane. VASTUS lk 664.

4.7. 1000 loodud ettevõttest on esimese kolme aasta jooksul tegevuse lõpetanud 45 ettevõtet. Kui suur on tõenäosus, et vastloodud ettevõte ei tegutse kauem kui kolm aastat? VASTUS lk 664.

4.8. 2013. aastal jagunesid Eesti ettevõtted töötajate arvu järgi järgmiselt:[1]

Töötajate arv	Ettevõtete arv
väiksem kui 10	105 659
10–49	5793
50–249	1126
250 ja enam	182

Leida, kui suure tõenäosusega on juhuslikult valitud ettevõttes töötajate arv:
a) väiksem kui 10;
b) 250 ja enam.

VASTUS lk 664.

4.9. Juhuslikult välja valitud 100 ostjast 32 ei ostnud midagi, 45 ostjat tegid kuni 50-eurose ostu, 15 ostjat tegid 50 kuni 500-eurose ostu ja 8 ostsid kaupa rohkem kui 500 euro eest. Leida tõenäosus, et ostja
a) ei osta midagi;
b) ostab kaupa kuni 500 euro eest.

VASTUS lk 664.

4.10. Leida tõenäosus, et juhuslikult valitud palgatöötaja palk
a) on suurem kui mediaanpalk;

[1] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel ER025: statistilisse profiili kuuluvad ettevõtted töötajate arvu ja tegevusala (EMTAK2008) järgi.

b) on palkade 1. ja 3. kvartiili vahel;
c) on palkade 2. ja 3. detsiili vahel.

VASTUS lk 664.

4.11. Tõenäosus, et teatud kauba nõutav kogus nädalas ületab 500 tükki, on 0,04. Mitmel nädalal aastas võib nõutav kogus olla suurem kui 500 tk? Nädalaid on aastas 52. VASTUS lk 664.

4.12. Ettevõttes on 100 töötajat, kellest 40 on mehed. Küsitluse tulemusel selgus, et oma tööga on rahul 60 töötajat, kusjuures nendest 30 olid naised. Leida tõenäosus, et

a) mees ei ole tööga rahul;
b) naine on tööga rahul.

VASTUS lk 664.

4.13. Ettevõttes on viis osakonda. Tabelis on toodud ülevaade osakondade esitatud põhivara tellimustest.

	Osakonnad					
	Müük	Varustus	Tootmine	Raamatupidamine	Ladu	KOKKU
Mööbel	10	12	4	8	4	38
Inventar	1	3	9	1	1	15
Kontoritehnika	0	0	4	1	2	7

Avastati, et ühes tellimuses esineb viga. Leida tõenäosus, et vigane tellimus

a) oli mööbli tellimise kohta;
b) ei olnud mööbli tellimise kohta;
c) tuli laost;
d) tuli tootmisosakonnast;
e) tuli kas tootmisosakonnast või laost;
f) ei tulnud ei tootmisosakonnast ega laost;
g) oli inventari tellimise kohta, kui on teada, et see tuli tootmisosakonnast.

VASTUS lk 664.

Tehted tõenäosustega

4.14. Ülikoolis küsitleti 100 üliõpilast, kellele oli määratud stipendium. Selgus, et neist 40 töötasid, 25-l oli õppevõlgnevusi ning 15 nii töötasid kui olid ka õppevõlglased. Kui suur on tõenäosus, et stipendiumi saanud üliõpilane kas töötas või oli õppevõlglane? VASTUS lk 664.

4.15. Münti visatakse kaks korda. Kui suur on tõenäosus, et

a) mõlemal korral tuleb „kull“;
b) esimene kord tuleb „kiri“ ja teine kord „kull“;

c) ühel juhul tuleb „kiri“ ja teisel juhul „kull“?

VASTUS lk 664.

4.16. Kaks inimest viskavad mõlemad 60 korda münti. Mitmel korral peaksid nad mõlemad saama tulemuseks „kiri“? VASTUS lk 664.

4.17. Münti on rikutud, nii et „kulli“ tuleku tõenäosus on 0,2.

1. Kui suure tõenäosusega tuleb esimesel viskel „kiri“ ja teisel viskel „kull“?
2. Kui suure tõenäosusega tuleb kolmel viskel kolmest „kiri“?

VASTUS lk 664.

4.18. Testitakse uue seadme töökindlust. On kindlaks tehtud, et esimesel sisselülimisel on tõrke tekkimise tõenäosus 0,1. Kui tõrget ei tekkinud, on järgmisel katsel tõrke tõenäosus sama. Kui aga tekib tõrge, on järgmisel katsel tõrke tekkimise tõenäosus 0,9. Seega tõrke tekkimise tõenäosus sõltub sellest, milline oli eelmine sündmus. Leida järgmised tõenäosused:

a) tõrget ei teki kolmel korral järjest;

b) esimesel korral tekib tõrge, siis töötab edukalt kaks katset järjest;

c) esimesel katsel töötab, siis tõrge, siis töötab ja siis jälle tõrge.

VASTUS lk 664.

4.19. Ettevõtte klientide sooline, vanuseline ja elukohajärgne jaotus on toodud tabelis.

Kliendigrupp	Osakaal
Mehed	65%
Naised	35%
18–25-aastased	27%
26–45-aastased	49%
46–60-aastased	24%
Tallinnast ja Harjumaalt	68%
Mujalt Eestist	32%

Leida tõenäosus, et kaebuse esitanud klient oli

a) mees;

b) 26–45-aastane mees;

c) mujalt Eestist pärit 26–45-aastane mees.

VASTUS lk 664.

4.20. Turu-uuring on näidanud, et reedeõhtust meelelahutussaadet jälgib peredes regulaarselt 30% meestest ja 20% naistest. Lisaks on teada, et 12% abielupaaridest jälgivad saadet koos. Kui suur on tõenäosus, et vähemalt üks abikaasa on saate regulaarne jälgija? VASTUS lk 664.

4.21. Tõenäosus, et aktsiad A annavad tulu, on 0,7 ning tõenäosus, et aktsiad B annavad tulu, on 0,9. Kui suur on tõenäosus, et tulu

toovad mõlemad aktsiad? Eeldada, et aktsiate A ja B tulusused on sõltumatud. VASTUS lk 664.

4.22. Ettevõte saatis kiirtellimuse vajaminevate detailide kohta kahele tarnijale. Kui 4 päeva jooksul kumbagi tellimust ei täideta, tuleb tootmine peatada kuni esimese partii saabumiseni. Tõenäosus, et tarnija A täidab tellimuse 4 päeva jooksul, on 0,55. Tõenäosus, et tarnija B täidab tellimuse 4 päeva jooksul, on 0,35.

1. Kui suur on tõenäosus, et mõlemad tellimused täidetakse 4 päeva jooksul?
2. Kui suur on tõenäosus, et vähemalt üks tellimus täidetakse 4 päeva jooksul?
3. Kui suure tõenäosusega tuleb tootmine 4 päeva pärast peatada?

VASTUS lk 664.

4.23. Ettevõte taotleb kahte lepingut. Lepingu A sõlmimise tõenäosus on 0,4 ja lepingu B sõlmimise tõenäosus on 0,1, kusjuures lepingute A ja B sõlmimine ei ole teineteisega seotud. Leida järgmised tõenäosused:

a) sõlmitakse mõlemad lepingud;
b) ei sõlmita kumbagi lepingut;
c) sõlmitakse ainult üks leping.

VASTUS lk 664.

4.24. Aktsiaseltsi nõukogusse kuulub 15 liiget. Reorganiseerimiseks on vajalik vähemalt 80% nõukogu liikmete nõusolek. Tõenäosus, et iga üksik nõukogu liige hääletab poolt, on 0,7, ja see ei sõltu sellest, kuidas ülejäänud liikmed hääletavad. Kui suure tõenäosusega läheb reorganiseerimise ettepanek nõukogus läbi? VASTUS lk 664.

4.25. Veoauto teeb päevas kaks reisi. Täiskoorma saamise tõenäosus mõlemal reisil on 0,25, s.t täiskoorma saamise tõenäosus päeva teisel reisil ei sõltu esimesel reisil saadud koormast. Leida tõenäosus, et veok saab täpselt ühe täiskoorma päevas. VASTUS lk 664.

4.26. Enne valimisi uuriti toetust erinevatele parteidele. Küsitleti 1000 inimest, kellest 600 olid naised. Parteid A toetas 350 vastajat, parteid B 320 vastajat, parteid C 300 ja 30 vastajal puudus eelistus. Eeldades, et toetus ühele või teisele parteile ei sõltu vastaja soost, leida

a) mitu meest võis toetada parteid A;
b) mitu naist võis toetada parteid C;
c) mitu naist võis toetada kas parteid A või C.

VASTUS lk 664.

4.27. Sajast bakalaureusetöö kaitsnud üliõpilasest 20 kaitsesid töö hindele „5“ ja nende hulgas oli 15 sellist üliõpilast, kellel oli ka statistika hinne „5“. Üldse oli statistika hinne „5“ 25 üliõpilasel 100 hulgast.

1. Leida tõenäosus, et üliõpilane kaitseb bakalaureusetöö hindele „5“.
2. Leida tõenäosus, et üliõpilane kaitseb bakalaureusetöö hindele „5“, kui tal statistika hinne oli „5“.
3. Kas sündmused „bakalaureusetöö hinne 5“ ja „statistika hinne 5“ on sõltumatud või mitte?

VASTUS lk 664.

4.28. Kontrollimaks ettevõtetest maksukohustuslaste maksude arvestamise ja tasumise õigsust viis Maksu- ja Tolliamet läbi 2000 maksukontrolli ja avastas 130 maksupettuse juhtumit. Kontrollitavad ettevõtted valiti välja juhuslikult. Ehitusettevõtteid oli kontrollitud ettevõtete seas 250 ja maksupettuse juhtumeid ehitusettevõtetes oli 21. Leida tõenäosus, et maksupettus avastatakse

a) suvalise ettevõtte kontrollimisel;
b) ehitusettevõtte kontrollimisel.
c) Kui avastati maksupettus, siis kui suure tõenäosusega oli tegemist ehitusettevõttega?

VASTUS lk 664.

4.29. Tõenäosus, et uus teenus õnnestub edukalt viia Läti turule, on 0,7. Tõenäosus, et sama teenus õnnestub edukalt viia Leedu turule, on 0,6, ja tõenäosus, et see õnnestub edukalt mõlemas riigis, on 0,55. Kui Lätis õnnestus turuleviimine edukalt, siis kui suure tõenäosusega on see edukas ka Leedus? VASTUS lk 664.

4.30. Kui suur on tõenäosus, et kahelapselisel emal on mõlemad lapsed pojad? Arvutus teha kahel juhul.

1. Eeldada, et nii esimese kui ka teise lapse sünni korral on tütre ja poja sündimine ühesuguse tõenäosusega.
2. Eeldada, et esimese lapse korral on tütre ja poja sündimine ühesuguse tõenäosusega, kuid teise lapse sugu sõltub esimese lapse soost. Kui esimene laps oli poiss, siis teise lapse korral on suurem tõenäosus poisi sündimiseks. Sama tüdrukute korral. Vastavad tinglikud tõenäosused:

$$P(\text{„2. poeg“}|\text{„1. poeg“}) = P(\text{„2. tütar“}|\text{„1. tütar“}) = 0{,}6.$$

VASTUS lk 664.

4.31. Kui suure tõenäosusega on kahelapselisel emal üks poeg ja üks tütar, kui kehtib eelmises ülesandes tehtud 2. eeldus? VASTUS lk 664.

4.32. Tootmise käivitamisel tuleb seadistada tööpinke. Korrektse seadistamise tõenäosus on 0,90. Kui tööpink on seadistatud korrektselt, siis defektsete detailide osakaal on 5%. Kui aga seadistus pole korrektne, on defektsete detailide osakaal 75%.

1. Kui suur on tõenäosus, et ühe detaili kontrollimisel avastatakse defekt?
2. Kvaliteedikontroll valis välja ühe detaili ja avastas, et see on defektne. Kui suure tõenäosusega on tööpink valesti seadistatud?

VASTUS lk 664.

4.33. Jalgpallimeeskond mängib 55% mängudest koduväljakul ja 45% külas. Kodus mängides on võidu tõenäosus 0,80, külas mängides võidetakse aga tõenäosusega 0,65. Eelmisel laupäeval peetud mängus saavutati võit. Kui suure tõenäosusega toimus see mäng koduväljakul? VASTUS lk 664.

4.34. Pärast mõningast rämpspostifiltriga e-posti kliendiprogrammi kasutamist on kõik saabunud kirjad jaotatud kaheks: rämpsposti kausta on kasutaja suunanud 65% kõikidest saabunud kirjadest. Automaatselt suunab rämpspostifilter vastavasse kausta need kirjad, mille korral tõenäosus, et tegemist on rämpskirjaga, ületab väärtuse 0,8. Sõna „deal“ esineb 85% rämpsposti kuuluvatest kirjadest ja 10% ülejäänud kirjades. Kui saabuv kiri sisaldab sõna „deal“, kas see suunatakse filtri poolt rämpsposti kausta? VASTUS lk 664.

4.35. Pank planeerib korrastada krediitkaartide väljastamist. Eelmistel aastatel 5% kaardiomanikest ei suutnud oma krediitkaardi võlgnevusi ära maksta ja pank sai kahjumit. Panga juhtkond on analüüsi teel leidnud, et suvalise krediitkaardi omaniku maksevõimetuks jäämise tõenäosus on 0,05. Lisaks sellele on kindlaks tehtud, et nendel, kes maksevõimetuks ei muutunud, esines tõenäosusega 0,2 siiski ühe või mitme kuu kaardihooldustasu võlgu jäämist. Maksevõimetuks jäänutel oli see tõenäosus muidugi 1. Leida, kui suure tõenäosusega kaotab maksevõime see klient, kes on jäänud võlgu kaardihooldustasu. Pank otsustas, et sulgeb need kaardid, kelle omanikel on tõenäosus jääda maksevõimetuks suurem kui 0,20. Kas on põhjust sulgeda kaart, mille omanik jäi võlgu kaardihooldustasu? VASTUS lk 664.

4.36. Olgu meil kaks sõltuvat sündmust A ja B ning olgu teada tõenäosused $P(B|A)$, $P(B|\bar{A})$, $P(A)$ ning $P(\bar{A}) = 1 - P(A)$. Lähtudes ülesannete 4.33–4.35 lahenduskäikudest, panna kirja valem tingliku tõenäosuse $P(A|B)$ arvutamiseks. VASTUS lk 664.

Peatükk 5

Juhusliku suuruse jaotusseadused

Eelmises peatükis tutvusime tõenäosuse mõiste ja selle kasutamisega. Tõenäosust on lihtne leida klassikalistel juhtudel nagu täringuviskamine või kaarditõmbamine kaardipakist. Kuidas leida aga tõenäosust, et autobuss on täis, et järjekorras tuleb seista kauem kui viis minutit või et aktsia hind tõuseb rohkem kui 5%? Järgnevalt käsitleme mõningaid erinevate juhuslike suuruste käitumist modelleerivaid jaotusseadusi.

5.1. Diskreetse juhusliku suuruse jaotusfunktsioon

Juhuslik suurus on suurus, mis katse tulemusel omandab juhuslikult ühe ja ainult ühe oma võimalikest väärtustest. Juhuslik suurus võib olla:

- **diskreetne** — väärtused on isoleeritud, erinevad üksteisest mingi lõpliku arvu võrra;
- **pidev** — väärtused täidavad mingi vahemiku täielikult ära.

Juhuslikke suurusi tähistatakse tavaliselt suurte tähtedega ja nende väärtusi väikeste tähtedega. Kui juhuslik suurus X omandab väärtuse x_i, kirjutame $X = x_i$. Suurte ja väikeste tähtede eristamine on oluline, sest meil võib olla tegemist ühe juhusliku suuruse X erinevate väärtustega $x_1, x_2, \ldots$ või erinevate juhuslike suurustega $X_1, X_2, \ldots$, mille väärtushulgad on vastavalt $\{x_{11}, x_{12}, \ldots\}$, $\{x_{21}, x_{22}, \ldots\}, \ldots$.

Näide 5.1. Kaebuste arv nädalas

Teenindusettevõte registreerib kõik klientidelt tulnud kaebused. Teeninduskvaliteedi parandamiseks analüüsiti kaebuste arvu nä-

dalas 50 nädala jooksul. Sagedustabelis on toodud analüüsi tulemused.

Kaebuste arv nädalas x_i	0	1	2	3	4	KOKKU
Nädalate arv e. sagedus f_i	10	18	15	6	1	50

Esitame tabelis toodud andmed uuesti, kasutades sageduse asemel statistilist tõenäosust:

$$p_i = \frac{\text{sündmuse toimumise sagedus}}{\text{katsete arv}} = \frac{f_i}{\sum f_j}.$$

Kaebuste arv nädalas x_i	0	1	2	3	4	KOKKU
Tõenäosus p_i	0,20	0,36	0,30	0,12	0,02	1

Suurus „Kaebuste arv nädalas“ on juhuslik suurus. Viimases tabelis on toodud selle juhusliku suuruse kõigi väärtuste esinemise tõenäosused. See tabel kirjeldab seda juhuslikku suurust täielikult ja seda esitust nimetatakse vastava juhusliku suuruse jaotusseaduseks.

Jaotusseadus

Diskreetse juhusliku suuruse X **jaotusseaduseks** nimetatakse vastavust juhusliku suuruse kõikvõimalike väärtuste x_i ja nende tõenäosuste p_i vahel.

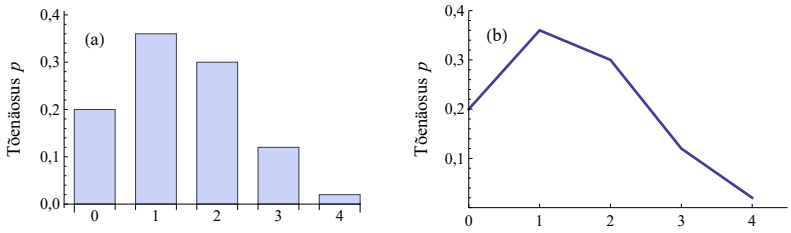
Tabelina esitatud jaotusseadus on **jaotustabel**, mis üldjuhul näeb välja nii:

Juhusliku suuruse X väärtused x_i	x_1	x_2	...	x_k
Väärtuse x_i esinemise tõenäosus p_i	p_1	p_2	...	p_k

Seda, kuidas tõenäosus jaotub erinevate väärtuste vahel, võib esitada ka diagrammil, kus horisontaalteljele kantakse juhusliku suuruse väärtused ning vertikaalteljele tõenäosused (joonis 5.1). Jaotuspolügoonil on erinevatele väärtustele vastavad tõenäosused joonega ühendatud vaid piltlikkuse huvides.

Mõnikord õnnestub jaotusseadust väljendada analüütilisel kujul, kus väärtuse x_i tõenäosus on mingi funktsioon sellest väärtusest:

$$p_i = f(x_i). \tag{5.1}$$



Joonis 5.1. Jaotusseaduse graafiline esitamine: (a) tulpdiagramm, (b) jaotuspolügoon

Näide 5.2. Silmade arvu jaotusseadus kahe täringu korral

Kahe täringu viskamisel võib summaarne silmade arv olla 2 kuni 12, kusjuures erinevate silmade arvu saamise tõenäosus on erinev. Suurus „silmade arv kokku“ on juhuslik, mille väärtused võivad olla vahemikus 2 kuni 12. Kõige väiksem tõenäosus on saada 2 või 12, kõige suurem tõenäosus on saada 7.
Erinevaid katsetulemusi on kokku $n = 6 \cdot 6 = 36$ ja need on esitatud tabeli 5.1 veerus „Erinevad võimalused kahe täringu korral“. Soodsate katsetulemuste arv m_i iga väärtuse jaoks on toodud tabeli kolmandas veerus. Viimases veerus on leitud tõenäosused $p_i = m_i/n$, tulemused on ümardatud.

Tabel 5.1. Silmade arvu jaotusseadus kahe täringu viskamisel

Silmade arv kokku x_i	Erinevad võimalused kahe täringu korral	Soodsate katsetulemuste arv m_i	Tõenäosus $p_i = m_i/n$
2	$1+1$	1	0,028
3	$1+2,\ 2+1$	2	0,056
4	$1+3,\ 2+2,\ 3+1$	3	0,083
5	$1+4,\ 2+3,\ 3+2,\ 4+1$	4	0,111
6	$1+5,\ 2+4,\ 3+3,\ 4+2,\ 5+1$	5	0,139
7	$1+6,\ 2+5,\ 3+4,\ 4+3,\ 5+2,\ 6+1$	6	0,167
8	$2+6,\ 3+5,\ 4+4,\ 5+3,\ 6+2$	5	0,139
9	$3+6,\ 4+5,\ 5+4,\ 6+3$	4	0,111
10	$4+6,\ 5+5,\ 6+4$	3	0,083
11	$5+6,\ 6+5$	2	0,056
12	$6+6$	1	0,028

Soodsate katsetulemuste arvu m_i leidmiseks võime kirja panna valemi:

$$m_i = 6 - |7 - x_i|,$$

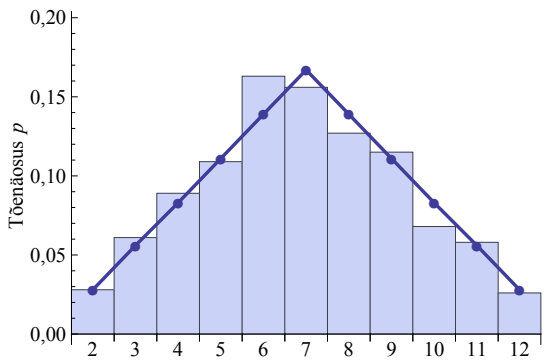
kus $|7 - x_i|$ on absoluutväärtus. Tõenäosuse jaoks saame siis valemi

$$p_i = \frac{m_i}{n} = \frac{6 - |7 - x_i|}{36}. \qquad (5.2)$$

See valem võimaldab leida juhusliku suuruse X väärtuse x_i esinemise tõenäosust ning on selle juhusliku suuruse jaotusseaduse analüütiline esitus.

Empiiriline ja teoreetiline jaotus

Tabelis 5.1 toodud tõenäosuste ja valemi (5.2) kirjapanekul ei viinud me läbi ühtegi katset. See tõenäosusjaotus on saadud teoreetilise arutluskäigu tulemusel ja on **teoreetiline jaotus**. Kui heidame kaht täringut näiteks tuhat korda, s.t viime läbi katsete seeria, siis saame **empiirilise jaotuse**. Katsete arvu suurenemisel kehtib sama suurte arvude seadus (4.2) mis statistilise ja teoreetilise tõenäosuse korral: empiiriline jaotus läheneb teoreetilisele jaotusele. Joonisel 5.2 vastab joon teoreetilisele jaotusele, mis on toodud tabelis 5.1. Tulbad vastavad tuhandel viskel saadud empiirilisele jaotusele.



Joonis 5.2. Teoreetiline (joon) ja empiiriline (tulbad) silmade arvu jaotus kahe täringu viskamisel. Teoreetiline jaotus on saadud valemist (5.2), empiirilise jaotuse saamiseks sooritati kahe täringuga tuhat viset

See, et juhuslik suurus X omandab mingi väärtuse x_i väärtuste hulgast $\{x_1, x_2, \ldots\}$, on sündmuste täielik süsteem, s.t kindlasti realiseerub üks sündmustest. Teiste sõnadega: see, et juhuslik suurus omandab mingi väärtuse, on kindel sündmus. Kindla sündmuse esinemise tõenäosus on aga 1.

Jaotusseadus peab rahuldama **normeerimistingimust**:

$$\sum_i p_i = 1, \tag{5.3}$$

kus p_i on i-nda väärtuse x_i esinemise tõenäosus ja summeerimine toimub üle kõigi juhusliku suuruse väärtuste.

Normeerimistingimus

Kui me soovime leida tõenäosust, et juhusliku suuruse väärtus on väiksem-võrdne mingist arvust a, siis liidame vastavad tõenäosused

$$P(X \leq a) = \sum_{x_i \leq a} P(X = x_i). \tag{5.4}$$

Leidmaks tõenäosust, et juhusliku suuruse väärtus on suurem mingist arvust a, kasutame normeerimistingimust. Jagame väärtuste hulga kaheks $X \leq a$ ja $X > a$. Sellisel juhul saame normeerimistingimusest, et

$$P(X \leq a) + P(X > a) = 1, \tag{5.5}$$

millest vastandsündmuse tõenäosus

$$P(X > a) = 1 - P(X \leq a). \tag{5.6}$$

Juhusliku suuruse X **jaotusfunktsioon** $F(x)$ on tõenäosus, et juhusliku suuruse X väärtus on väiksem-võrdne mingist reaalarvust x:

$$F(x) = P(X \leq x). \tag{5.7}$$

Jaotusfunktsioon

Jaotusfunktsiooni ingliskeelne termin on *cumulative distribution function* ning sellest tulenevalt kasutatakse tihti tähistust *cdf*. Mõningates õpikutes on jaotusfunktsioon defineeritud range võrratuse abil: $F(a) = P(X < a)$. Enamik jaotusfunktsiooni omadustest sellest ei sõltu ning pidevate juhuslike suuruste korral langevad need kaks definitsiooni kokku. Käesolevas õpikus oleme valinud definitsiooni (5.7) peamiselt sellepärast, et tabelarvutuses leitakse jaotusfunktsioon just niimoodi.

Diskreetse juhusliku suuruse jaotusfunktsioon on võrdne väärtuste x_i tõenäosuste summaga, kus summeeritakse kõik liidetavad, mille korral kehtib tingimus $x_i \leq a$:

$$F(a) = \sum_{x_i \leq a} P(X = x_i). \tag{5.8}$$

Erinevate probleemide lahendamisel on tihti vaja leida tõenäosust, et juhusliku suuruse X väärtus on

- **väiksem-võrdne** mingist arvust a: $P(X \leq a)$;
- **suurem** mingist arvust b: $P(X > b)$;
- mingis **vahemikus** a kuni b: $P(a < X \leq b)$.

Jaotusfunktsiooni teades on võimalik leida kõik need tõenäosused.

N05Jaotused
N5.3

Näide 5.3. Jaotusfunktsioon ja selle kasutamine

152 isiku poolt tehtud võimekuse testil saadud punktide arvu jaotust võib esitada jaotusfunktsiooni abil (vt tabel 5.2). Kumulatiivse sageduse saamiseks sagedused f_i liidetakse kuni vaadeldava väärtuse sageduseni (kaasa arvatud). Näiteks kolmandas reas oleva väärtuse 30 kumulatiivne sagedus on

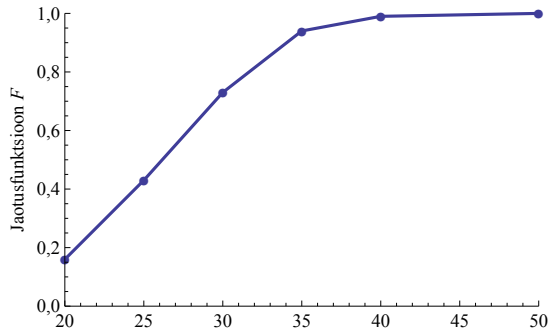
$$24 + 42 + 45 = 66 + 54 = 111.$$

Viimases veerus esitatud kumulatiivne suhteline sagedus ehk jaotusfunktsioon näitab kuni vastava väärtuseni esinevate väärtuste esinemise tõenäosust. Näiteks väärtusele 30 vastav jaotusfunktsiooni väärtus

$$F(30) = \frac{111}{152} \approx 0{,}73.$$

Tabel 5.2. Võimekuse testil saadud punktide jaotusfunktsiooni leidmine

Punktid	Sagedus f_i	Kumulatiivne sagedus	Kumulatiivne suhteline sagedus ehk jaotusfunktsioon F
20	24	24	0,16
25	42	66	0,43
30	45	111	0,73
35	32	143	0,94
40	7	150	0,99
50	2	152	1,00



Kasutades jaotusfunktsiooni väärtusi, võime leida järgmised tõenäosused.

1. Tõenäosus, et testil saadud punktide arv on kuni 25:

$$P(X \leq 25) = F(25) = 0{,}43.$$

2. Tõenäosus, et testil saadud punktide arv on suurem kui 25:

$$P(X > 25) = 1 - F(25) = 1 - 0{,}43 = 0{,}57.$$

3. Tõenäosus, et testil saadud punktide arv jääb vahemikku 31 kuni 40:

$$P(30 < X \leq 40) = F(40) - F(30) = 0{,}99 - 0{,}73 = 0{,}26.$$

Jaotusfunktsiooni definitsioonist (5.7) järeldub, et

$$P(X \leq a) = F(a). \tag{5.9}$$

Arvestades seoseid (5.6) ja (5.7), võime kirjutada

$$P(X > a) = 1 - P(X \leq a) = 1 - F(a). \tag{5.10}$$

Tõenäosus, et juhuslik suurus omandaks väärtusi poollõigus $(a, b]$, on võrdne jaotusfunktsiooni muuduga selles poollõigus:

$$P(a < X \leq b) = F(b) - F(a). \tag{5.11}$$

5.2. Keskväärtus

Kasutades tõenäosust, on võimalik üldistada kaalutud aritmeetilise keskmise (2.6) mõistet. Erinevate väärtuste kaalude asemel võetakse kasutusele nende väärtuste esinemise tõenäosused.

Näide 5.4. Nädalas esitatud kaebuste arvu keskväärtus

Näites 5.1 oli esitatud teenindusettevõttes registreeritud kaebuste arvu jaotusseadus. Leiame keskmise kaebuste arvu nädalas. Esitame uuesti näites 5.1 toodud tabeli, kus on iga väärtuse x_i esinemise sagedus f_i ehk nädalate arv. Et leida keskmine kaebuste arv nädalas (kaalutud aritmeetiline keskmine), on viimasel real leitud korrutised.

Kaebuste arv nädalas x_i	0	1	2	3	4	KOKKU
Nädalate arv e. sagedus f_i	10	18	15	6	1	50
$f_i x_i$	0	18	30	18	4	70

Keskmine kaebuste arv nädalas on leitav kaalutud aritmeetilise keskmise abil:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{70}{50} = 1{,}4.$$

Nüüd kasutame keskmise leidmiseks jaotusseadust, kus on iga väärtuse esinemise tõenäosus p_i.

Kaebuste arv nädalas x_i	0	1	2	3	4	KOKKU
Tõenäosus p_i	0,20	0,36	0,30	0,12	0,02	1
$p_i x_i$	0	0,36	0,60	0,36	0,08	**1,4**

Näeme, et summa $\sum p_i x_i = 1{,}4$ on võrdne keskmise kaebuste arvuga nädalas, mille eelnevalt leidsime kaalutud aritmeetilise keskmise abil. Aga summat $\sum p_i x_i$ ei saa me nimetada aritmeetiliseks keskmiseks, seda nimetatakse keskväärtuseks.

Olgu meil juhusliku suuruse X kõikvõimalike väärtuste hulk $\{x_1, x_2, \ldots, x_i, \ldots\}$. Suurusel X võib olla mingi kindel väärtus x_i teatud tõenäosusega $p_i = P(X = x_i)$.

Keskväärtus

Kui juhusliku suuruse X väärtuse x_i esinemise tõenäosus on p_i, siis selle juhusliku suuruse **keskväärtus** ehk **oodatav väärtus** (*expected value*)

$$\mu = E[X] = \sum p_i x_i, \tag{5.12}$$

kus summeerimine toimub üle juhusliku suuruse kõikide väärtuste.

Täringu viskamisel võib silmade arv omandada väärtusi $\{1, 2, 3, 4, 5, 6\}$ ja kõigi vääruste esinemise tõenäosus on $1/6$. Silmade arvu keskväärtus on

$$\mu = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = \\ = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3{,}5. \tag{5.13}$$

Sellele arvule läheneb täringu silmade arvu aritmeetiline keskmine, kui visete arv suureneb. Suure arvu katsete korral on keskväärtus ligikaudu võrdne katsetulemuste aritmeetilise keskmisega.

Keskväärtust tähistatakse üsna mitmel erineval moel. Levinumad tähistused on

$$\mathbf{E}X,\ E[X],\ \mu,\ \mu_X,\ <x>.$$

Viimane tähistus on levinud peamiselt füüsikas. Esimest kaht tähistust kasutatakse siis, kui on vaja esile tuua juhuslik suurus X, mille keskväärtusega on tegemist. Sümbol E tähistab keskväärtuse operaatorit ja tuleb ingliskeelsest sõnast *expectation* (ootus). Esmakordselt kasutas seda W. A. Whitworth oma raamatus „*Choice and Chance*“ 1901. aastal[1]. Käesolevas õpikus kasutame paralleelselt tähistusi μ ja $E[\ldots]$.

Keskväärtust võib leida erinevatest juhuslikest suurustest, samuti juhuslike suuruste summast, korrutisest või muudest juhuslike suurustega tehtud tehete tulemustest ning keskväärtuse operaatori kasutamine näitab korrektselt, millest keskväärtus leitakse. Näiteks juhusliku suuruse X ruudu keskväärtus

$$E[X^2] = \sum p_i x_i^2.$$

Dispersiooni definitsioonvalemi (3.3) võib keskväärtuse abil esitada kujul

$$\sigma^2 = E\left[(X - E[X])^2\right]. \tag{5.14}$$

Keskväärtuse põhiomadused

Toome ära **keskväärtuse põhiomadused**. Nende tõestused võib leida näiteks õpikutest Aksel Jõgi „Tõenäosusteooria“ või Kalev Pärna „Tõenäosusteooria algkursus“.

1. Konstandi c keskväärtus võrdub selle konstandiga:

$$E[c] = c. \tag{5.15}$$

2. Homogeensus: konstandi c võib tuua keskväärtuse operaatori ette:

$$E[cX] = cE[X]. \tag{5.16}$$

3. Aditiivsus: juhuslike suuruste summa keskväärtus on nende keskväärtuste summa:

$$E[X + Y] = E[X] + E[Y]. \tag{5.17}$$

Seda võib laiendada k juhusliku suuruse jaoks:

$$E\left[\sum_{i=1}^{k} X_i\right] = \sum_{i=1}^{k} E[X_i]. \tag{5.18}$$

[1] Kuna enne II maailmasõda olid Mandri-Euroopas peamisteks teaduskeelteks saksa ja prantsuse keel, siis mõningate allikate järgi tuleb sümbol E saksakeelsest sõnast *Erwartung* või prantsuskeelsest sõnast *espérance*, mis mõlemad tähendavad ootust.

4. Multiplikatiivsus: sõltumatute juhuslike suuruste korrutise keskväärtus on nende keskväärtuste korrutis:

$$E[XY] = E[X] \cdot E[Y]. \tag{5.19}$$

Ka multiplikatiivsust võib laiendada rohkem kui kahe sõltumatu juhusliku suuruse jaoks:

$$E\left[\prod_{i=1}^{k} X_i\right] = \prod_{i=1}^{k} E[X_i]. \tag{5.20}$$

Oodatav väärtus

Majandus- ja finantsanalüüsis kasutatakse termini „keskväärtus“ asemel sageli terminit „**oodatav väärtus**“. Oodatav väärtus on suurus, mis majandusotsuste tegemisel loob seose tõenäosuse ja mingi arvuliselt mõõdetava majandussuuruse X (nt kasum, tulu) vahel.

Näide 5.5. Oodatav kasum

Ettevõte soovib toota ja turustada kokkupandavaid elamuid. On vaja ehitada tootmisettevõte ning valida võib väikese, keskmise ja suure ettevõtte vahel. Selliste elamute nõudlus võib kasvada suureks või jääda väikeseks. Kuna suure või väikese nõudluse esinemise tõenäosus pole teada, kasutatakse võrdsete võimaluste printsiipi. Seega võetakse mõlema sündmuse tõenäosuseks 0,5. Kasumi suurust aastas (mln eurot) erinevate alternatiivide ja nõudluse korral on hinnatud järgnevalt:

Alternatiiv	Suur nõudlus, tõenäosus 0,5	Väike nõudlus, tõenäosus 0,5
Väike ettevõte	8	7
Keskmine ettevõte	14	5
Suur ettevõte	20	−9

Tabelist näeme, et kui valime väikese ettevõtte rajamise, siis tõenäosusega 0,5 võib kasum olla 8 miljonit eurot ja tõenäosusega 0,5 võib see olla 7 miljonit eurot. Keskmise ettevõtte korral võib tõenäosusega 0,5 saada kasumit 14 miljonit eurot ja tõenäosusega 0,5 võib see olla 5 miljonit eurot. Kuidas nüüd neid alternatiive omavahel võrrelda? Selleks oletame, et tabelis toodud situatsioon ei esine mitte üks kord, vaid näiteks 1000 korda järjest. Milline on sellisel juhul keskmine kasum, kui valime iga kord väikese ettevõtte rajamise? Kuna nõudluse mõlema seisundi tekkimise tõenäosus on 0,5, siis ligikaudu 500 juhul 1000 peaks olema suur nõudlus ja 500 juhul väike. Keskmine kasum väikese

ettevõtte korral

$$\frac{500 \cdot 8 + 500 \cdot 7}{1000} = 0{,}5 \cdot 8 + 0{,}5 \cdot 7 = 7{,}5.$$

Kui me aga iga kord valime keskmise ettevõtte, siis on keskmine kasum

$$\frac{500 \cdot 14 + 500 \cdot 5}{1000} = 0{,}5 \cdot 14 + 0{,}5 \cdot 5 = 9{,}5.$$

Mõistlik on valida selline alternatiiv, mis analoogse situatsiooni pideval kordamisel annab keskmiselt suurema tulemuse.
Sama tulemuseni jõuame, kui jätame ära oletuse, et sama situatsioon esineb 1000 korda järjest, ja kasutame arvutustes kohe tõenäosusi.

$$\text{Väike ettevõte: } 0{,}5 \cdot 8 + 0{,}5 \cdot 7 = 7{,}5.$$
$$\text{Keskmine ettevõte: } 0{,}5 \cdot 14 + 0{,}5 \cdot 5 = 9{,}5.$$
$$\text{Suur ettevõte: } 0{,}5 \cdot 20 + 0{,}5 \cdot (-9) = 5{,}5.$$

Valida tuleks keskmise ettevõtte rajamine, sest selle korral on oodatav kasum kõige suurem.

Oodatavat väärtust kasutatakse otsustamise teoorias kriteeriumina valiku tegemisel erinevate käitumisvariantide vahel. Näiteks investori käitumisteoorias on tüüpiliseks olukorraks püüd kindlaks määrata, milline investeering annab suurema oodatava tulu, kui on hinnatud erinevate majandussituatsioonide realiseerumise tõenäosused ja neile vastavad rahavood.

Näide 5.6. Oodatav tulumäär

Tabelis on toodud ootused kahe ettevõtte aktsiate käitumise kohta erinevates majandusolukordades.

Üldine majandusolukord	Tõenäosus	Ettevõtte A aktsiate tulumäär r_A	Ettevõtte B aktsiate tulumäär r_B
Tõus	0,2	45%	25%
Stabiilne	0,5	20%	0%
Langus	0,3	5%	−10%

Leiame oodatavad tulumäärad mõlema ettevõtte aktsiate jaoks:

$$E[r_A] = 0{,}2 \cdot 45\% + 0{,}5 \cdot 20\% + 0{,}3 \cdot 5\% = 20{,}5\%,$$
$$E[r_B] = 0{,}2 \cdot 25\% + 0{,}5 \cdot 0\% + 0{,}3 \cdot (-10\%) = 2\%.$$

Näeme, et kui otsustame osta ainult ühe ettevõtte aktsiaid, siis tuleb otsustada ettevõtte A kasuks, sest nende aktsiate oodatav tulumäär on suurem.
Kui me aga moodustame aktsiaportfelli, kus 75% aktsiaid on ettevõtte A aktsiad ja 25% ettevõtte B aktsiad, siis saame leida oodatava portfellitulu:

$$E[r_p] = 0{,}75 \cdot 20{,}5\% + 0{,}25 \cdot 2\% = 15{,}875\% \approx 15{,}9\%.$$

Kui moodustame väärtpaberiportfelli, siis oodatava portfellitulu võime leida seosest

$$E[r_p] = \sum w_i E[r_i], \tag{5.21}$$

kus w_i on teatud aktsiate osakaal portfellis ja $E[r_i]$ vastavate aktsiate oodatav tulumäär.

Diskreetse juhusliku suuruse dispersiooni leidmisel lähtutakse kaalutud dispersiooni valemist (3.4), kus viiakse läbi järgmised teisendused:

- aritmeetilise keskmise $\bar{x}$ asemel võetakse kasutusele keskväärtus μ;
- kaalude f_i asemel võetakse kasutusele tõenäosused $p_i = f_i/n$, kus $n = \sum f_j$.

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_j} = \frac{\sum (x_i - \mu)^2 f_i}{n} = \sum (x_i - \mu)^2 \frac{f_i}{n} = \sum (x_i - \mu)^2 p_i.$$

Dispersioon

Diskreetse juhusliku suuruse X **dispersioon**

$$\sigma^2 = \sum (x_i - \mu)^2 p_i, \tag{5.22}$$

kus p_i on väärtuse x_i esinemise tõenäosus ning summeerimine toimub üle kõigi juhusliku suuruse väärtuste x_i.

Paneme tähele, et valemis (5.22) on tegemist avaldise $(x_i - \mu)^2$ keskväärtusega. Seepärast esitatakse dispersiooni valem tihti keskväärtuse kaudu:

$$\sigma^2 = E\left[(X - E[X])^2\right]. \tag{5.23}$$

Olgu X täringu viskamisel saadud silmade arv. Leiame selle dispersiooni ja standardhälbe. Iga väärtuse tõenäosus $p_i = 1/6$ ning valemist (5.13) on silmade arvu keskväärtus $\mu = 3{,}5$. Dispersioon valemist

(5.22):

$$\sigma^2 = (1-3{,}5)^2 \cdot \frac{1}{6} + (2-3{,}5)^2 \cdot \frac{1}{6} + (3-3{,}5)^2 \cdot \frac{1}{6} + (4-3{,}5)^2 \cdot \frac{1}{6} + \\ + (5-3{,}5)^2 \cdot \frac{1}{6} + (6-3{,}5)^2 \cdot \frac{1}{6} = \frac{1}{6}(6{,}25 + 2{,}25 + 0{,}25 + 0{,}25 + \\ + 2{,}25 + 6{,}25) = \frac{17{,}5}{6} \approx 2{,}92.$$

Täringu silmade arvu standardhälve

$$\sigma = \sqrt{2{,}92} \approx 1{,}7.$$

Näide 5.7. Oodatava tulumäära standardhälve ja väärtpaberi riskitase

Näites 5.6 leidsime oodatava tulumäära kahe aktsia jaoks. Peale tulumäära huvitab investorit ka see, kumb aktsia on riskantsem, s.t kumma aktsia tulumäära hajuvus on suurem. Väärtpaberi riskitaset iseloomustab tulumäära standardhälve (või dispersioon). Mida suurem on standardhälve, seda suurem on tulumäära hajuvus ja järelikult seda suurem ka risk.
Kasutades näites 5.6 toodud andmeid ja leitud oodatavaid tulumäärasid, leiame tulumäära standardhälbe mõlema aktsia korral. Et ei tekiks segadust protsentide ruutuvõtmisel, kasutame arvutustes kümnendmurdusid. Aktsia A tulumäära dispersioon

$$\sigma_A^2 = 0{,}2 \cdot (0{,}45 - 0{,}205)^2 + 0{,}5 \cdot (0{,}2 - 0{,}205)^2 + \\ + 0{,}3 \cdot (0{,}05 - 0{,}205)^2 \approx 0{,}0841$$

ja standardhälve

$$\sigma_A = \sqrt{0{,}0841} = 0{,}29 = 29\%.$$

Aktsia B tulumäära dispersioon

$$\sigma_B^2 = 0{,}2 \cdot (0{,}25 - 0{,}02)^2 + 0{,}5 \cdot (0 - 0{,}02)^2 + 0{,}3 \cdot (-0{,}1 - 0{,}02)^2 \approx 0{,}0677$$

ja standardhälve

$$\sigma_B = \sqrt{0{,}0677} \approx 0{,}26 = 26\%.$$

Vastus: aktsia A tulumäära standardhälve on 29% ja aktsia B tulumäära standardhälve 26%. Aktsia A riskitase on suurem kui aktsial B.

5.3. Pidev juhuslik suurus

Pidev juhuslik suurus võib omandada mistahes reaalarvulist väärtust arvtelje mingil lõigul. Näiteks järjekorras seismise aeg, telefonikõne pikkus, pere elektrienergia tarbimine kuus ja väärtpaberi tulumäär päevas on pidevad juhuslikud suurused. Tõenäosuse käsitlemine ja sellega seotud arvkarakteristikute leidmine on pideval juhul erinev diskreetsest juhust. Selle mõistmiseks vaatame üht näidet.

Oletame, et teie kaaslane valib juhuslikult ühe arvu hulgast $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. See hulk on siis diskreetse juhusliku suuruse X väärtuste hulk. Eeldame, et teie kaaslasel ei ole ühegi arvu suhtes eelistust ning kõik valikud on võrdvõimalikud. Sellisel juhul võime kasutada klassikalise tõenäosuse valemit (4.4). Näiteks tõenäosus, et kaaslane valis arvu 2, on

$$P(X = 2) = \frac{1}{10},$$

sest soodsate võimaluste arv $m = 1$ ja kõikide võimaluste arv $n = 10$. Samasugune on kõikide ülejäänud väärtuste tõenäosus:

$$P(X = i) = \frac{1}{n} = \frac{1}{10}, \quad \text{kus} \;\; i \in \{1, 2, \ldots, 10\}.$$

Kui kaaslane peab reeglitest kinni, siis tõenäosus, et ta valis mõne muu arvu, on võrdne nulliga:

$$P(X = k) = 0, \quad \text{kus} \;\; k \notin \{1, 2, \ldots, 10\}.$$

Muudame nüüd reegleid nii, et valida tuleb üks reaalarv, mis asub vahemikus nullist üheni (kaasa arvatud). Eeldame jälle, et kõik valikud on võrdvõimalikud. Valituks võib osutuda mõni ratsionaalarv nagu $1/2$, $1/4$ või $6/7$. Aga valik võib olla ka mõni irratsionaalarv nagu näiteks $1/\sqrt{2}$ või $\pi/4$. Kui juhusliku suuruse X väärtus võib olla suvaline reaalarv selles vahemikus, $x \in [0, 1]$, siis võimalikke valikuid on lõpmatu arv, $n = \infty$. Kui nüüd soovime leida näiteks tõenäosuse, et kaaslane valis arvu 0,5, ja kasutame klassikalise tõenäosuse valemit, siis saame tõenäosuseks nulli:

$$P(X = 0{,}5) = \frac{1}{\infty} = 0.$$

Samamoodi ükskõik millise teise reaalarvu $x \in [0, 1]$ jaoks:

$$P(X = x) = 0, \quad \text{kus} \;\; x \in [0, 1].$$

Kuid ka tõenäosus, et kaaslane valis mõne reaalarvu, mis sellesse lõiku ei kuulu, on null:

$$P(X = x) = 0, \quad \text{kus} \;\; x \notin [0, 1].$$

Jõudsime kummalise tulemuseni: suvalise reaalarvu valiku tõenäosus on null! Veel kummalisem on aga see, et summa üle kõikide lõiku $[0, 1]$ kuuluvate väärtuste tõenäosuste peab olema 1, sest mingi valiku pidi kaaslane tegema.

Sellisele kummalisele tulemusele jõudsime sellepärast, et püstitasime küsimuse valesti. Pideva juhusliku suuruse X korral ei saa leida tõenäosust, et see omandab mingi üksiku konkreetse väärtuse, sest see tõenäosus on null. Nullist erinev on tõenäosus, et X on mingi konkreetse väärtuse lähedal, teatud vahemikus.

Tõenäosus, et pideva juhusliku suuruse X väärtus langeks vahemikku $(x, x + \Delta x)$ pikkusega Δx, on määratud jaotusfunktsiooni F vastavate väärtuste vahega:

$$P(x < X < x + \Delta x) = F(x + \Delta x) - F(x) = \Delta F.$$

Kuna erineva pikkusega vahemike jaoks on vahemikku langemise tõenäosus erinev, on otstarbekas kasutada tõenäosust ühiku kohta, mida nimetatakse ka **suhteliseks sagedustiheduseks**:

$$f^* = \frac{\Delta F}{\Delta x}. \tag{5.24}$$

Kui soovime leida suhtelist sagedustihedust mingis punktis, siis vahemiku laius Δx läheneb nullile. Suhtelise sagedustiheduse valemist (5.24) tuleb sel juhul võtta piirväärtus ja saadud suurust nimetatakse pideva juhusliku suuruse jaotustiheduseks ehk tõenäosustiheduseks (*probability density function*, *pdf*).

Jaotustihedus

Pideva juhusliku suuruse **jaotustihedus** on jaotusfunktsiooni tuletis:

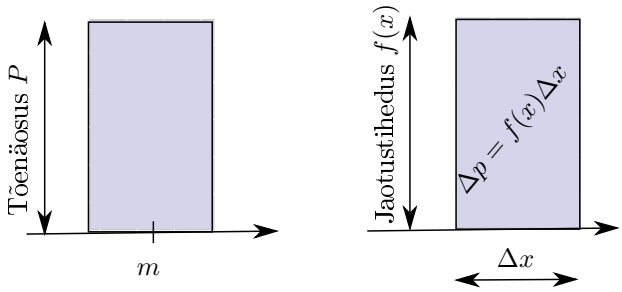
$$f(x) = \lim_{\Delta x \to 0} \frac{\Delta F}{\Delta x} = F'(x). \tag{5.25}$$

Kui diskreetse juhusliku suuruse jaotusseaduse korral esitatakse üksikutele väärtustele vastavad tõenäosused, siis pideva juhusliku suuruse korral nii ei saa, sest iga väärtuse tõenäosus on 0. Pideva juhusliku suuruse jaotusseadus esitatakse jaotustiheduse abil. Kui jaotusseaduse illustreerimiseks kasutame diagrammi, siis vertikaalteljel on nüüd tõenäosuse asemel jaotustihedus (joonis 5.3).

Praktikas huvitab meid aga tõenäosus. Olgu vahemikku Δx langemise tõenäosus $\Delta p = f(x)\Delta x$ (joonis 5.3 (b)). Kuna jaotustihedus $f(x)$ võib vahemiku Δx piires muutuda, tuleb vaadelda lõpmata väikest vahemikku ehk diferentsiaali dx. Sellele vastav tõenäosus on

$$dp = f(x)dx. \tag{5.26}$$

(a) Diskreetne suurus (b) Pidev suurus



Joonis 5.3. (a) Diskreetse suuruse korral näitab tulba kõrgus tõenäosust ja tulba laius pole oluline. (b) Pideva suuruse korral on tulba kõrguseks jaotustihedus ja tulba pindala on võrdne vahemikku Δx langemise tõenäosusega Δp

Lõplikku vahemikku (a, b) langemise tõenäosus on sel juhul integraal

$$P(a < X < b) = \int_a^b dp. \tag{5.27}$$

Tehes integraalis (5.27) asenduse (5.26), saame vahemikku langemise tõenäosuse avaldada jaotustiheduse kaudu.

Tõenäosus pideval juhul

Tõenäosus, et pideva juhusliku suuruse X väärtus jääb vahemikku (a, b), on

$$P(a < X < b) = \int_a^b f(x)dx, \tag{5.28}$$

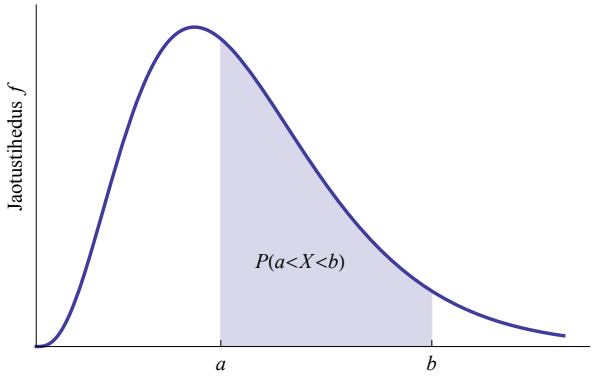
kus $f(x)$ on jaotustihedus.

Loomulikult võib tõenäosuse leidmiseks kasutatava integraali (5.28) panna kohe kirja valemi (5.3) põhjal, arvestades, et integreerimine on diferentseerimise pöördoperatsioon. Geomeetriliselt on integraal (5.28) võrdne antud vahemikus jaotustiheduse kõvera alla jääva pindalaga (joonis 5.4).

Jaotusfunktsioon on päratu integraal jaotusfunktsioonist radades miinus lõpmatusest väärtuseni a:

$$F(a) = P(X < a) = \int_{-\infty}^{a} f(x)dx. \tag{5.29}$$

Valemites (5.27), (5.28), (5.29) kasutasime rangeid võrratusi („<“ ja „>“), s.t otspunktid polnud kaasa haaratud. Kuna pideva suuruse



Joonis 5.4. Tõenäosus $P(a < X < b)$ on jaotustiheduse kõvera alla jääva varjutatud osa pindala

korral $P(X = a) = 0$, siis $P(X \leq a) = P(X < a)$ ja $P(a \leq X \leq b) = P(a < X < b)$, s.t pole vahet, kas me vahemiku otspunkti haarame kaasa või mitte.

Jaotusseaduse normeerimistingimuses (5.3) tuleb summeerimiselt üle minna integreerimisele.

Normeerimistingimus pideval juhul

Jaotustiheduse $f(x)$ **normeerimistingimus** pideva juhusliku suuruse korral, kui juhuslikul suurusel võivad olla väärtused vahemikus $(-\infty, \infty)$, on

$$\int_{-\infty}^{\infty} f(x)dx = 1. \tag{5.30}$$

Kui juhuslikul suurusel võivad olla väärtused lõigul (a, b), siis normeerimistingimus on

$$\int_{a}^{b} f(x)dx = 1. \tag{5.31}$$

Geomeetriliselt tähendab normeerimistingimus, et jaotuskõvera alla jääv

$$\text{kogupindala} = 1.$$

Pideva juhusliku suuruse mood on väärtus, mille korral jaotusfunktsioon on kõige suurem, mis on jaotuskõvera maksimumkoht (joonis 5.5(a)). Mediaan Me jaotab väärtuste vahemiku kaheks nii, et kummalegi poole jäämise tõenäosus on 0,5: $P(X < Me) = P(X > Me) = 0{,}5$. Jaotuskõvera graafikul jaotab mediaan kõveraaluse pindala kaheks võrdseks osaks, kummagi osa pindala on 0,5 (joonis 5.5(b)).

Analoogselt mediaaniga defineeritakse pideva suuruse korral suvalist järku **kvantiilid** x_p nii, et

$$P(X < x_p) = p.$$

Mood, mediaan ja kvantiilid pideval juhul

Pideva juhusliku suuruse asendikeskmised.
Mood on jaotustiheduse globaalne maksimumkoht:

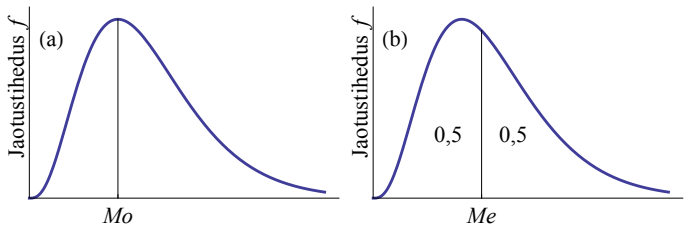
$$Mo: \quad f(Mo) = \max[f(x)]. \tag{5.32}$$

Mediaan on väärtus, millest väiksemate väärtuste esinemise tõenäosus on 0,5:

$$Me: \quad \int_{-\infty}^{Me} f(x)dx = 0{,}5. \tag{5.33}$$

p**-kvantiil** x_p on väärtus, millest väiksemate väärtuste esinemise tõenäosus on p:

$$x_p: \quad \int_{-\infty}^{x_p} f(x)dx = p. \tag{5.34}$$



Joonis 5.5. (a) Mood on jaotustiheduse globaalne maksimumkoht. (b) Mediaan jaotab jaotuskõveraaluse piirkonna kaheks võrdseks osaks, nii et kummagi osa pindala on 0,5

Täiendkvantiil

Pideva juhusliku suuruse korral on tihti kasutusel ka **täiendkvantiil** x_α, millest suuremate väärtuste esinemise tõenäosus on α: $P(X > x_\alpha) = \alpha$. Täiendkvantiil on integraali alumine raja:

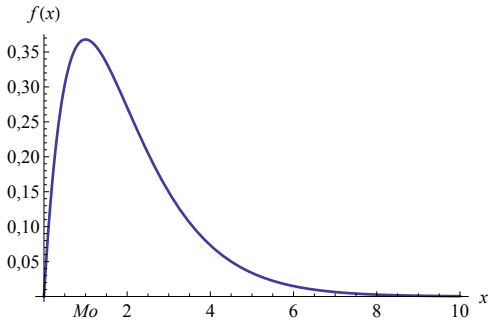
$$x_\alpha: \quad \int_{x_\alpha}^{\infty} f(x)dx = \alpha. \tag{5.35}$$

Näide 5.8. Moodi leidmine

Olgu juhusliku suuruse X jaotustihedus $f(x) = xe^{-x}$, kui $x > 0$. Leiame moodi.
Mood on jaotustiheduse maksimumkoht. Maksimumkoha leidmiseks võtame jaotustihedusest tuletise ja paneme selle võrduma nulliga. Tuletise võtmisel kasutame korrutise tuletise valemit:

$$f'(x) = \left(xe^{-x}\right)' = x'e^{-x} + x\left(e^{-x}\right)' = e^{-x} - xe^{-x} = (1-x)e^{-x}.$$

On näha, et $f'(x) = 0$, kui $x = 1$, mis on jaotusfunktsiooni maksimumkoht. Järelikult mood on 1.



Pideva juhusliku suuruse keskväärtuse leidmiseks lähtume valemist (5.36), kus i-nda väärtuse tõenäosuse p_i asemel kirjutame lõpmatult väikesesse vahemikku dx kuulumise tõenäosuse $dp = f(x)dx$ ja läheme summeerimiselt üle integreerimisele.

Keskväärtus pideval juhul

Pideva juhusliku suuruse **keskväärtus** on

$$\mu = \int_{-\infty}^{\infty} xf(x)dx, \tag{5.36}$$

kus $f(x)$ on jaotustihedus.

Kui jaotustihedus on nullist erinev vaid lõigus $[a, b]$, siis on integreerimisrajadeks lõigu otspunktid a ja b:

$$\mu = \int_{a}^{b} xf(x)dx. \tag{5.37}$$

Pideva juhusliku suuruse keskväärtuse korral kehtivad samad omadused, mis on toodud peatükis 5.2.

Näide 5.9. Normeerimistingimuse kontroll ning asendikeskmiste ja keskväärtuse leidmine

Allugu juhuslik suurus X vahemikus $[1, 10]$ jaotusele, mille jaotustihedus on

$$f(x) = \frac{1}{x \ln 10}.$$

Kontrollida, kas kehtib normeerimistingimus (5.31) ning leida mood, mediaan ja keskväärtus.

1. Normeerimistingimuse kontrollimiseks leiame integraali (5.31), kus rajadeks on vahemiku otspunktid $a = 1$ ja $b = 10$.

$$\int_1^{10} \frac{1}{x \ln 10} dx = \frac{1}{\ln 10} \int_1^{10} \frac{1}{x} dx = \frac{1}{\ln 10} [\ln x]_1^{10} =$$
$$= \frac{1}{\ln 10} (\ln 10 - \ln 1) = \frac{\ln 10}{\ln 10} = 1.$$

Vastus: normeerimistingimus kehtib.

2. Moodi leidmisel arvestame seda, et funktsioon $\frac{1}{x}$ on $x > 0$ korral monotoonselt kahanev funktsioon. Järelikult mood on vahemiku $[1, 10]$ vasakpoolne otspunkt 1. Vastus: mood on 1.

3. Mediaani leidmiseks tuleb integraalis (5.33) alumine raja võtta 1, sest kui $x < 1$, siis $f(x) = 0$. Seejärel tuleb leida integraal ning panna see võrduma 0,5-ga. Saadakse võrrand mediaani Me jaoks, mis tuleb lahendada.

$$\int_1^{Me} \frac{1}{x \ln 10} dx = \frac{1}{\ln 10} \int_1^{Me} \frac{1}{x} dx = \frac{1}{\ln 10} [\ln x]_1^{Me} = \frac{\ln Me}{\ln 10}.$$

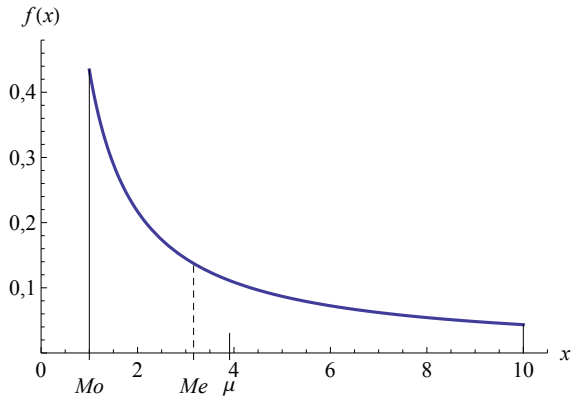
$$\begin{aligned} \frac{\ln Me}{\ln 10} &= 0{,}5 \\ \ln Me &= 0{,}5 \ln 10 \\ \ln Me &= \ln \sqrt{10} \\ Me &= \sqrt{10} \end{aligned}$$

Vastus: mediaan on $\sqrt{10} \approx 3{,}16$.

4. Keskväärtuse leidmiseks kasutame valemit (5.36), kus integreerimisrajadeks on vahemiku otspunktid:

$$\mu = \int_1^{10} x \frac{1}{x \ln 10} dx = \frac{1}{\ln 10} \int_1^{10} dx = \frac{9}{\ln 10} \approx 3{,}9.$$

Vastus: keskväärtus on ligikaudu 3,9.



Pideva juhusliku suuruse dispersiooni leidmiseks lähtume diskreetse juhusliku suuruse dispersiooni valemist (5.22). Pidevale juhule üleminekuks asendame tõenäosused p_i lõpmatult väikesesse vahemikku dx langemise tõenäosusega $dp = f(x)dx$ ja summeerimiselt läheme üle integreerimisele:

$$\sum (x_i - \mu)^2 p_i \to \int (x-\mu)^2 dp = \int (x-\mu)^2 f(x)dx.$$

Pideva juhusliku suuruse **dispersioon**

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx, \tag{5.38}$$

kus μ on keskväärtus ja $f(x)$ jaotustihedus.

Dispersioon pideval juhul

Arvestades keskväärtuse valemit (5.36) võime dispersiooni esitada keskväärtuse kaudu:

$$\sigma^2 = E\left[(X - E[X])^2\right]. \tag{5.39}$$

See langeb kokku valemiga (5.23). Järelikult valem (5.39) on üldine dispersiooni valem, mis kehtib sõltumata sellest, kas juhuslik suurus on pidev või diskreetne.

Kasutades alapeatükis 3.8 toodud statistiliste momentide formalismi, on võimalik leida valemid ka pideva jaotuse asümmeetriakordaja ning püstakuse leidmiseks (vt näiteks Ivar Tammeraidi õpikut „Tõenäosusteooria ja matemaatiline statistika“).

5.4. Teoreetilised jaotusseadused

Näites 5.2 tuletasime jaotusseaduse, millele allub kahe täringu viskamisel saadav silmade arv. Reaalses elus on palju erinevaid juhuslikke suurusi, mis käituvad erinevalt, s.t neil on erinevad jaotusseadused. Nende käitumise kirjeldamiseks on matemaatilises statistikas konstrueeritud ligikaudu 100 erinevat teoreetilist jaotusseadust. Tähtsamaid on lähemalt kirjeldatud Aksel Jõgi õpikus „Tõenäosusteooria“ ning paljudega saab tutvuda portaalis *WolframMathWorld* [2].

See, kui juhuslik suurus X allub mingile jaotusseadusele, pannakse tavaliselt kirja järgmiselt:

$$X \sim \text{jaotusseadus (jaotusseaduse parameetrid)}.$$

Teades seda, millisele jaotusseadusele juhuslik suurus allub, on võimalik prognoosida juhusliku suuruse käitumist ning leida erinevaid tõenäosusi. Mingi empiirilise juhusliku suuruse analüüsimisel:

1) valitakse arvatavalt sobiv teoreetiline jaotusseadus. Seda, kas antud juhuslik suurus allub valitud jaotusseadusele või mitte, on võimalik testida (vt näiteks χ^2-test alapeatükis 7.12);
2) vaatlusandmete põhjal leitakse valitud teoreetilise jaotusseaduse parameetrid;
3) vajalike tõenäosuste leidmiseks kasutatakse teoreetilist jaotusseadust, mille parameetrid on määratud vaatlusandmete põhjal.

Järgnevalt on loetletud mõningaid olulisemaid juhusliku suuruse jaotusseadusi (vt ka lisa A.5).

- Diskreetsed:
  - ühtlane jaotus;
  - Bernoulli jaotus;
  - binoomjaotus;
  - Poissoni jaotus.
- Pidevad:
  - pidev ühtlane ehk ristkülikjaotus;
  - eksponentjaotus;
  - normaaljaotus;
  - Studenti ehk t-jaotus;
  - χ^2 (hii-ruut) jaotus;
  - Fisheri ehk F-jaotus.

Teatud juhtudel kasutatakse diskreetse juhusliku suuruse kirjeldamiseks pideva suuruse jaotusseadust. Seda võib teha siis, kui diskreetse suuruse väärtuste arv on vaadeldavas vahemikus väga suur. Näiteks tunni aja jooksul poodi astuvate ostjate arv on diskreetne juhuslik suurus, mis võib omandada naturaalarvulisi väärtusi. Kui me analüüsime

[2] `http://mathworld.wolfram.com/topics/StatisticalDistributions.html`

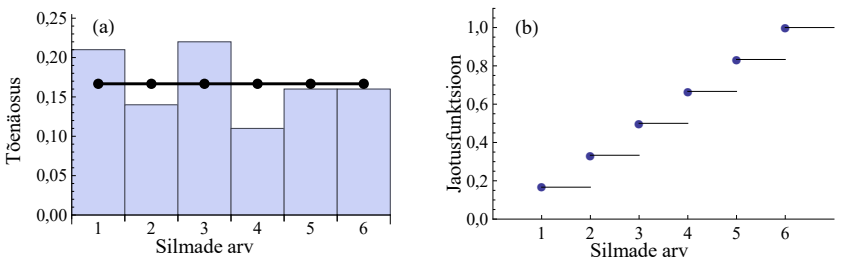
väikest maapoodi, mida tunni aja jooksul külastab ehk maksimaalselt 20 inimest, tuleb kasutada diskreetse suuruse jaotusseadust. Kui aga vaatluse all on näiteks suurlinna ostukeskus, kus külastajate arv tunnis on keskmiselt 1000, võib kasutada pideva suuruse jaotusseadust.

Järgnevalt tutvume lähemalt tähtsamate majandus- ja ärindusvaldkonnas ning statistiliste meetodite kasutamisel vajaminevate jaotustega. t-, χ^2- ja F-jaotusi käsitleme lähemalt valikvaatlustega tutvumisel ja hüpoteeside kontrollimisel (peatükid 6 ja 7).

5.5. Diskreetne ja pidev ühtlane jaotus

Kui juhuslik suurus võib omandada erinevaid diskreetseid väärusi ja kõigi väärtuste tõenäosus on ühesugune, on tegemist **diskreetse ühtlase jaotusega**. Diskreetne ühtlane jaotus väljendab võrdsete võimaluste printsiipi. Kui juhuslikul suurusel on n võimalikku väärtust, siis

$$p_1 = p_2 = \ldots = p_n = \frac{1}{n}. \tag{5.40}$$



Joonis 5.6. Näide diskreetsest ühtlasest jaotusest: täringuviskel saadud silmade arv. (a) Tõenäosusjaotus, kus joon näitab teoreetilist tõenäosust 1/6 ja tulbad kirjeldavad 100 viskel saadud tulemuste empiirilist jaotust. (b) Teoreetilise jaotuse jaotusfunktsioon

Tuntuim diskreetne ühtlane jaotus on täringu viskamisel saadav silmade arv. Erinevate silmade arvu saamise tõenäosused on ühesugused:

$$p(1) = p(2) = ... = p(6) = \frac{1}{6}.$$

Täringu viskamisel saadavate silmade arvu keskväärtus oli arvutuse (5.13) põhjal 3,5. Arv 3,5 on ka selle jaotuse mediaan. Ka juhusliku täisarvu valimine hulgast $\{1, 2, \ldots, 10\}$ (näide alapeatüki 5.3 alguses) vastab diskreetsele ühtlasele jaotusele konstantse tõenäosusega $p_i = 0{,}1$ ja keskväärtusega $\mu = 5{,}5$.

Kuna ühtlase jaotuse korral on kõikide väärtuste tõenäosus ühesugune, siis mood ühtlasel jaotusel puudub.

Analoogne ühtlane jaotus võib esineda ka pideva juhusliku suuruse korral.

Ristkülikjaotus

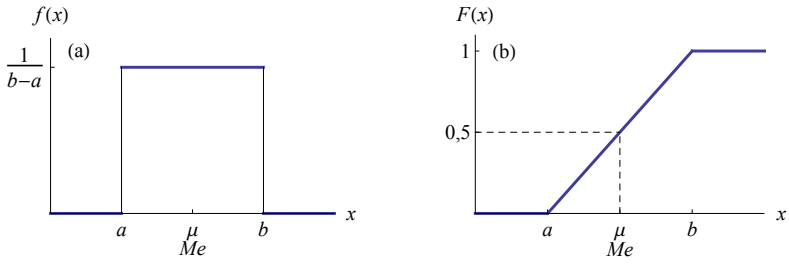
Pidevat ühtlast jaotust tuntakse ka nime all **ristkülikjaotus** lõigul $[a, b]$. Selle jaotuse jaotustihedus on lõigul $[a, b]$ nullist erinev konstant ja väljaspool seda lõiku 0:

$$f(x) = \begin{cases} 0, & \text{kui } x < a \\ \dfrac{1}{b-a}, & \text{kui } a \le x \le b \\ 0, & \text{kui } x > b \end{cases} \tag{5.41}$$

Kui X võib omada väärtusi lõigul $[0, 1]$, siis vastav jaotustihedus sellel lõigul $f(x) = 1/(1-0) = 1$. Kui aga väärtused asuvad lõigul $[100, 300]$, siis jaotustihedus sellel lõigul $f(x) = 1/(300-100) = 0{,}005$.

Jaotustiheduse selline sõltuvus lõigu otspunktidest tuleneb normeerimistingimusest (5.31):

$$\int_a^b f(x)dx = \int_a^b \frac{1}{b-a}dx = \frac{1}{b-a}\int_a^b dx = \frac{1}{b-a}(b-a) = 1.$$



Joonis 5.7. Ristkülikjaotuse jaotustihedus (a) ja jaotusfunktsioon (b). Jaotustiheduse graafikul (a) oleva ristküliku pindala rahuldab jaotustiheduse normeerimistingimust, $\dfrac{1}{b-a}(b-a) = 1$

Järgnevalt mõningaid näiteid ristkülikjaotuse kasutamise kohta.

- Ristkülikjaotus kirjeldab arvude ümardamisel tekkivat viga. Kui avaldatud SKP väärtus on näiteks 4,9 mld eurot, siis tegelik väärtus võib olla vahemikus $4{,}9 \pm 0{,}05$ mld eurot, kusjuures jaotustihedus on selles vahemikus konstantne.
- Tihti palutakse küsitlusel märkida, millisesse vanusevahemikku küsitletava vanus kuulub. Kui vastaja on valinud vastusevariandi „20–30-aastane“, siis tema tegelik vanus ankeedi analüüsija jaoks on selles vahemikus ühtlaselt jaotunud juhuslik suurus.

- Mediaani ja kvartiilide leidmisel intervallitud variatsioonridade korral eeldatakse, et vastav tunnus on intervalli piires jaotunud ühtlaselt.
- Mõnikord eeldatakse, et mingi kauba hind varieerub erinevatel müüjatel teatud vahemikus ühtlaselt (Stigler, 1961).
- Ristkülikjaotus on oluline mitmesuguste arvutisimulatsioonide puhul. Tarkvara abil genereeritakse ühtlase jaotusega juhuslikke arve lõigul $[0,1]$ ja nendest lähtutakse ülejäänud jaotuste saamisel. Simulatsioone kasutatakse näiteks riskianalüüsis.

Tabelarvutuses on vahemikus $[0,1]$ ühtlaselt jaotunud juhusliku suuruse väärtuste genereerimiseks funktsioon **RAND**. Parameetrid sellel funktsioonil puuduvad ja tabelarvutuse lahtrisse kirjutatakse lihtsalt =RAND(). Kui me soovime saada 100 arvu, siis kopeerime selle funktsiooni 100 lahtrisse. Uute väärtuste saamiseks tuleb tööleht lasta uuesti arvutada (Excelis *Calculate Sheet* või kiirklahv Shift+F9).

Kui juhuslik suurus allub ühtlasele pidevale jaotusele, siis märgitakse seda järgmiselt:

$$X \sim U(a,b),$$

kus a ja b on jaotuse parameetrid, sest need määravad üheselt ära jaotusfunktsiooni.

Ristkülikjaotuse jaotusfunktsiooni saab leida valemist (5.29):

$$F(c) = \int_a^c f(x)dx = \int_a^c \frac{1}{b-a}dx = \frac{1}{b-a}\int_a^c dx = \frac{c-a}{b-a}.$$

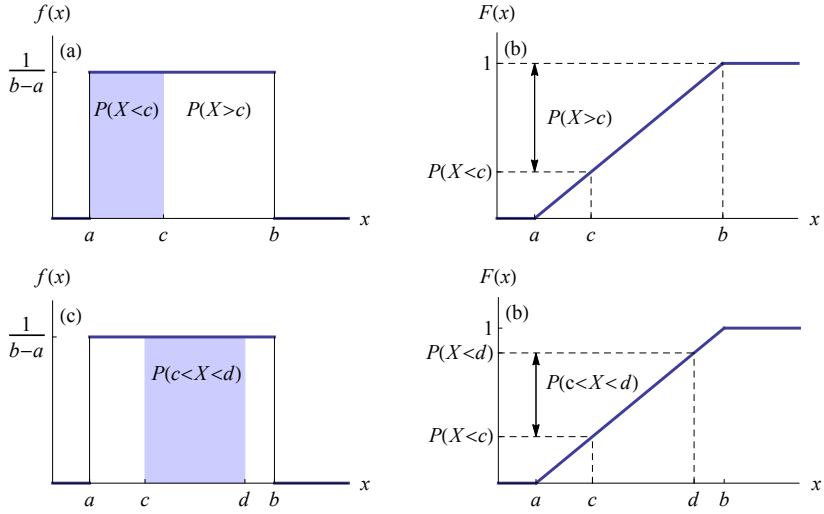
Lõigul $[a,b]$ esineva ristkülikjaotuse jaotusfunktsioon on

$$F(x) = \frac{x-a}{b-a}. \tag{5.42}$$

Geomeetriliselt on jaotusfunktsiooni väärtus kohal x võrdne sellise ristküliku pindalaga, mille laius on $x-a$ ja kõrgus $1/(b-a)$. Kui c ja d on lõigult $[a,b]$, siis tõenäosused (vt ka joonis 5.8):

$$\begin{aligned} P(X<c) &= \frac{c-a}{b-a}, \\ P(X>c) &= 1-\frac{c-a}{b-a} = \frac{b-c}{b-a}, \\ P(c<X<d) &= \frac{c-a}{b-a} - \frac{d-a}{b-a} = \frac{d-c}{b-a}. \end{aligned}$$

Viimane valem kehtib muidugi juhul, kui $c<d$.



Joonis 5.8. Jaotustiheduse graafikul (a) on tõenäosus $P(X < c)$ võrdne varjutatud ristküliku pindalaga ning tõenäosus $P(X > c)$ võrdne varjutamata ristküliku pindalaga. Jaotusfunktsiooni graafikult (b) saab tõenäosuse $F(c) = P(X < c)$ määrata jaotusfunktsiooni teljelt. Tõenäosus $P(X > c) = 1 - P(X < c)$. Tõenäosus $P(c < X < d)$ on jaotustiheduse graafikul (c) varjutatud ristküliku pindala ning jaotusfunktsiooni graafikul (d) kahe tõenäosuse vahe $P(X < d) - P(X < c)$

Pideva juhusliku suuruse keskväärtuse (5.36) ja mediaani (5.34) valemitest saab leida valemid ristkülikjaotuse keskväärtuse ja mediaani arvutamiseks, kui panna sinna jaotustihedus (5.41).

Ristkülikjaotuse keskväärtus, mediaan, dispersioon

Lõigul $[a, b]$ esineva ristkülikjaotuse keskväärtus μ ja mediaan Me langevad kokku:

$$\mu = \frac{1}{2}(a + b), \tag{5.43}$$

$$Me = \frac{1}{2}(a + b). \tag{5.44}$$

Ristkülikjaotuse dispersioon on

$$\sigma^2 = \frac{1}{12}(b - a)^2 \tag{5.45}$$

ja standardhälve

$$\sigma = \frac{1}{\sqrt{12}}(b - a). \tag{5.46}$$

Valemite (5.43) ja (5.44) tõestuse jätame lugeja hooleks (ülesanne 5.16). Dispersiooni (5.45) leidmiseks kasutatakse valemit (5.38) ja vastav tuletuskäik on toodud lisas A.2.

5.6. Bernoulli jaotus

Alapeatükis 3.9 tutvusime kaheväärtuselise tunnusega. Selleks võib olla sugu (mees/naine), tööhõive seisund (töötab/ei tööta), jaatav või eitav vastus mingile küsimusele jpt.

Kõik kaheväärtuselised tunnused saab kodeerida arvude 0 ja 1 abil. Selline juhuslik suurus X, mis võib omada vaid kahte väärtust 0 ja 1, on **Bernoulli jaotusega**[3]. Väärtuse 1 esinemise tõenäosus $p = P(X = 1)$ määrab ära ka väärtuse 0 esinemise tõenäosuse $q = P(X = 0) = 1 - p$. Järelikult on Bernoulli jaotusel üks parameeter p. Kui juhuslik suurus X on Bernoulli jaotusega, siis tähistame seda järgmiselt:

$$X \sim Be(p).$$

Bernoulli jaotusele alluva juhusliku suuruse X saab siduda mingi sündmuse A toimumisega järgmiselt:

$$X = \begin{cases} 1, & \text{kui katse tulemusena esineb sündmus } A, \\ 0, & \text{kui katse tulemusena esineb sündmus } \bar{A}, \end{cases}$$

kus $\bar{A}$ on sündmuse A vastandsündmus. Sellisel juhul on X sündmuse A indikaator. Sellist katset, millel võib olla vaid kaks tulemust, nimetatakse **Bernoulli katseks**. Näiteks on Bernoulli katseteks mündi viskamine, juhuslikult valitud inimese soo määramine, juhuslikult valitud toote nõuetele vastavuse määramine (vastab/ei vasta).

Bernoulli jaotuse keskväärtus ja dispersioon

Bernoulli jaotuse $Be(p)$ **keskväärtus**

$$\mu = p \tag{5.47}$$

ja **dispersioon**

$$\sigma^2 = p(1 - p). \tag{5.48}$$

Need valemid tuletasime me kaheväärtuselise tunnuse analüüsimise juures alapeatükis 3.9 (valemid (3.20) ja (3.25)).

[3] Jacob Bernoulli (1605–1705), Šveitsi matemaatik.

5.7. Binoomjaotus

Enamasti ei piirduta ühe Bernoulli katsega, vaid tehakse katsete seeria. Näiteks kontrollitakse mitme tooteeksemplari vastavust kvaliteedinõuetele. Kui me teeme järjest mitu Bernoulli katset, siis võime defineerida uue juhusliku suuruse, mis võrdub sündmuse A esinemise sagedusega n katse korral. Sellise juhusliku suuruse analüüsimisel on suurem praktiline tähtsus.

Näide 5.10. Edukad müügikõned

Telemarketingiga tegelevas ettevõttes töötav telefonimüüja teeb päevas 30 kõnet. Loeme edukaks telefonikõneks kõnet, mille tulemusena klient soetab pakutud toote. Kahe kuu jooksul on edukaid müügikõnesid esinenud järgmiselt:

Edukate kõnede arv päevas x_i	0	1	2	3	4	5	6
Päevade arv ehk sagedus f_i	2	6	10	10	7	4	2
Suhteline sagedus ehk statistiline tõenäosus	0,049	0,146	0,244	0,244	0,171	0,098	0,049

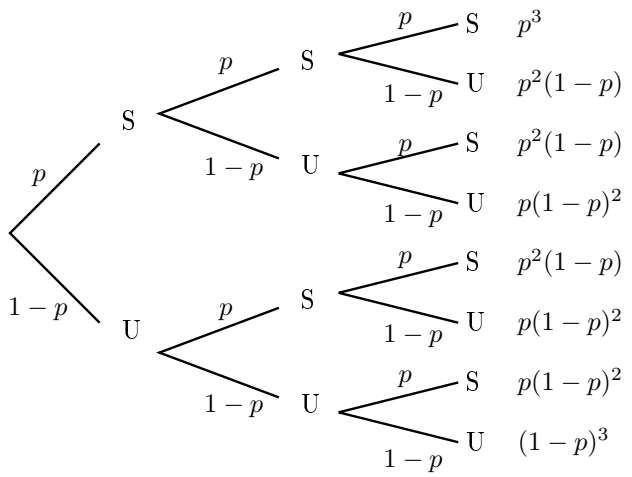
Edukate kõnede arv $\sum f_i = 41$ päeva jooksul on $\sum f_i x_i = 116$. Kokku on 41 päeva jooksul tehtud $41 \cdot 30 = 1230$ kõnet. Siit tõenäosus, et müügikõne on edukas:

$$p = \frac{116}{1230} = 0{,}094.$$

Edukate kõnede arv päevas X on juhuslik suurus, mis võib omandada väärtusi $m = \{0, 1, 2, 3, \ldots, n\}$, kus n on kõigi kõnede arv päevas. Ilmselt sõltub mingi konkreetse väärtuse esinemise tõenäosus nii kõnede arvust päevas kui ka tõenäosusest p, kas üksik kõne on edukas. Tabelis toodud statistiline tõenäosus annab meile selle juhusliku suuruse empiirilise jaotusseaduse. Kuidas leida vastav teoreetiline jaotusseadus?

Iga müügikõnet võime vaadelda kui katset, millel saab olla kaks võimalikku tulemust: „edukas“, „ei ole edukas“. Järelikult iga müügikõnet iseloomustab kaheväärtuseline tunnus. Üht kõnet võime vaadelda kui Bernoulli katset ja järjestikused müügikõned moodustavad Bernoulli katsete seeria. Tähistame edukat kõnet tähega S (*success*) ja selle tõenäosus on p. Mitteedukas kõne olgu U (*unsuccess*) ja selle tõenäosus on

$1-p$. Kolme järjestikuse kõne tulemused võime skemaatiliselt kujutada Bernoulli katsete puul, mis on toodud joonisel 5.9.



Joonis 5.9. Bernoulli katsete puu

Joonise 5.9 viimases veerus on vastava tee läbimise tõenäosus. Näiteks tee SSS tõenäosus on p^3, SSU tõenäosus $p^2(1-p)$ ja samasugune on ka teede SUS ning USS tõenäosused. Järelikult tõenäosus, et kolmest kõnest kolm on edukad, on p^3. Tõenäosus, et kolmest kõnest kaks on edukad, on $3p^2(1-p)$ ja tõenäosus, et kolmest kõnest üks on edukas, on $3p(1-p)^2$ (teed SUU, USU ja UUS).

Binoomjaotuse esinemine

Kui
- katse (*trial*) tulemus võib olla positiivne (*success*) või negatiivne;
- katset korratakse mitu korda, registreerides positiivsete tulemuste arvu;
- katsed on sõltumatud, s.t eelneva katse tulemus ei mõjuta järgnevate katsete tulemusi;
- positiivse tulemuse esinemise tõenäosus on kõikide katsete korral ühesugune;

siis positiivsete tulemuste arv on juhuslik suurus, mis allub **binoomjaotusele**.

Tähistame tõenäosust, et positiivsete tulemuste arv X omandab väärtuse m järgmiselt: $P(X = m)$. Leiame selle tõenäosuse erinevate m väärtuste jaoks. Olgu positiivse tulemuse tõenäosus üksikul katsel p, negatiivse tulemuse tõenäosus on siis $1-p$. Katsete arv on n.

Olgu $\mathbf{m = 0}$, s.t positiivseid tulemusi on 0. Kõik n katset annavad negatiivse tulemuse, mille tõenäosus on $1-p$. Tegemist on n sõltuma-

tu sündmuse korrutamisega ning järelikult tõenäosused korrutatakse (valem(4.16)):

$$P(X=0) = \underbrace{(1-p)\cdot(1-p)\cdot(1-p)\cdot\ldots\cdot(1-p)}_{n} = (1-p)^n.$$

Olgu $\mathbf{m=1}$, s.t positiivseid tulemusi on n katse hulgas 1. See võib juhtuda mitmel erineval moel. Tähistades positiivset tulemust tähega S ja negatiivset tulemust tähega U, koostame järgmise tabeli.

Tulemuste järjestus	Tõenäosus	
S U U ... U	$p\cdot\underbrace{(1-p)\cdot\ldots\cdot(1-p)}_{n-1}=$	$p\cdot(1-p)^{n-1}$
U S U U ... U	$(1-p)\cdot p\cdot(1-p)^{n-2}=$	$p\cdot(1-p)^{n-1}$
U U S U ... U	$(1-p)\cdot(1-p)\cdot p\cdot(1-p)^{n-3}=$	$p\cdot(1-p)^{n-1}$
...	...	
U U U ... S	$(1-p)^{n-1}\cdot p=$	$p\cdot(1-p)^{n-1}$

On selge, et erinevaid võimalusi on n tükki ja meid rahuldab „kas see või teine või kolmas või ...“. Tegemist on üksteist välistavate sündmuste liitmisega ja tõenäosused liidetakse. Tulemuseks saame

$$P(X=1) = np(1-p)^{n-1}.$$

Olgu $\mathbf{m=2}$, s.t positiivseid tulemusi on n katse hulgas 2. Sarnaselt eelnevale situatsioonile koostame tabeli, kus esitame erinevad võimalused.

Tulemuste järjestus	Tõenäosus	
S S U U ... U	$p\cdot p\cdot\underbrace{(1-p)\cdot\ldots\cdot(1-p)}_{n-2}=$	$p^2\cdot(1-p)^{n-2}$
S U S U U ... U	$p\cdot(1-p)\cdot p\cdot(1-p)^{n-3}=$	$p^2\cdot(1-p)^{n-2}$
U S S U U ... U	$(1-p)\cdot p\cdot p\cdot(1-p)^{n-3}=$	$p^2\cdot(1-p)^{n-2}$
...	...	
U U U ... S S	$(1-p)^{n-2}\cdot p^2=$	$p^2\cdot(1-p)^{n-2}$

Kui palju on nüüd erinevaid võimalusi? Selleks tuleb esitada küsimus: mitu võimalust on kahe elemendi valikuks n elemendi hulgast? Vastus on: kombinatsioonide arv n elemendist kahe kaupa. Tähistatakse seda tavaliselt C_n^2. Otsitav tõenäosus on järelikult

$$P(X=2) = C_n^2 p^2 (1-p)^{n-2}.$$

Suuremaid m väärtusi me ei vaata, püüame saadud tulemusi üldistada. Selleks kirjutame leitud tõenäosused uuesti välja. Arvestame seda, et $C_n^0 = 1$ (on ainult üks võimalus mitte ühegi elemendi valikuks

n hulgast) ja $C_n^1 = n$ (ühe elemendi valikuks n elemendi hulgast on n võimalust).

$$P(X=0) = (1-p)^n = C_n^0 p^0 (1-p)^n$$
$$P(X=1) = np(1-p)^{n-1} = C_n^1 p^1 (1-p)^{n-1}$$
$$P(X=2) = C_n^2 p^2 (1-p)^{n-2}.$$

Üldistades neid tulemusi suvalise arvu m jaoks, võime kirja panna

$$P(X=m) = C_n^m p^m (1-p)^{n-m}.$$

Tõenäosus binoomjaotuse korral

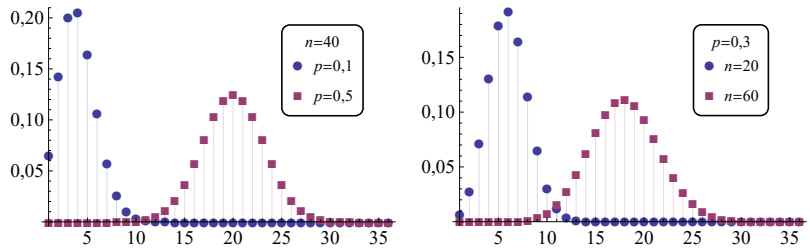
Olgu katse tulemus tõenäosusega p positiivne ja tõenäosusega $1-p$ negatiivne. Katset n korda sõltumatult korrates võib saada m positiivset ja $n-m$ negatiivset tulemust. Positiivsete tulemuste arv m on diskreetne juhuslik suurus, mille tõenäosus leitakse **Bernoulli valemist**:

$$P(X=m) = C_n^m p^m (1-p)^{n-m}, \tag{5.49}$$

kus C_n^m on kombinatsioonide arv n elemendist m kaupa:

$$C_n^m = \frac{n!}{m!(n-m)!}. \tag{5.50}$$

Tuletame meelde, et $n!$ tähistab arvu n faktoriaali: $n! = 1 \cdot 2 \cdot \ldots \cdot n$. Näiteks $4! = 1 \cdot 2 \cdot 3 \cdot 4 = 24$. Eraldi on defineeritud arvu 0 faktoriaal: $0! = 1$.



Joonis 5.10. Binoomjaotus sõltub kahest parameetrist. Vasakul on kaks jaotust, kus katsete arv on ühesugune: $n = 40$, aga positiivse tulemuse tõenäosus p on erinev. Parempoolsel graafikul on tõenäosus $p = 0{,}3$ ühesugune ja katsete arv n erinev

Binoomjaotusel on kaks parameetrit: katsete arv n ja positiivse tulemuse tõenäosus üksikul katsel p. Need määravad täielikult binoomjaotusele alluva juhusliku suuruse väärtuste tõenäosused, s.t määravad ära jaotusseaduse. Kui mingi juhuslik suurus X allub binoomjaotusele parameetritega n ja p, tähistatakse seda järgmiselt:

$$X \sim B(n,p).$$

Kui me teeme ainult ühe katse, s.t $n = 1$, siis saame Bernoulli jaotuse:

$$B(1,p) = Be(p).$$

Näide 5.11. Defektiga toodete arvu tõenäosusjaotus

Olgu defektiga toote esinemise tõenäosus 0,2. Valime suvaliselt välja 10 toodet. Leiame, kui suur on tõenäosus, et nende 10 hulgas

a) pole ühtegi defektiga toodet;
b) üks toode on defektiga;
c) kaks toodet on defektiga.

Defektiga toodete arv allub binoomjaotusele $B(10,0{,}2)$. Leiame valemi (5.49) abil tõenäosused, et n toote hulgas on defektiga tooteid m tükki, kus m võib omada väärtusi 0, 1, 2 ning $n = 10$ ja $p = 0{,}2$.

a) Kui 10 toote hulgas pole ühtegi defektiga toodet, siis $m = 0$.

$$P(X=0) = C_{10}^{0} \cdot 0{,}2^{0} \cdot (1-0{,}2)^{10-0} = \frac{10!}{0!10!} \cdot 1 \cdot 0{,}8^{10} = 0{,}107.$$

Vastus: tõenäosus, et 10 toote hulgas pole ühtegi defektiga toodet, on 0,107.

b) Kui defektseid tooteid on üks, siis $m = 1$.

$$P(X=1) = C_{10}^{1} \cdot 0{,}2^{1} \cdot (1-0{,}2)^{10-1} = \frac{10!}{1!9!} \cdot 0{,}2 \cdot 0{,}8^{9} = 0{,}268.$$

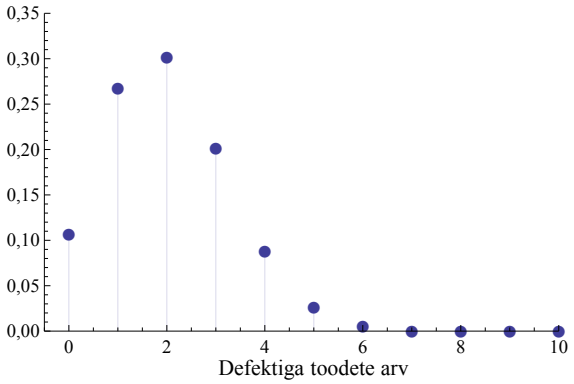
Vastus: tõenäosus, et 10 toote hulgas on üks defektiga toode, on 0,268.

c) Kui defektseid tooteid on kaks, siis $m = 2$.

$$P(X=2) = C_{10}^{2} \cdot 0{,}2^{2} \cdot (1-0{,}2)^{10-2} = \frac{10!}{2!8!} \cdot 0{,}2^{2} \cdot 0{,}8^{8} = 0{,}302.$$

Vastus: tõenäosus, et 10 toote hulgas on kaks defektiga toodet, on 0,302.

Analoogselt võime leida tõenäosused ka ülejäänud m väärtuste jaoks. Defektiga toodete arvu tõenäosusjaotus on esitatud joonisel 5.11.



Joonis 5.11. Defektiga toodete arvu tõenäosusjaotus

See, kumb katsetulemus lugeda positiivseks, kumb negatiivseks, pole oluline. Olgu näiteks tulemus A positiivne ja tulemus B negatiivne. Tulemuse A esinemiste arvu m tõenäosuse saame siis valemite (5.49) ja (5.50) põhjal, kus p on A esinemise tõenäosus üksikul katsel:

$$P(X = m) = \frac{n!}{m!(n-m)!} p^m (1-p)^{n-m}. \tag{5.51}$$

Loeme nüüd positiivseks tulemuse B. Siis m on tulemuse B esinemiste arv ja p tulemuse B esinemise tõenäosus üksikul katsel. Tulemuse A esinemiste arv on $n-m$ ja tulemuse A esinemise tõenäosus üksikul katsel $1-p$. Leiame uuesti tulemuse A esinemiste arvu tõenäosuse:

$$P(X = n-m) = \frac{n!}{(n-m)!m!} (1-p)^{n-m} p^m. \tag{5.52}$$

Saime sama tulemuse, mis (5.51) puhul.

Tabelarvutuses leiab binoomjaotusele vastava tõenäosuse või jaotusfunktsiooni väärtuse funktsioon **BINOM.DIST**, mille parameetrid on järgmised:
Number_s — positiivsete sündmuste arv m;
Trials — katsete arv n;
Probability_s — üksiku positiivse tulemuse tõenäosus p;
Cumulative — tõenäosuse leidmisel 0, jaotusfunktsiooni väärtuse leidmisel 1.

Parameetri *Cumulative* väärtus toimib järgmiselt:
Cumulative $= 0$ korral leitakse tõenäosus $P(x = m)$;
Cumulative $= 1$ korral leitakse jaotussfunktsiooni väärtus $F(m) = P(x \le m) = P(x = 0) + P(x = 1) + \ldots + P(x = m)$.

Kui binoomjaotuse parameetriteks on katsete arv n ja positiivse tulemuse tõenäosus p, siis erinevate tõenäosuste leidmiseks kasutatakse tabelarvutuse funktsiooni BINOM.DIST järgmiselt:

$$\begin{aligned} P(X = m) &= \text{BINOM.DIST}(m; n; p; 0), \\ P(X \leq m) &= \text{BINOM.DIST}(m; n; p; 1), \\ P(X > m) &= 1 - \text{BINOM.DIST}(m; n; p; 1). \end{aligned}$$

N05Jaotused
N5.12

Näide 5.12. Laenud ettevõtetele ja krediidirisk

Kõige levinum risk panganduses, mis laenu andmisega kaasneb, on krediidirisk. See on risk, et laen jääb pangale tagastamata. Laenuvõtjate kontrollimisel kasutavad pangad tihti krediidiinfo kogumisele spetsialiseerunud asutusi, mis omistavad ettevõtetele krediidireitinguid. Tuntumad nendest on Moody's, Standard & Poor's, Fitch, Eestis Krediidiinfo AS.
Ühe panga laenuportfelli kuulub 250 ettevõtet, mille krediidireiting Moody'si järgi on Baa. Keskmine laenu pikkus on viis aastat. Ettevõtte reiting Baa tähendab, et viie aasta jooksul on maksejõuetuse tekkimise tõenäosus 1,94% (Cantor, Hamilton ja Tennant, 2007). Leiame tõenäosuse, et viie aasta jooksul muutub maksejõuetuks 250 ettevõttest

a) mitte ükski ettevõte;
b) maksimaalselt viis ettevõtet;
c) üle viie ettevõtte.

Määrame binoomjaotuse parameetrite väärtused. Maksejõuetuks jäämise tõenäosus on positiivse sündmuse tõenäosus $p = 0{,}0194$. Katsete arv on vastava krediidireitinguga laenuvõtjate arv $n = 250$. Tõenäosuste leidmiseks kasutame tabelarvutust.
a) Tuleb leida tõenäosus, et maksejõuetuks jääb 0 ettevõtet:

$$P(X = 0) = \text{BINOM.DIST}(0; 250; 0{,}0194; 0) \approx 0{,}0075.$$

Vastus: tõenäosus, et maksejõuetuks ei jää ükski ettevõte 250-st, on 0,0075.
b) Tuleb leida tõenäosus, et maksejõuetuks jääb maksimaalselt viis ettevõtet:

$$P(X \leq 5) = \text{BINOM.DIST}(5; 250; 0{,}0194; 1) \approx 0{,}643.$$

Vastus: tõenäosus, et maksejõuetuks jääb kuni viis ettevõtet, on 0,643.
c) Tuleb leida tõenäosus, et maksejõuetuks jääb rohkem kui viis ettevõtet:

$$P(X > 5) = 1 - 0{,}643 = 0{,}357.$$

Vastus: tõenäosus, et maksejõuetuks jääb rohkem kui viis ettevõtet, on 0,357.

Olgu panga juhtkonna poolt seatud eesmärk, et halbade (s.t tagastamata jäänud) laenude osakaal ei tohiks olla suurem kui 4% laenuportfellist. Kui suur on tõenäosus, et see piir ületatakse? Kuna 4% 250-st on 10 ettevõtet, tuleb leida tõenäosus, et maksejõuetuks jääb rohkem kui 10 ettevõtet:

$$P(X > 10) = 1 - \text{BINOM.DIST}(10; 250; 0{,}0194; 1) \approx 0{,}01.$$

Vastus: tõenäosus, et halbade laenude osakaal ületab etteantud piiri 4% laenuportfellist, on 0,01.

Binoomjaotuse keskväärtuse leidmiseks arutleme järgmiselt. Kui me teeme n katset, millel võib olla kaks võimalikku tulemust, siis iga katse tulemus vastab Bernoulli jaotust omavale juhuslikule suurusele X_i. Seega, binoomjaotusele alluv suurus $Y \sim B(n, p)$ on üksikute Bernoulli jaotusele $Be(p)$ alluvate juhuslike suuruste X_i summa $Y = X_1 + X_2 + \ldots + X_n$. Kuna positiivse tulemuse tõenäosus p jääb järjestikuste katsete ajal konstantseks, siis Bernoulli jaotusele alluvate suuruste keskväärtus on ühesugune (vt valem (5.47)):

$$E[X_1] = E[X_2] = \ldots = E[X_n] = p$$

ja dispersioon (valem (5.48))

$$\sigma^2_{X_1} = \sigma^2_{X_2} = \ldots = \sigma^2_{X_n} = p(1-p).$$

Arvestades keskväärtuse aditiivsuse omadust (5.18), võime kirjutada

$$E[Y] = E[X_1] + E[X_2] + \ldots + E[X_n] = p + p + \ldots + p = np.$$

Dispersiooni leidmisel arvestame seda, et järjestikused katsed ning järelikult suurused X_i on sõltumatud. Sellisel juhul on summa dispersioon võrdne dispersioonide summaga:

$$\sigma^2_Y = \sigma^2_{X_1} + \sigma^2_{X_2} + \ldots + \sigma^2_{X_n} = p(1-p) + p(1-p) + \ldots + p(1-p) = np(1-p).$$

Kui katsete arv on n ja positiivse tulemuse tõenäosus p, siis **binoomjaotuse keskväärtus** on

$$\mu = pn \tag{5.53}$$

Binoomjaotuse keskväärtus ja standardhälve

ja **standardhälve**

$$\sigma = \sqrt{np(1-p)}. \tag{5.54}$$

Valemist (5.53) võime avaldada positiivse tulemuse tõenäosuse

$$p = \frac{\mu}{n}. \tag{5.55}$$

Nüüd saame Bernoulli valemi (5.49) kirjutada binoomjaotuse keskväärtuse kaudu:

$$P(X = m) = C_n^m \left(\frac{\mu}{n}\right)^m \left(1 - \frac{\mu}{n}\right)^{n-m}. \tag{5.56}$$

Näide 5.13. Kindlustusettevõtte riskide juhtimine

Kindlustusettevõtte riskide juhtimisel kasutatakse mitmeid tõenäosusmudeleid. Lihtsa mudeli korral eeldatakse, et

1) kõikidel kindlustuslepingutel on kahjujuhtumi tõenäosus ühesugune (kehtib juhul, kui kindlustusportfellis on ainult üht liiki kindlustustooted);
2) kahjujuhtumid toimuvad sõltumatult (kehtib enamasti alati).

Sellisel juhul allub kahjujuhtumite arv aastas binoomjaotusele $B(n, p)$, kus n on sõlmitud lepingute arv ning p kindlustusjuhtumi tekkimise tõenäosus aastas. Kindlustusmaksete ja vajaliku omakapitali suuruse määramisel võib lähtuda kahest põhireeglist (Beneplanc ja Rochet, 2011, ptk 5).

1. Kindlustusmaksetest saadav tulu peab olema vähemalt nii suur kui kindlustushüvitiste väljamaksmisest tingitud oodatav kulu. See tingimus garanteerib, et kindlustusandjal on piisavalt raha võimalike kindlustushüvitiste väljamaksmiseks. Kui keskmine hüvitis ühe kahjujuhtumi kohta on c, siis tulu aastas

$$R \geq c\mu, \tag{5.57}$$

   kus μ on kahjujuhtumite arvu keskväärtus aastas.
2. Kuna kahjujuhtumite arv võib olla keskväärtusest suurem, peab eksisteerima omakapital K, mis moodustab turvavaru. Selle miinimumväärtuseks võetakse

$$K \geq cs\sigma, \tag{5.58}$$

   kus s on turvakordaja (*safety coefficient*) ja σ kahjujuhtumite arvu standardhälve[a].

Kaks tingimust võib kokku võtta:

$$R + K \geq c(\mu + s\sigma).$$

Kui kahjuhüvitisi tuleb maksta rohkem kui kindlustusettevõttel on varasid, ootab ettevõtet pankrot. Võttes näiteks $s = 10$, siis Tšebõšovi teoreemist (vt alaptk 3.5) tõenäosus, et kindlustusjuhtumite arv erineb keskväärtusest rohkem kui 10σ võrra, on väiksem kui $1/10^2$. See tähendab, et pankroti tõenäosus on sel juhul väiksem kui 1%. Tegelikult on see veel väiksem, sest Tšebõšovi teoreemi järgi leitud tõenäosus jaguneb kaheks: vähem kui $\mu - 10\sigma$ ja rohkem kui $\mu + 10\sigma$. Lisaks sellele kehtib Tšebõšovi teoreem üldjuhul ja ei arvesta konkreetset jaotusseadust.
Olgu ühel kindlustusettevõttel 5000 lepingut, millel kindlustusjuhtumi tekkimise tõenäosus aastas on 5%. Keskmine hüvitis ühe kahjujuhtumi kohta on 1200 eurot. Leida minimaalne vajalik tulu aastas ning minimaalne omakapitali suurus, kui $s = 10$.
Aastas aset leidvate kahjujuhtumite arvu keskväärtus (valemist (5.53))

$$\mu = 0{,}05 \cdot 5000 = 250$$

ning minimaalne vajalik tulu (valemist (5.57))

$$R = 1200 \cdot 250 = 1200 \cdot 270 = 300000.$$

Kahjujuhtumite arvu standardhälve (valemist (5.54))

$$\sigma = \sqrt{5000 \cdot 0{,}05 \cdot (1 - 0{,}05)} \approx 15$$

ning minimaalne omakapital (valemist (5.58))

$$K = 1200 \cdot 10 \cdot 15 = 180000.$$

Vastus: minimaalne vajalik tulu aastas on 300 tuhat eurot ning minimaalne omakapitali suurus 180 tuhat eurot.

[a] Eestis on kindlustusandja omavahendite miinimum määratud kindlustustegevuse seadusega.

Näites 5.13 kasutasime minimaalse omakapitali leidmiseks valemit (5.58), kus kordaja s võeti ette. On võimalik ka teistsugune lähenemine: võtta ette kriteerium β, nii et pankroti tekkimise tõenäosus oleks sellest väiksem. Siis peab kehtima tingimus

$$P(X > m) < \beta,$$

kus X on kahjujuhtumite arv aastas. Vastandsündmuse jaoks kehtib tingimus

$$P(X \leq m) > \alpha,$$

kus $\alpha = 1 - \beta$. Viimane on tingimus jaotusfunktsiooni väärtuse $F(m)$ jaoks: tuleb leida vähim m väärtus, mille korral

$$F(m) > \alpha. \tag{5.59}$$

See on jaotusfunktsiooni väärtuse leidmise pöördülesanne.

Tabelarvutuses aitab viimast probleemi lahendada funktsioon **BINOM.INV**. See funktsioon leiab minimaalse positiivsete tulemuste arvu m, mille korral kehtib tingimus (5.59). Funktsiooni parameetrid on järgmised:

Trials — katsete arv n;
Probability_s — üksiku positiivse tulemuse tõenäosus p;
Alpha — kriteerium α tingimusest (5.59), $0 \le \alpha \le 1$.

Näide 5.14. Kindlustusettevõtte minimaalne omakapital

N05Jaotused
N5.14

Leiame näites 5.13 toodud andmetele tuginedes, kui suur peab olema selle kindlustusettevõtte omakapital, et pankroti tekkimise tõenäosus oleks väiksem kui 1%.
Lepingute arv $n = 5000$ ja kahjujuhtumi tõenäosus $p = 0{,}05$. Kriteerium $\beta = 0{,}01$, järelikult $\alpha = 1 - 0{,}01 = 0{,}99$. Minimaalne kahjujuhtumite arv m, mille korral on täidetud tingimus (5.59):

$$m = \text{BINOM.INV}(5000; 0{,}05; 0{,}99) = 286.$$

Sellise arvu kahjujuhtumite hüvitamiseks vajalike ressursside suurus (tulu aastas R pluss omakapital K)

$$R + K = 1200 \cdot 286 = 343200.$$

Kuna vajalik tulu aastas oli $R = 300000$, siis

$$K = 343200 - 300000 = 43200.$$

Vastus: selleks, et näites 5.13 kirjeldatud kindlustusettevõtte pankroti tekkimise tõenäosus oleks väiksem kui 1%, peab minimaalne omakapital olema 43,2 tuhat eurot. See on oluliselt väiksem kui näites 5.13 leitud 180 tuhat eurot, kus me kasutasime kriteeriumit $K \ge 10\sigma$.

5.8. Poissoni jaotus

Mingi aja jooksul poodi sisenevate ostjate arv on juhuslik suurus. Kui me loendame sisenenud ostjaid järjestikuste viieminutiliste intervallide

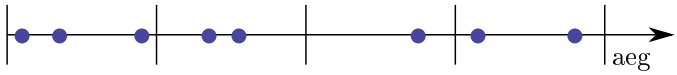
jooksul, võime saada näiteks arvud $m = \{5, 7, 2, 0, 5, 6\}$. Selle vaatluse põhjal siseneb viie minuti jooksul keskmiselt 4,17 inimest.

Enamasti teeb otsuse, kas siseneda poodi või mitte, iga inimene sõltumatult teiste käitumisest. Sellisel juhul on erinevate inimeste poodi sisenemised sõltumatud sündmused.

Poissoni jaotuse esinemine

Poissoni jaotus kirjeldab diskreetset juhuslikku suurust, mille väärtuseks on haruldaste sündmuste arv m teatud intervallis. Sündmused toimuvad

- juhuslikult ja üksteisest sõltumatult;
- harva;
- kindlas ajaintervallis või ruumipiirkonnas.



Joonis 5.12. Kindlas ajavahemikus esinevate sõltumatute sündmuste arv on juhuslik suurus, mis allub Poissoni jaotusele

Me võime 5-minutiliste intervallide asemel vaadelda lühemaid intervalle, näiteks sekundilisi, ja vaadata, kas mingis sekundis keegi astub poodi või mitte. Sellisel juhul on tegemist kaheväärtuselise tunnusega: „siseneb/ei sisene“, mis allub binoomjaotusele. Positiivsete tulemuste „sisenes“ arvu m tõenäosuse saame järelikult leida valemist (5.56), sest meil on leitud 5 minuti jooksul sisenevate inimeste arvu keskväärtus $\mu = 4{,}17$. Sekundilisi intervalle on viies minutis 300, järelikult katsete arv $n = 300$. Näiteks tõenäosus, et viie minuti jooksul astub poodi 6 inimest, on

$$P(X = 6) = C_{300}^{6} \left(\frac{4{,}17}{300}\right)^{6} \left(1 - \frac{4{,}17}{300}\right)^{294} \approx 0{,}1133. \qquad (5.60)$$

Aga kuidas arvutada tõenäosust siis, kui sekundi jooksul astub poodi korraga kaks inimest? Sel juhul ei saa me kasutada binoomjaotust, sest binoomjaotusele alluv suurus peab olema kaheväärtuseline („siseneb/ei sisene“). Ei ole võimalust „siseneb kaks inimest“. Sellest probleemist ülesaamiseks jagame 5-minutilise intervalli veel lühemateks osadeks, millega suurendame katsete arvu n.

Lisas A.3 on näidatud, mis juhtub valemiga (5.56), kui katsete arv n läheneb lõpmatusele. Kui kasutame keskväärtuse jaoks tähistust $\lambda = \mu$, siis

$$\lim_{n\to\infty} \left[C_n^m \left(\frac{\lambda}{n}\right)^m \left(1 - \frac{\lambda}{n}\right)^{n-m} \right] = e^{-\lambda} \frac{\lambda^m}{m!}. \qquad (5.61)$$

Näeme, et tõenäosus $P(X = m)$ ei sõltu enam katsete arvust. **Binoomjaotuse piirjuht** katsete arvu lähenemisel lõpmatusele on **Poissoni jaotus**. Oma nimetuse on see jaotus saanud Prantsuse matemaatiku ja füüsiku Siméon Denis Poissoni (1781–1849) järgi.

Poissoni valem

Kui λ on keskmine haruldaste sündmuste arv mingis intervallis, siis m sündmuse toimumise tõenäosus selles intervallis on leitav **Poissoni valemi** abil:

$$P(X = m) = e^{-\lambda}\frac{\lambda^m}{m!}, \tag{5.62}$$

kus $\lambda > 0$ on Poissoni jaotuse keskväärtus ja $e = 2{,}718\ldots$ on Euleri arv.

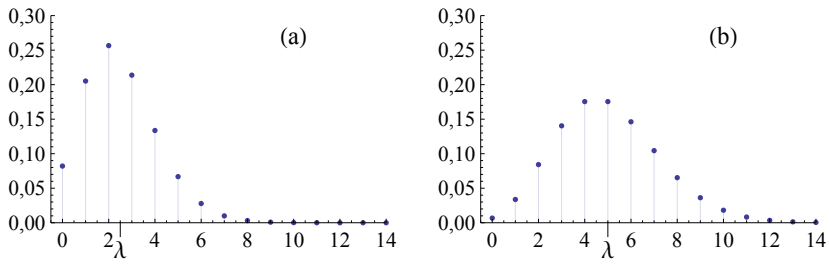
Kasutame Poissoni valemit (5.62) leidmaks tõenäosust, et poodi astub viie minuti jooksul kuus inimest, kui sisenejate arvu keskväärtus $\lambda = 4{,}17$:

$$P(X = 6) = e^{-4{,}17}\frac{4{,}17^6}{6!} \approx 0{,}1128. \tag{5.63}$$

Lõpmatu arvu katsete korral saame ligikaudu sama tulemuse, mis binoomjaotuse kasutamisel katsete arvu $n = 300$ korral (arvutus (5.60)).

Poissoni jaotusel on üks parameeter: keskväärtus λ, mis määrab täielikult ära Poissoni jaotusele alluva juhusliku suuruse jaotusseaduse (vt joonis 5.13). Kui mingi juhuslik suurus X allub Poissoni jaotusele keskväärtusega λ, tähistatakse seda järgmiselt:

$$X \sim \mathrm{Pois}(\lambda).$$



Joonis 5.13. Kaks erinevat Poissoni jaotust: (a) $\lambda = 2{,}5$; (b) $\lambda = 5$

Kuna Poissoni jaotusele alluva juhusliku suuruse väärtused saadakse sündmuste loendamisel, siis võib see juhuslik suurus omandada vaid naturaalarvulisi väärtusi, $m \in \{0; 1; 2; \ldots\}$. Poissoni jaotuse keskväärtus λ võib aga olla positiivne murdarv. Jooniselt 5.13 on näha,

et väikeste keskväärtuste korral on Poissoni jaotuse kõver parempoolse asümmeetriaga. See on seletatav sellega, et jaotus on vasakult poolt piiratud: loendamisel saadav tulemus ei saa olla väiksem kui null. Suuremate keskväärtuste korral muutub jaotus sümmeetrilisemaks, sest null on keskväärtusest kaugemal.

Näide 5.15. Telefonikõnede arvu jaotus

Telefonifirma Bell registreeris seitsme päeva jooksul, kui palju kõnesid tehti taksofonidest viieminutiliste intervallide jooksul. Iga päev vaadeldi 20 intervalli ajavahemikus 12:00–14:00 ja vaatluse all oli neli taksofoni, millest tehtud kõnede arv summeeriti. Tulemused on toodud tabelis (Thorndike, 1926). Näiteks selliseid viieminutilisi intervalle, mille jooksul ei tehtud ühtegi kõnet, oli 140 intervalli hulgas viis. Selliseid intervalle, mille jooksul tehti neljast taksofonist kokku kolm kõnet, oli 38.

Kõnede arv intervallis	0	1	2	3	4	5	6	7
Intervallide arv ehk sagedus f	5	33	24	38	23	9	4	4

Kokku vaadeldi 140 intervalli, mille jooksul tehti kõnesid kokku 384. Järelikult keskmine kõnede arv viieminutilise intervalli jooksul on

$$\lambda = \frac{\sum f_i m_i}{\sum f_i} = \frac{384}{140} = 2{,}74.$$

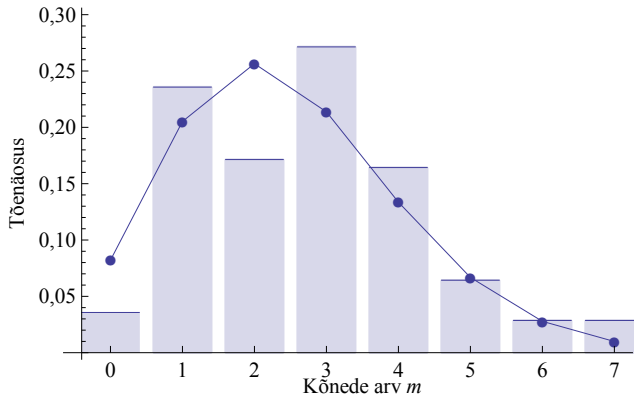
Mingi aja jooksul tehtud telefonikõnede arv on juhuslik suurus. Kuna vaatluse all olid tänaval seisvad taksofonid, siis järjestikused kõned olid üksteisest sõltumatud, tehtud erinevate inimest poolt. Järelikult telefonikõnede arvu jaotust kirjeldab Poissoni jaotus.

Võrdluseks leiame iga m väärtuse jaoks statistilise tõenäosuse

$$p_i^* = \frac{f_i}{384}$$

ja keskväärtusega $\lambda = 2{,}74$ Poissoni jaotusele vastava teoreetilise tõenäosuse. Tulemused (ümardatult) on toodud järgnevas tabelis ja joonisel 5.14.

Kõnede arv intervallis	0	1	2	3	4	5	6	7
Statistiline tõenäosus p^*	0,036	0,236	0,171	0,271	0,164	0,064	0,029	0,029
Teoreetiline tõenäosus	0,064	0,177	0,242	0,221	0,152	0,083	0,038	0,015



Joonis 5.14. Viieminutiliste intervallide jooksul tehtud kõnede arvu jaotus. Tulbad vastavad statistilisele tõenäosusele, joon näitab vastava Poissoni jaotuse alusel leitud teoreetilist tõenäosust

Poissoni jaotust kasutatakse tihti nõudluse prognoosimisel ja varude juhtimisel, liiklusvoogude prognoosimisel, ressursside planeerimisel (töötajate arv klienditeeninduses), töökindluse teoorias, füüsikas, infotehnoloogias — kõikjal, kus teatud aja jooksul võib toimuda juhuslik arv sõltumatuid sündmusi (kaupluse või teenindussaali külastamine, ostuotsused, infopäringud, rikete esinemine, radioaktiivne lagunemine jne). On vaja vaid leida sündmuste arvu keskväärtus mingi aja jooksul ja me saame leida tõenäosuse, et sündmuste arv selles ajavahemikus on võrdne mingi väärtusega, suurem või väiksem mingist väärtusest.

Näide 5.16. Sisenevate ostjate arvu tõenäosuse leidmine

Poodi sisenevate ostjate keskmine arv minutis on 3,4. Eeldades, et sisenejate arv allub Poissoni jaotusele, leida tõenäosus, et ühe minuti jooksul

a) ei sisene ühtki ostjat;
b) siseneb täpselt üks ostja;
c) siseneb kaks või rohkem ostjat.

Kasutame valemit (5.62), kus keskväärtus $\lambda = 3{,}4$.

a) Kui keegi ei sisene, siis $m = 0$ ja

$$P(X = 0) = e^{-3,4}\frac{3,4^0}{0!} = e^{-3,4} \approx 0{,}0334.$$

Vastus: tõenäosus, et minuti jooksul ei sisene ühtki ostjat, on 0,0334.

b) Kui siseneb 1 ostja, siis $m = 1$ ja

$$P(X = 1) = e^{-3,4}\frac{3,4^1}{1!} = e^{-3,4} \cdot 3,4 \approx 0,1135.$$

Vastus: tõenäosus, et minuti jooksul siseneb täpselt üks ostja, on 0,1135.
c) Kui soovime leida tõenäosust, et minuti jooksul siseneb kaks või rohkem ostjat, arvestame jaotusseaduse normeerimistingimusega: kõikvõimalike väärtuste tõenäosuste summa on 1. Järelikult

$$P(X = 0) + P(X = 1) + P(X \geq 2) = 1.$$

Sellest tingimusest

$$\begin{aligned} P(X \geq 2) &= 1 - (P(X = 0) + P(X = 1)) = \\ &= 1 - (0,0334 + 0,1135) = 0,8531. \end{aligned}$$

Vastus: tõenäosus, et minuti jooksul siseneb kaks või rohkem ostjat, on 0,8531.

Poissoni jaotuse keskväärtus on antud mingi kindla intervalli kohta. Kui on vaja leida tõenäosus mingi teise pikkusega intervalli jaoks, tuleb kasutada sellele intervallile vastavat keskväärtust. Kui keskväärtus ühe minuti kohta on λ, siis keskväärtus viie minuti kohta on 5λ.

Intervalli muutmine

Kui mingis intervallis $X \sim \text{Pois}(\lambda)$, siis k intervallis $Y \sim \text{Pois}(k\lambda)$.

Näide 5.17. Telefonikõnede esinemissagedus erinevates ajavahemikes

Näites 5.15 oli toodud andmed telefonikõnede arvu kohta viie minuti pikkustes intervallides. Viie minuti jooksul tehtud kõnede arvu keskväärtus on $\lambda = 2,74$.
Leiame järgmised tõenäosused:

a) ühe minuti jooksul ei tehta ühtegi kõnet;
b) tunni aja jooksul tehakse üle 50 kõne.

Nende tõenäosuste leidmiseks kasutame vastavate intervallide jaoks leitud keskväärtusi.

a) üheminutilisi intervalle on viies minutis 5. Järelikult keskmine kõnede arv ühe minuti jooksul on

$$\frac{2{,}74}{5} = 0{,}548.$$

Kui ei tehta ühtegi kõnet, siis kõnede arv $m = 0$. Vastav tõenäosus

$$P(X = 0) = e^{-0{,}548}\frac{0{,}548^0}{0!} = e^{-0{,}548} = 0{,}578.$$

Vastus: tõenäosus, et ühe minuti jooksul ei tehta ühtegi kõnet, on 0,578;
b) tunnis on viieminutilisi intervalle 20, järelikult kõnede arvu keskväärtus tunnis peab olema 20 korda suurem:

$$2{,}74 \cdot 20 = 54{,}8.$$

Et leida, kui suure tõenäosusega on kõnede arv suurem kui 50, arvestame jaotusseaduse normeerimistingimust:

$$P(X = 0) + P(X = 1) + \ldots + P(X = 50) + P(X > 50) = 1.$$

Siit

$$P(X > 50) = 1 - (P(X = 0) + P(X = 1) + \ldots + P(X = 50)) = \\ = 1 - P(X \leq 50).$$

Tõenäosuse $P(X \leq 50)$ leidmiseks kasutame tabelarvutuse funktsiooni POISSON.DIST(50;54,8;1), s.t parameeter *Cumulative* on 1. See annab tulemuseks 0,98. Nüüd

$$P(X > 50) = 1 - 0{,}98 = 0{,}02.$$

Vastus: tõenäosus, et tunni aja jooksul tehakse rohkem kui 50 kõnet, on 0,02.

Tabelarvutuses on Poissoni jaotuse tõenäosuse või jaotusfunktsiooni leidmiseks funktsioon **POISSON.DIST**. Parameeter *Cumulative* on tõenäosuse leidmisel 0. Jaotusfunktsiooni väärtuse $F(m) = P(X \leq m)$ leidmiseks tuleb panna parameetri *Cumulative* väärtuseks 1. Näiteks

$$\text{POISSON.DIST}(2;1{,}5;1) = \text{POISSON.DIST}(0;2{,}5;0) + \\ + \text{POISSON.DIST}(1;2{,}5;0) + \text{POISSON.DIST}(2;2{,}5;0),$$

sest

$$P(X \le 2) = P(X = 0) + P(X = 1) + P(X = 2).$$

Kui Poissoni jaotuse keskväärtus on λ, siis erinevate tõenäosuste leidmiseks kasutatakse tabelarvutuses funktsiooni POISSON.DIST järgmiselt:

$$\begin{aligned} P(X = m) &= \text{POISSON.DIST}(m; \lambda; 0), \\ P(X \le m) = F(m) &= \text{POISSON.DIST}(m; \lambda; 1), \\ P(X > m) = 1 - F(m) &= 1 - \text{POISSON.DIST}(m; \lambda; 1). \end{aligned}$$

Näide 5.18. Bussireisijate arv

N05Jaotused
N5.18

Linnadevahelist bussiliini teenindava bussiettevõtte statistika näitab, et keskmiselt reisib sellel liinil 36,7 reisijat. Liini teenindavas bussis on 49 kohta. Leida

a) tõenäosus, et bussi tuleb täpselt 49 reisijat;
b) tõenäosus, et bussi tuleb kuni 49 reisijat;
c) tõenäosus, et bussi peale soovib tulla rohkem inimesi, kui on kohti ja vähemalt üks inimene jääb maha;
d) kui buss teeb aastas 240 reisi, siis mitmel reisil võib mõni inimene maha jääda?

Tõenäosuste leidmiseks kasutame tabelarvutust.

a) Tuleb leida tõenäosus, et bussi tuleb täpselt 49 reisijat:

$$P(X = 49) = \text{POISSON.DIST}(49; 36{,}7; 0) \approx 0{,}00883.$$

Vastus: tõenäosus, et bussi tuleb täpselt 49 reisijat, on 0,00883.

b) Tuleb leida tõenäosus, et bussi tuleb kuni 49 reisijat:

$$P(X \le 49) = \text{POISSON.DIST}(49; 36{,}7; 1) \approx 0{,}979.$$

Vastus: tõenäosus, et bussi tuleb kuni 49 reisijat, on 0,979.

c) Tuleb leida tõenäosus, et sõita soovijaid on rohkem kui 49:

$$P(X > 49) = 1 - 0{,}979 = 0{,}021.$$

Vastus: tõenäosus, et sõita soovijaid on rohkem kui bussis kohti, on 0,021.

d) Reiside arv kokku on 240 ja tõenäosus, et mõni inimene jääb maha, on 0,021. Järelikult selliste väljumiste arv

$$0{,}021 \cdot 240 \approx 5{,}1.$$

Vastus: mahajääjaid võib olla umbes viiel reisil aastas.

Poissoni jaotuse dispersioon

Poissoni jaotuse **dispersioon** on võrdne jaotuse keskväärtusega

$$\sigma^2_{Pois} = \lambda \tag{5.64}$$

ja standardhälve

$$\sigma_{Pois} = \sqrt{\lambda}. \tag{5.65}$$

Seost (5.64) saab kasutada kontrollimaks, kas mingi suurus allub Poissoni jaotusele. Kui see seos ligikaudu kehtib, siis on tegemist Poissoni jaotusega. Täpsemaks testimiseks võib aga kasutada χ^2-testi (vt ptk 7.12).

Kui me soovime kasutada Poissoni jaotust mingi empiirilise juhusliku suuruse tõenäosuste arvutamiseks, tuleb eelnevalt läbi viia vaatlus ja selle alusel leida jaotuse keskväärtus. Vaatluse käigus loendatakse sündmuste arv paljudes ühesuguse pikkusega intervallides. Intervallid peavad olema valitud nii, et jaotuse keskväärtus oleks konstantne. Näites 5.15 kasutatud allika korral vaadeldi 20 intervalli erinevatel päevadel ajavahemikus 12:00–14:00. Kui vaadelda mõnda teist ajavahemikku, näiteks 18:00–20:00, on jaotuse keskväärtus teistsugune. Samamoodi poodi sisenevate ostjate arvu jaotust analüüsides peame leidma piirid, mille jooksul keskväärtus ei muutu. Näiteks kellaaegadel 10:00–12:00 ja 16:00–18:00 on jaotuse keskväärtus ilmselt erinev. Seda arvestatakse kassapidajate töö planeerimisel.

Poissoni jaotust kasutades on massteeninduse ehk **järjekorrateoorias** (*queuing theory*) tuletatud praktilised valemid järjekorra pikkuse ja keskmise ooteaja leidmiseks (vt näiteks Evald Übi ja Kadrin Keres „Rakendusmatemaatika“ või Tiiu Paas „Kvantitatiivsed meetodid majanduses“). Lähtesuurusteks on keskmine sisenejate arv ehk sisendvoo intensiivsus ja keskmine teenindatud klientide arv (teenindaja töökiirus). Valemeid saab kasutada lihtsamate, ühe teenindajaga süsteemide analüüsimisel. Keerulisemate süsteemide korral, kus on mitu teenindajat või kus tuleb järjestikku läbida mitu teeninduspunkti, kasutatakse vastava tarkvara abil läbiviidud simulatsioone. Selline modelleerimine võimaldab optimeerida teenindusprotsessi.

5.9. Eksponentjaotus

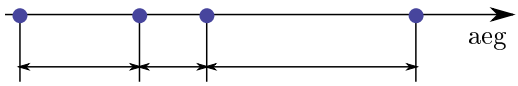
Kui mingis ajavahemikus toimuvate sõltumatute sündmuste arv allub Poissoni jaotusele, siis ajavahemik nende sündmuste vahel allub **eksponentjaotusele**. Eksponentjaotusele võib alluda näiteks

- ajavahemik kahe järjestikuse kliendi saabumise vahel;
- ajavahemik kahe järjestikuse päringu vahel veebilehelt;

- uue seadme (raadio, arvuti) tööaeg esimese rikkeni[4];
- aeg mingi demograafilise sündmuseni (inimese surm, kahe lapse sünni vaheline aeg sünnitusmajas).

Eksponentjaotuse esinemine

Eksponentjaotusele allub ajavahemik kahe sõltumatult toimuva samaliigilise järjestikuse sündmuse vahel.



Joonis 5.15. Eksponentjaotusele allub ajavahemik kahe järjestikuse sõltumatu sündmuse vahel

Protsessi, mille käigus toimuvate sündmuste arv allub Poissoni jaotusele ja sündmustevaheline aeg eksponentjaotusele, nimetatakse **Poissoni protsessiks**. Aega kahe sündmuse vahel nimetatakse ka **ooteajaks** (*waiting time*).

Eksponentjaotuse jaotustihedus

Eksponentjaotuse **jaotustihedus**

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{kui } x \geq 0 \\ 0, & \text{kui } x < 0 \end{cases}, \qquad (5.66)$$

kus λ on jaotuse parameeter ja $e = 2{,}718\ldots$ on Euleri arv.

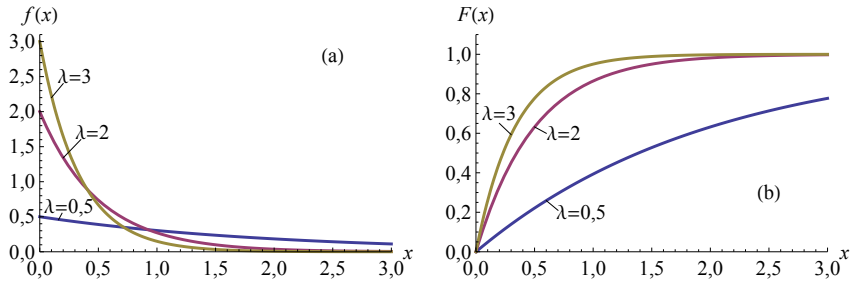
Erinevate parameetri λ väärtustega jaotustiheduse graafikud on toodud joonisel 5.16(a). Paneme tähele, et jaotustiheduse graafik lõikab y-telge punktis $(0, \lambda)$. Seda näeme ka jaotustiheduse valemist (5.66):

$$f(0) = \lambda e^{-\lambda \cdot 0} = \lambda e^{0} = \lambda.$$

Kui juhuslik suurus X allub eksponentjaotusele parameetriga λ, märgitakse seda järgmiselt:

$$X \sim \mathrm{Exp}(\lambda).$$

[4]Uue seadme rike võib olla põhjustatud defektsetest komponentidest või tootmisprotsessi parameetrite kõikumisest. Seadme vananedes hakkab rolli mängima kulumine ning edaspidi võib rikke tekkimise tõenäosus sõltuda seadme elueast. Sellisel juhul kasutatakse näiteks gammajaotust või Weibulli jaotust.



Joonis 5.16. Eksponentjaotuse jaotustiheduse (a) ja jaotusfunktsiooni (b) graafikud erineva λ korral. Jaotustiheduse graafik lõikab y-telge punktis $(0, \lambda)$

N05 Jaotused
N5.19

Näide 5.19. Päringud veebiserverist

2003. aasta algul sai Audentese Ülikoolis valmis õppeinfosüsteemi veebiliides, mis võimaldas üliõpilastel vaadata oma tulemusi. Ajavahemikul 28.01.2003–25.02.2003 tehti selle kaudu 1124 päringut. Tabelis on toodud veebiserveri logifailist pärinevad andmed. Neid juhtumeid, kus kahe päringu vaheline aeg oli väiksem kui 1 minut, oli 170. Neid, kus kahe järjestikuse päringu vaheline aeg oli 1–3 minutit, oli 185 jne. Aegade aritmeetiline keskmine tuli 14,46 minutit ja standardhälve 18,11 minutit.

Aeg kahe päringu vahel, min	Vahemiku laius Δx	Sagedus n_i	Suhteline sagedus w_i	Suhteline sagedus-tihedus f_i^*	Eksponent-jaotuse jaotustihedus $f(x)$
0–1	1	170	0,151 2	0,151 2	0,066 8
1–3	2	185	0,164 6	0,082 3	0,060 2
3–5	2	129	0,114 8	0,057 4	0,052 4
5–15	10	284	0,252 7	0,025 3	0,034 6
15–50	35	281	0,250 0	0,007 14	0,007 31
50–80	30	62	0,055 2	0,001 839	0,000 772
80–100	20	13	0,011 6	0,000 578	0,000 137

Iga vahemiku jaoks leiti suhteline sagedus

$$w_i = \frac{n_i}{N},$$

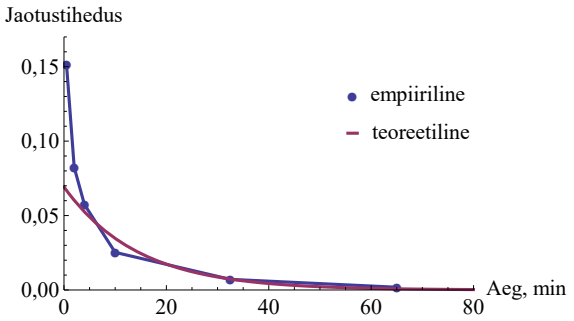
kus vaatluste koguarv $N = \sum n_i = 1124$. Et võrrelda, kas jaotus sarnaneb eksponentjaotusele, leiti ka empiiriline jaotustihedus: suhteline sagedus jagatud klassi laiusega:

$$f_i^* = \frac{w_i}{\Delta x}.$$

Viimases veerus on võrdluseks leitud eksponentjaotuse jaotustihedus $f(x)$. Jaotuse parameetri leidmiseks kasutati empiirilistest andmetest leitud aritmeetilist keskmist:

$$\lambda = 1/14{,}46 \approx 0{,}0692.$$

Tabeli kahe viimase veeru põhjal on koostatud diagramm, kus markeritega joon tähistab empiirilist jaotust. Empiirilise jaotuse esitamiseks kasutatakse suhtelist sagedustihedust (tabeli eelviimane veerg).



Eksponentjaotuse jaotusfunktsiooni leidmiseks kasutame valemit (5.29). Alumiseks rajaks võtame 0, sest $x < 0$ korral on jaotustihedus $f(x) = 0$.

$$\int_0^a f(x)dx = \int_0^a \lambda e^{-\lambda x}dx = \lambda \int_0^a e^{-\lambda x}dx = \lambda \left[-\frac{1}{\lambda}e^{-\lambda x}\right]_0^a =$$
$$= -\left(e^{-\lambda a} - e^0\right) = 1 - e^{-\lambda a}.$$

Eksponentjaotuse **jaotusfunktsioon**

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{kui } x \geq 0 \\ 0, & \text{kui } x < 0 \end{cases}, \qquad (5.67)$$

kus λ on jaotuse parameeter.

Eksponentjaotuse jaotusfunktsioon

Jaotusfunktsiooni graafikud erineva λ korral on toodud joonisel 5.16(b). Jaotusfunktsiooni kasutades on võimalik leida järgmisi tõenäosusi:

$$P(X < a) = F(a), \qquad (5.68)$$
$$P(X > a) = 1 - F(a), \qquad (5.69)$$
$$P(a < X < b) = F(b) - F(a). \qquad (5.70)$$

„Mäluta“ jaotus

Eksponentjaotuse tähtis omadus on „mälu“ puudumine, s.t eksponentjaotus on **„mäluta“ jaotus**. Milles see väljendub? Olgu näiteks T aeg maapoes kahe järjestikuse ostja sisenemise vahel, $T \sim \text{Exp}(\lambda)$. Oletame, et müüja on järgmist ostjat oodanud juba vähemalt s minutit. Kui suur on tõenäosus, et ta peab ootama veel vähemalt t minutit? Meil tuleb leida tinglik tõenäosus $P(T > s+t|T > s)$. Kasutame tingliku tõenäosuse leidmise valemit (4.10) ja valemeid (5.69) ning (5.67):

$$P(T > s+t|T > s) = \frac{P(T > s+t)}{P(T > s)} = \frac{1-(1-e^{-\lambda(s+t)})}{1-(1-e^{-\lambda s})} = \\ = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t}.$$

Tõenäosus, et tuleb veel oodata vähemalt t minutit, ei sõltu s väärtusest, s.t sellest, kui kaua on juba oodatud. See tähendabki „mälu“ puudumist.

Eksponentjaotuse keskväärtus ja standardhälve

Eksponentjaotuse **keskväärtus** ja **standardhälve** on võrdsed:

$$\mu = \frac{1}{\lambda}, \tag{5.71}$$

$$\sigma = \frac{1}{\lambda}. \tag{5.72}$$

Praktikas saab vaatlusandmete põhjal leida ajavahemike keskväärtuse μ. Eksponentjaotuse kasutamiseks tuleb siis arvutada jaotuse parameeter $\lambda = 1/\mu$.

N05Jaotused N5.20

Näide 5.20. Piduriketaste eluiga

Auto piduriketaste eluiga allub eksponentjaotusele keskväärtusega 10 aastat. Kui piduriketaste garantiiaeg on 5 aastat, kui suure tõenäosusega kestavad pidurikettad kauem kui garantiiaeg?
Kui keskväärtus $\mu = 10$ aastat, siis parameeter $\lambda = 1/10 = 0{,}1$. Tõenäosus, et eluiga on suurem kui 5 aastat, leitakse valemitest (5.69) ja (5.67):

$$P(X > 5) = 1 - F(5) = e^{-0{,}1 \cdot 5} \approx 0{,}61.$$

Vastus: tõenäosus, et pidurikettad kestavad kauem kui garantiiaeg 5 aastat, on 0,61.

Tabelarvutuses on eksponentjaotuse jaotustiheduse ja jaotusfunktsiooni leidmiseks funktsioon **EXPON.DIST**. Jaotustiheduse leidmisel

on parameeter *Cumulative* 0, jaotusfunktsiooni leidmiseks tuleb *Cumulative* väärtuseks panna 1. Kui eksponentjaotuse parameeter on λ, siis tõenäosuste leidmiseks kasutatakse vastavat tabelarvutusfunktsiooni järgmiselt:

$$\begin{aligned} P(X < a) &= F(a) && = \text{EXPON.DIST}(a;\lambda;1),\\ P(X > a) &= 1 - F(a) && = 1 - \text{EXPON.DIST}(a;\lambda;1),\\ P(a < X < b) &= F(b) - F(a) && = \text{EXPON.DIST}(b;\lambda;1) - \\ & && -\text{EXPON.DIST}(a;\lambda;1). \end{aligned}$$

Mõnikord tuleb lahendada probleem, kus tõenäosus ja seega jaotusfunktsiooni väärtus $F(x)$ on ette antud ning tuleb leida sellele vastav juhusliku suuruse väärtus x. Siis tuleb valemist (5.67) avaldada x.

$$\begin{aligned} F(x) &= 1 - e^{-\lambda x}\\ e^{-\lambda x} &= 1 - F(x)\\ -\lambda x &= \ln(1 - F(x)). \end{aligned}$$

Viimasest avaldisest saame, et

$$x = -\frac{\ln(1 - F(x))}{\lambda}. \tag{5.73}$$

Näide 5.21. Garantiiaja pikkuse leidmine

Analüüsides rikete statistikat, on tootja kindlaks teinud, et keskmiselt on tema toodetud seadmel 0,1 riket aastas. Kui pikka garantiiaega peaks ta oma seadmele lubama, et garantiiajal tagastavaid seadmeid ei oleks üle 5%?
Eeldame, et aeg, mis kulub rikke tekkimiseni, allub eksponentjaotusele. Meil on vaja leida a, nii et $P(X < a) = F(a) = 0{,}05$. Eksponentjaotuse parameeter $\lambda = 0{,}1$. Kasutame valemit (5.73):

$$x = -\frac{\ln(1 - F(x))}{\lambda} = -\frac{\ln(1 - 0{,}05)}{0{,}1} \approx 0{,}513.$$

Vastus: garantiiajaks tuleks panna 0,5 aastat.

5.10. Normaaljaotus

Järgnevalt vaatame pideva juhusliku suuruse üht sagedamini kasutatavat teoreetilist jaotust, mida nimetatakse normaaljaotuseks (*normal distribution*) ehk Gaussi jaotuseks (*Gaussian distribution*). Saksa matemaatik Carl Friedrich Gauss (1777–1855) formuleeris selle jaotuse kui

üldise mudeli eksperimendi vigade jaotuse modelleerimiseks. XIX sajandil demonstreeris Šoti teadlane James Clerk Maxwell (1831–1879), et normaaljaotus pole mitte ainult sobiv matemaatiline mudel, vaid see jaotus esineb ka looduses. Termin „normaaljaotus“ võeti kasutusele XIX sajandi lõpul märkimaks, et see kirjeldab tüüpilist, standardset jaotust.

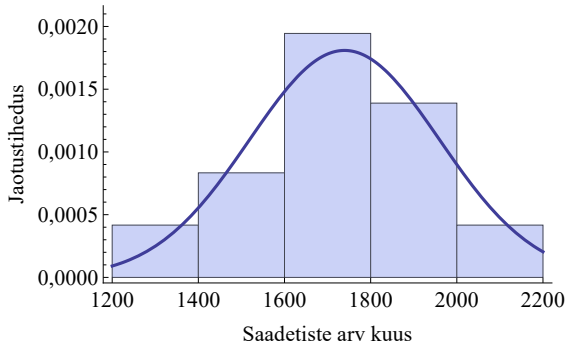
N05Jaotused
N5.22

Näide 5.22. Kullerfirma saadetiste arvu jaotus

Kullerfirma viis läbi oma töökoormuse analüüsi aastatel 2000 kuni 2003. Analüüsiti kohaletoimetatud väikesaadetiste arvu kuus. Saadetiste arvu aritmeetiline keskmine oli $\bar{x} = 1739{,}4$ ja standardhälve $\sigma = 220{,}5$ saadetist kuus. Et saada paremat ülevaadet töökoormuse varieerumisest, rühmitati vaatlusandmed klassidesse laiusega $\Delta x = 200$. Iga klassi jaoks leiti suhteline sagedus $w_i = n_i/N$, kus vaatluste koguarv $N = \sum n_i = 36$. Et võrrelda, kas jaotus sarnaneb normaaljaotusele, leiti ka empiiriline jaotustihedus: suhteline sagedus jagatud klassi laiusega, $f_i^* = w_i/\Delta x$.

Sagedusklass	Sagedus n_i	Suhteline sagedus w_i	Empiiriline jaotustihedus f_i^*	Normaaljaotusele vastav jaotustihedus f_i
1200–1400	3	0,083	0,000417	0,000249
1400–1600	6	0,167	0,000833	0,001004
1600–1800	14	0,389	0,001944	0,001780
1800–2000	10	0,278	0,001389	0,001387
2000–2200	3	0,083	0,000417	0,000475

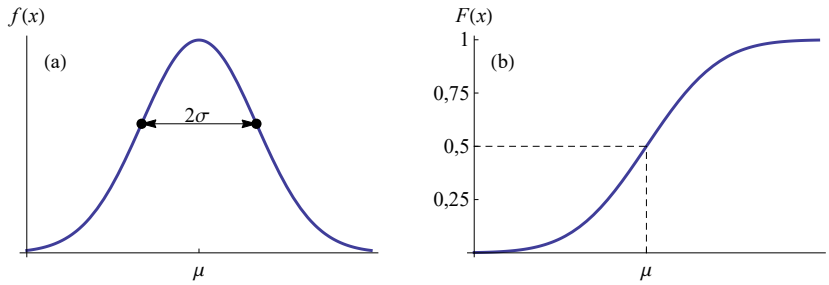
Väikesaadetiste arvu jaotus on esitatud joonisel, kus empiirilist jaotustihedust näitavatele tulpadele on lisatud keskväärtusega 1739,4 ja standardhälbega 220,5 määratud normaaljaotuse jaotuskõver.



Saadetiste arvu kuus võime vaadelda pideva juhusliku suurusena, mille konkreetne väärtus sõltub väga paljudest faktoritest. On võimalik näidata, et selline suurus allub normaaljaotusele, mille tüüpiline jaotustiheduse graafik on toodud joonisel 5.17(a) ja jaotusfunktsiooni graafik joonisel 5.17(b).

Juhuslik suurus allub **normaaljaotusele**, kui see
- on mõjutatud paljude faktorite poolt;
- iga üksikfaktori mõju on väike;
- puudub domineeriv faktor.

Normaaljaotuse esinemine



Joonis 5.17. Normaaljaotuse jaotustiheduse (a) ja jaotusfunktsiooni graafik (b). Jaotustiheduse graafikut (a) nimetatakse ka Gaussi kõveraks. Kõvera laius käänupunktide juures on kahekordne standardhälve (vt lisa A.4)

Kui mingi juhuslik suurus allub normaaljaotusele, siis öeldakse, et see suurus on **normaalselt jaotunud**. Normaalselt jaotunud on tihti näiteks tootmisliinilt tulnud toodete kaal ja mõõtmed, mingi tegevuse jaoks kulunud aeg, toodangu maht, poe külastajate arv päevas, väärtpaberi tulusus, samuti inimeste füüsilised parameetrid ja vaimsed võimed (vt näiteks jooniseid 5.18 ja 5.19).

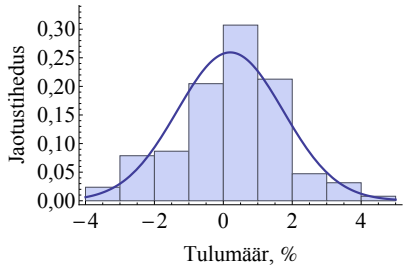
Normaaljaotuse **jaotustihedus** väärtusel x:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{5.74}$$

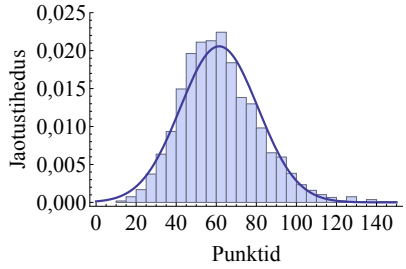
kus μ on jaotuse keskväärtus ja σ standardhälve.

Normaaljaotuse jaotustihedus

Keskväärtus määrab normaaljaotuskõvera asukoha, standardhälbest sõltub, kui lai on jaotuskõver (joonised 5.20 ja 5.21). Kuna jaotuskõvera alla jääv pindala on alati üks (tõenäosus, et juhusliku suuruse



Joonis 5.18. Xeroxi aktsia tulumäär päevas 8. veebr kuni 9. august 2013, vaatluste arv 127. Aritmeetiline keskmine on 0,2% ja standardhälve 1,5%. Joon on vastava normaaljaotuse jaotustihedus

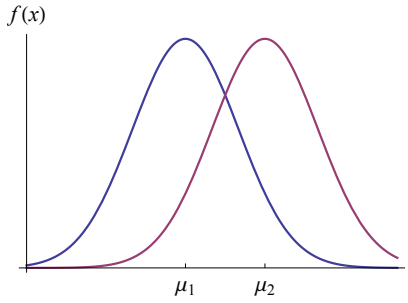


Joonis 5.19. Koolinoorte matemaatikavõistluse „Känguru 2002“ juunioride tulemuste jaotus. Osavõtjaid 2141. Aritmeetiline keskmine on 61,4 ja standardhälve 19,4. Lisatud on vastav normaaljaotuskõver

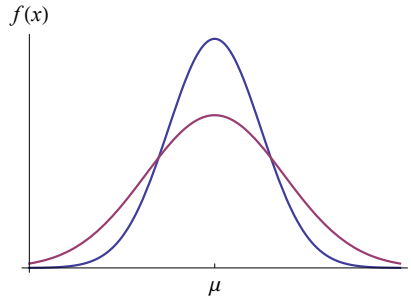
väärtus jääb vahemikku $(-\infty, \infty)$, on üks), siis mida hajuvam on antud juhuslik suurus, seda madalam on jaotuskõver. Kui leida keskväärtusele μ vastav jaotustiheduse väärtus, siis valemis (5.74) $x = \mu$ ja arvu e astmenäitaja on 0. Siis

$$f(\mu) = \frac{1}{\sigma\sqrt{2\pi}}. \tag{5.75}$$

Jaotustiheduse väärtus maksimumpunktis on määratud ainult standardhälbega. On näha, et mida suurem on standardhälve σ, seda väiksem on $f(\mu)$ ja järelikult seda madalamal on jaotustiheduse graafiku tipp.



Joonis 5.20. Kaks erineva keskväärtusega normaaljaotuskõverat



Joonis 5.21. Kaks erineva standardhälbega normaaljaotuskõverat

Jooniselt 5.17 (a) on näha, et normaaljaotuse keskväärtuse kohal on jaotustiheduse graafik kõige kõrgem. Järelikult seal asub ka normaaljaotuse mood. Samuti on näha, et normaaljaotuskõver on sümmeetriline ning keskväärtus μ jagab kõveraaluse piirkonna võrdsete pindala-

dega osadeks. Järelikult asub mediaan samas kohas. Seda on võimalik näidata ka matemaatiliselt, lähtudes jaotustiheduse valemist (5.74) (vt lisa A.4).

Normaaljaotuse korral:

$$\text{aritmeetiline keskmine} = \text{mood} = \text{mediaan}.$$

Asümmeetria kordaja

$$A = 0 \tag{5.76}$$

ja püstakuse kordaja ehk ekstsess

$$E = 0. \tag{5.77}$$

Mood, mediaan, asümmeetria ja püstakus normaaljaotuse korral

Sageli kasutatav Pearsoni püstakusekordaja on normaaljaotuse korral 3.

Erinevate tõenäosuste leidmiseks tuleb kasutada jaotusfunktsiooni, mis vastavalt valemile (5.29) on integraal jaotustihedusest:

$$F(a) = \int_{-\infty}^{a} f(x)dx.$$

Normaaljaotuse **jaotusfunktsioon** väärtusel a:

$$F(a) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx, \tag{5.78}$$

kus μ on jaotuse keskväärtus ja σ standardhälve.

Normaaljaotuse jaotusfunktsioon

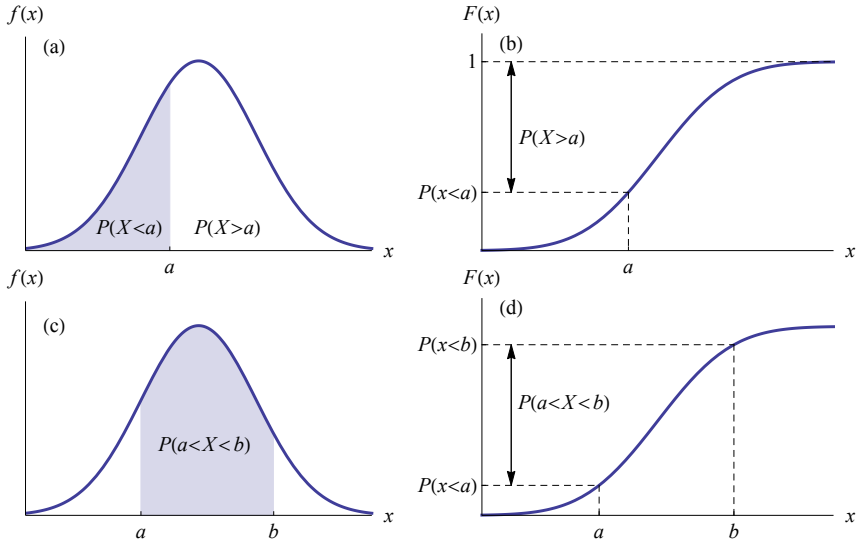
Valemis (5.78) olevat integraali ei ole võimalik avaldada elementaarfunktsioonide kaudu, sest integraal $\int e^{-x^2} dx$ on mitteelementaarne integraal. Seepärast ei saa normaaljaotusele alluva suuruse korral kasutada tõenäosuste arvutamiseks valemit ja kalkulaatorit. Selle integraali arvutamiseks kasutatakse ligikaudseid numbrilisi meetodeid, mida rakendab vastav tarkvara.

Nii nagu iga pideva jaotuse korral, on jaotustiheduse graafikul tõenäosus võrdne graafiku alla jääva vastava piirkonna pindalaga (vt joonis 5.22 (a) ja (c)).

Normaaljaotuse jaotustiheduse või jaotusfunktsiooni leidmiseks kasutatakse tabelarvutuses funktsiooni **NORM.DIST**. Jaotustiheduse väärtuste leidmisel on parameeter *Cumulative* 0, jaotusfunktsiooni väärtuste leidmisel *Cumulative* 1. Kui normaaljaotuse keskväärtus

on μ ja standardhälve σ, siis tabelarvutuses saab tõenäosusi leida järgmiselt:

$$\begin{aligned} P(X<a) &= F(a) && = \text{NORM.DIST}(a;\mu;\sigma;1),\\ P(X>a) &= 1-F(a) && = 1-\text{NORM.DIST}(a;\mu;\sigma;1),\\ P(a<X<b) &= F(b)-F(a) && = \text{NORM.DIST}(b;\mu;\sigma;1)-\\ & && -\text{NORM.DIST}(a;\mu;\sigma;1). \end{aligned}$$



Joonis 5.22. Jaotustiheduse graafikul (a) on tõenäosus $P(X<a)$ võrdne varjutatud piirkonna pindalaga ning tõenäosus $P(X>c)$ võrdne kõvera all oleva varjutamata piirkonna pindalaga. Jaotusfunktsiooni $F(x)$ graafikul (b) saab tõenäosuse $F(a)=P(X<a)$ määrata jaotusfunktsiooni teljelt. Tõenäosus $P(X>a)=1-P(X<a)$. Tõenäosus $P(a<X<b)$ on jaotustiheduse graafikul (c) varjutatud piirkonna pindala ning jaotusfunktsiooni graafikul (d) kahe tõenäosuse vahe $P(X<b)-P(X<a)$

Näide 5.23. Kullerfirma saadetiste arvu tõenäosuste leidmine

N05Jaotused
N5.23

Kasutades näites 5.22 toodud andmeid kullerfirma väikesaadetiste kohta ja eeldades, et saadetiste arv kuus allub normaaljaotusele, leiame, kui suure tõenäosusega on kuus

a) alla 1500 saadetise;
b) üle 1500 saadetise;
c) saadetiste arv 1500 ja 2000 vahel.

Selleks kasutame normaaljaotust, mille keskväärtus ja standardhälve on leitud empiiriliste andmete põhjal.

a) Ühes kuus kättetoimetatud väikesaadetiste arvu keskväärtus oli $\mu = 1739{,}4$ ja standardhälve $\sigma = 220{,}5$ saadetist kuus.

Otsitava tõenäosuse annab meile jaotusfunktsioon $F(1500)$. Selle väärtuse leidmiseks kasutame tabelarvutuses funktsiooni NORM.DIST(1500;1739,4;220,5;1):

$$P(X < 1500) = F(1500) \approx 0{,}139.$$

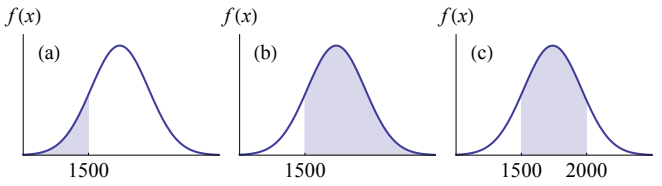
Vastus: tõenäosus, et saadetiste arv kuus on alla 1500, on 0,139 (joonisel 5.23 (a) varjutatud piirkonna pindala).
b) Tõenäosus, et kuus on üle 1500 saadetise (joonis 5.23 (b)):

$$P(X > 1500) = 1 - F(1500) = 1 - 0{,}139 = 0{,}861.$$

Vastus: tõenäosus, et saadetiste arv kuus on suurem kui 1500, on 0,861 (joonisel 5.23 (b) varjutatud piirkonna pindala).
c) Et leida, kui suure tõenäosusega on saadetiste arv 1500 ja 2000 vahel, leiame jaotusfunktsioonide $F(1500)$ ja $F(2000)$ vahe (joonis 5.23 (c)):

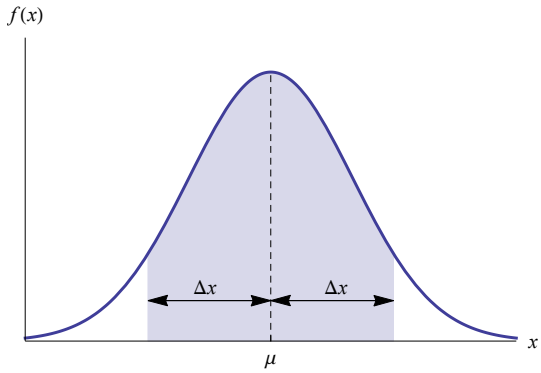
$$P(1500 < X < 2000) = F(2000) - F(1500) = 0{,}881 - 0{,}139 = 0{,}742.$$

Vastus: tõenäosus, et saadetiste arv kuus on 1500 ja 2000 vahel, on 0,742 (joonisel 5.23 (c) varjutatud piirkonna pindala).



Joonis 5.23. (a) Varjutatud ala pindala on tõenäosus, et saadetiste arv on alla 1500. (b) Varjutatud ala pindala on tõenäosus, et saadetiste arv on üle 1500. (c) Varjutatud ala pindala on tõenäosus, et saadetiste arv on 1500 ja 2000 vahel

Sageli on vaja leida tõenäosust, et normaaljaotusele alluva suuruse väärtus jääb teatud vahemikku, mille keskel on keskväärtus (vt joonis 5.24). Seda vahemikku tähistatakse tavaliselt $\mu \pm \Delta x$, kus Δx on absoluutne kõrvalekalle. Vahemiku alumine piir on siis $\mu - \Delta x$ ning ülemine piir $\mu + \Delta x$. Mõnikord antakse aga ette suhteline kõrvalekalle, mis on teatud protsent keskväärtusest. Sellisel juhul tuleb eelnevalt leida absoluutne kõrvalekalle Δx, seejärel vahemiku alumine ja ülemine piir ning siis saab jaotusfunktsiooni väärtuste vahe abil leida vahemikku langemise tõenäosuse.



Joonis 5.24. Varjutatud piirkonna pindala on tõenäosus, et juhusliku suuruse väärtus jääb vahemikku $\mu \pm \Delta x$

Näide 5.24. Kõrgemasse kvaliteediklassi kuuluvate toodete osakaal

Toote A läbimõõt peab olema 5 mm ja toode kuulub kõrgemasse klassi, kui selle tegelikud mõõtmed ei erine nominaalmõõtmest rohkem kui 10%. Toodete mõõtmed alluvad normaaljaotusele ja on kindlaks tehtud, et standardhälve on 0,8 mm. Kui suur on oodatav kõrgemasse klassi kuuluvate toodete arv päevas, kui päevas toodetakse toodet A 40 tükki?
Suhtelisele kõrvalekaldele 10% vastab nominaalmõõtmete 5 mm korral absoluutne kõrvalekalle 0,5 mm. Leiame, kui suure tõenäosusega jääb läbimõõt vahemikku (4,5 mm,5,5 mm). Selleks kasutame vastava normaaljaotuse jaotusfunktsiooni väätuste vahet, mille leiame tabelarvutusfunktsiooniga NORM.DIST:

$$F(5{,}5) - F(4{,}5) = 0{,}734 - 0{,}266 = 0{,}468.$$

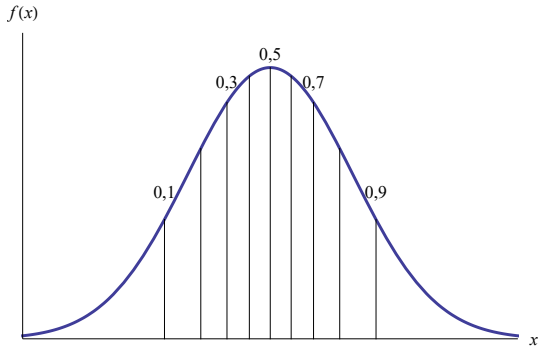
Järelikult kuulub keskmiselt 46,8% toodetest kõrgemasse klassi. Kui toodetakse 40 toodet, siis

$$40 \cdot 0{,}468 \approx 18{,}7.$$

Vastus: 40 tootest kuulub kõrgemasse kvaliteediklassi ligikaudu 19 toodet.

Mõnikord on vaja lahendada tõenäosuse leidmise pöördülesanne: antakse ette tõenäosus p ja leida tuleb juhusliku suuruse X selline väärtus a, et

$$P(X < a) = p.$$

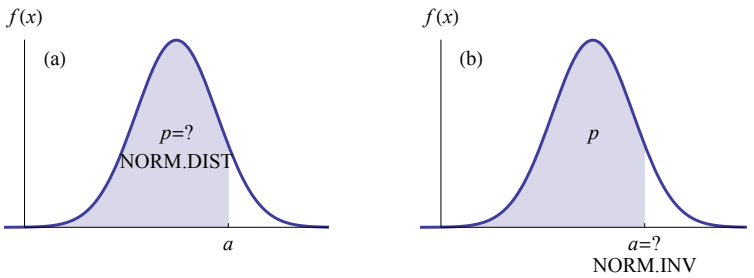


Joonis 5.25. Normaaljaotuse detsiilid. Detsiili tähistava joone kohal on märgitud detsiilile vastav tõenäosus p

Kui näiteks $p = 0{,}1$ ja $P(X < a) = 0{,}1$, siis a on esimene detsiil. Kui $P(X < a) = 0{,}2$, siis a on teine detsiil jne. Joonis 5.25 näitab normaaljaotuse detsiilide paiknemist. Horisontaaltelje juures on märgitud detsiilile vastav tõenäosus p. Igasse kõveraalusesse joontega eraldatud piirkonda jääb 10% kõikidest väärtustest. See tähendab, et kõikide piirkondade pindalad on 0,1.

Sellist normaaljaotusele alluva juhusliku suuruse väärtust x, mille korral jaotusfunktsioon $F(x) = p$, kus p on antud, võimaldab tabelarvutuses leida funktsioon **NORM.INV**. Tähistades jaotusfunktsiooni pöördfunktsiooni F^{-1}, võime tabelarvutuse jaoks kirjutada (vt ka joonis 5.26):

$$\begin{aligned} p = F(x) &= \text{NORM.DIST}(x; \mu; \sigma; 1), \\ x = F^{-1}(p) &= \text{NORM.INV}(p; \mu; \sigma). \end{aligned}$$



Joonis 5.26. Tabelarvutuse funktsioonide kasutamine normaaljaotuse korral. NORM.DIST leiab väärtusele a vastava tõenäosuse p. NORM.INV leiab väärtuse a, kui sellele vastav tõenäosus p on antud

Näiteks kui normaaljaotuse keskväärtus $\mu = 100$ ja standardhälve $\sigma = 50$, siis arvule 30 vastav jaotusfunktsiooni väärtus

$$F(30) = \text{NORM.DIST}(30; 100; 50; 1) \approx 0{,}0808.$$

Kui me tahame aga leida, millise arvu x korral on jaotusfunktsiooni väärtus 0,0808, siis

$$x = \text{NORM.INV}(0{,}0808; 100; 50) = 30.$$

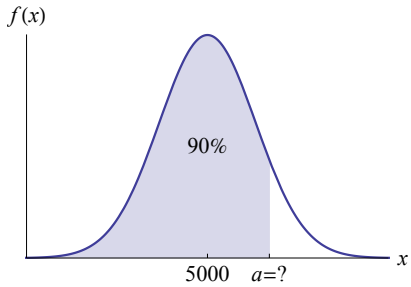
Näide 5.25. Piima tellimuskoguse leidmine

N05Jaotused
N5.25

Toidupoes müüakse päevas keskmiselt 5000 liitrit piima ja päevaste müügikoguste standardhälve on 400 liitrit. Eeldame, et piima müük allub normaaljaotusele. Kui palju tuleks päevas piima tellida, et piima jätkuks päeva lõpuni tõenäosusega 90%?
Meil tuleb leida selline väärtus a, et $P(X < a) = 0{,}9$, s.t üheksas detsiil. Selle leidmiseks kasutame tabelarvutuses funktsiooni NORM.INV:

$$\text{NORM.INV}(0{,}9; 5000; 400) \approx 5512{,}6.$$

Vastus: piima tuleks tellida ligikaudu 5513 liitrit.



Näites 5.25 ette antud tõenäosust, et kaupa jätkuks, nimetatakse **teenindustasemeks**. Vastandsündmus on kaubavaegus. Kui teenindustase on 90%, siis kaubavaeguse tekkimise tõenäosus on 10%. Praktikas võetakse teenindustase tihti ette ning seejärel leitakse sobiv tellimuse suurus. Kasutatava normaaljaotuse keskväärtus ja standardhälve leitakse eelmiste perioodide vaatlusandmete põhjal.

Näide 5.26. Varude juhtimine ja normaaljaotusele alluv nõudlus

N05Jaotused
N5.26

Jalatsikaupluse Kuldne King juhatajal tuleb ette valmistada tellimused järgmiseks kevad-suviseks hooajaks. Tellida tuleb parajasti nii palju kaupa, et jõuaks suve lõpuks maha müüa. Augustis on vaja laopind vabastada sügisjalatsite jaoks ja järelejäänud suvejalatsid tuleb müüa siis madalate hindadega. Mudeli A hind hooajal on 50 eurot. Augustis saab aga järelejäänud kingadest

lahti hinnaga 25 eurot, kusjuures mudeli omahind on kaupluse jaoks 35 eurot.
Eelmiste perioodide müügimahtude statistiline analüüs näitas, et keskmiselt müüakse seda mudelit hooaja jooksul 400 paari, kusjuures nõudlus allub normaaljaotusele ja standardhälve on 142 paari. Tuleb leida optimaalne tellimuskogus.
Juhataja arutleb järgmiselt. Iga hooaja jooksul müüdud lisa kingapaar annab lisatulu: $MR = 50 - 35 = 15$ eurot. Iga hooaja jooksul müümata jäänud kingapaar tekitab lisakulu: $MC = 35 - 25 = 10$ eurot. Kuna eeldame, et nõudlus allub normaaljaotusele, siis keskväärtus 400 paari on samal ajal ka mediaan. Kui tellimust suurendada 400 paarilt 401 paarini, siis tõenäosusega 0,5 võib nõudlus D olla suurem kui 400, $P(D > 400 = 0{,}5)$, ja võib saada lisatulu ning tõenäosusega 0,5 võib nõudlus olla väiksem kui 400 ning tekib lisakulu. Seega on lisatulu oodatav väärtus

$$E(MR)_{|400} = MR \cdot P(D > 400) = 15 \cdot 0{,}5 = 7{,}5$$

ja lisakulu oodatav väärtus

$$E(MC)_{|400} = MC \cdot P(D < 400) = 10 \cdot 0{,}5 = 5.$$

Kuna oodatav lisatulu on suurem kui oodatav lisakulu, on mõistlik tellimuskogust suurendada. Kogust tasub suurendada sellise väärtuseni Q, et oodatav lisakulu saaks võrdseks oodatava lisatuluga:

$$E(MR)_{|Q} = E(MC)_{|Q}$$
$$MR \cdot P(D > Q) = MC \cdot P(D \leq Q).$$

Arvestame, et $P(D > Q) = 1 - P(D \leq Q)$, ja avaldame tõenäosuse $P(D \leq Q)$:

$$MR \cdot (1 - P(D \leq Q)) = MC \cdot P(D \leq Q)$$
$$MR - MR \cdot P(D \leq Q)) = MC \cdot P(D \leq Q).$$

Viimasest võrdusest saame, et kauba piisavuse tõenäosus (nõudlus D on väiksem kui tellimuskogus Q):

$$P(D \leq Q) = \frac{MR}{MR + MC}. \tag{5.79}$$

Pannes valemisse (5.79) lisatulu ja lisakulu väärtused, saame leida tõenäosuse, mis näitab teenindustaset:

$$P(D \leq Q) = \frac{15}{15 + 10} = 0{,}6.$$

Teades tõenäosust, leiame optimaalse tellimuskoguse Q. Selleks kasutame tabelarvutuses funktsiooni NORM.INV:

$$Q = \text{NORM.INV}(0{,}6; 400; 142) = 436.$$

Vastus: optimaalne tellimuskogus on 436 paari.

Näites 5.26 toodud situatsiooni tuntakse ka ajalehepoisi probleemina (*newsboy problem*). Ajalehepoiss peab samuti hommikul otsustama, kui palju ajalehti ta sellel päeval müümiseks ostab. Lisaks sellisele üheks hooajaks tellimise probleemile kasutatakse normaaljaotust ka muude kaubavarude juhtimisega seotud probleemide korral.

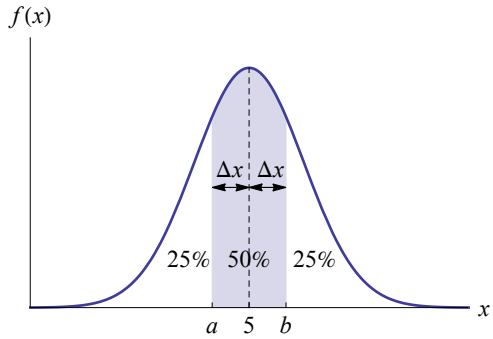
Näites 5.24 leidsime, kui suure tõenäosusega jääb normaaljaotusele alluva juhusliku suuruse väärtus vahemikku $\mu \pm \Delta x$. Sageli on vaja lahendada vastupidine probleem: on ette antud vahemikku $\mu \pm \Delta x$ langemise tõenäosus ning tuleb leida vahemiku poollaius Δx.

Näide 5.27. Lubatud kõrvalekalle nominaalsest

N05Jaotused
N5.27

Näites 5.24 oli toote nominaalne läbimõõt 5 mm ja standardhälve 0,8 mm. Mitu millimeetrit võib olla kõrvalekalle nominaalsest, kui tahame, et kõrgemasse kvaliteediklassi kuuluks 50% toodetest?
Otsitav vahemik on sümmeetriline keskväärtuse suhtes. Seega on meil vaja leida selline suurus Δx, et

$$P(5 - \Delta x < X < 5 + \Delta x) = 0{,}5.$$



Joonis 5.27. Kui vahemikku $5 \pm \Delta x$ jääb 50% kõikidest väärtustest, siis $P(x < a) = 0{,}25$ ja $P(x < b) = 0{,}75$, kus $a = 5 - \Delta x$ ja $b = 5 + \Delta x$

Jooniselt 5.27 on näha, et

$$P(X < 5 - \Delta x) = \frac{1 - 0{,}5}{2} = 0{,}25,$$
$$P(X < 5 + \Delta x) = 0{,}25 + 0{,}5 = 0{,}75.$$

Tähistame alumist piiri tähega $a = 5 - \Delta x$ ja ülemist piiri tähega $b = 5 + \Delta x$. Meil tuleb leida sellised väärtused a ja b, et

$$F(a) = 0{,}25 \text{ ja } F(b) = 0{,}75.$$

Tabelarvutuse vastava funktsiooni kasutamine annab järgmised tulemused:

$$F^{-1}(0{,}25) = \text{NORM.INV}(0{,}25; 5; 0{,}8) \approx 4{,}46,$$
$$F^{-1}(0{,}75) = \text{NORM.INV}(0{,}75; 5; 0{,}8) \approx 5{,}54.$$

Järelikult kõrvalekalle keskmisest võib olla

$$\Delta x = \frac{b - a}{2} = \frac{5{,}54 - 4{,}46}{2} = 0{,}54.$$

Vastus: et kõrgemasse kvaliteediklassi kuuluks 50% toodetest, tohib lubatud kõrvalekaldumine nominaalsest olla $\pm 0{,}54$ mm.

Näites 5.27 kasutatud lahendusskeemi võib kirja panna järgmiselt.

Kui on vaja leida normaaljaotuse $N(\mu, \sigma)$ keskväärtuse μ suhtes sümmeetriline vahemik $(\mu - \Delta x, \mu + \Delta x)$, millesse langemise tõenäosus oleks β, tuleb algul leida sellised väärtused a ja b, et vastavad jaotusfunktsioonid oleksid

$$F(a) = \frac{1 - \beta}{2} \text{ ja } F(b) = \frac{1 + \beta}{2}.$$

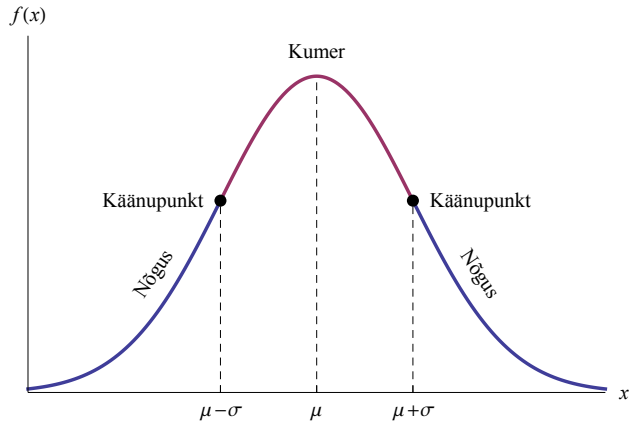
Vahemiku leidmine tõenäosuse järgi

Sellisel juhul on otsitava vahemiku poollaius

$$\Delta x = \frac{b - a}{2}.$$

Normaaljaotuskõvera alumine osa on nõgus ja ülemine osa kumer. Nõgusus läheb kumeruseks üle käänupunktides, mis on keskväärtusest standardhälbe kaugusel punktides $\mu - \sigma$ ja $\mu + \sigma$ (joonis 5.28). Selle tõestus on toodud lisas A.4. Järelikult kuni standardhälbe kauguse-

ni keskväärtusest kahaneb jaotustihedus kiirenevalt (kumer) ja edasi toimub aeglustuv kahanemine (nõgus).



Joonis 5.28. Normaaljaotuskõvera käänupunktid on standardhälbe kaugusel keskväärtusest

Suvalise juhusliku suuruse korral kehtis Tšebõšovi teoreem, mille alusel sai leida, kui suure tõenäosusega jääb juhusliku suuruse väärtus keskväärtusest kaugemale kui k standardhälvet (alaptk 3.5). Näiteks aritmeetilisest keskmisest kaugemal kui kaks standardhälvet asub kõige enam $1/4 = 25\%$ selle väärtustest. Kui aga juhuslik suurus allub normaaljaotusele, on vastav tõenäosus väiksem, ainult 4,8%.

Ükskõik, milliste väärtustega on normaaljaotuse parameetrid keskväärtus μ ja standardhälve σ, on standardhälbe kordsega määratud vahemikku jäämise tõenäosus ühesugune:

$$P(\mu - \sigma < X < \mu + \sigma) = F(\mu + \sigma) - F(\mu - \sigma) \approx 0{,}683, \quad (5.80)$$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = F(\mu + 2\sigma) - F(\mu - 2\sigma) \approx 0{,}954, \quad (5.81)$$

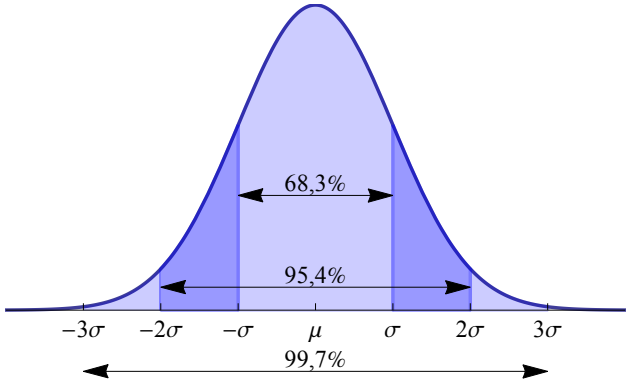
$$P(\mu - 3\sigma < X < \mu + 3\sigma) = F(\mu + 3\sigma) - F(\mu - 3\sigma) \approx 0{,}997, \quad (5.82)$$

kus $F(x)$ on normaaljaotuse jaotusfunktsioon (joonis 5.29).

σ-kordsed vahemikud

Tabel 5.3. Standardhälbega määratud vahemikku ja vahemikust välja jäämise tõenäosused normaaljaotuse korral. Protsendid on ligikaudsed

Vahemik	Vahemiku sees	Vahemikust väljas
$(\mu - \sigma,\ \mu + \sigma)$	68,3%	31,7%
$(\mu - 2\sigma,\ \mu + 2\sigma)$	95,4%	4,6%
$(\mu - 3\sigma,\ \mu + 3\sigma)$	99,7%	0,3%



Joonis 5.29. Kui suur osa kõikidest väärtustest jääb normaaljaotuse korral standardhälbe kordsega määratud vahemikesse

Näide 5.28. Aktsiahinna Bollingeri koridor

2014. aasta Pärnu Finantskonverentsil esines USA-st pärit aktsiahindade tehnilise analüüsi spetsialist John Bollinger, kes tutvustas oma 1980-ndatel loodud meetodit aktsiahindade muutumise jälgimiseks ja investeerimisotsuste tegemiseks (Abišala, 2015). Bollingeri koridori (*Bollinger Bands*) piirid on aktsiahinna libisevast keskmisest kummalegi poole kahe standardhälbe kaugusel. Kahekordse standardhälbe kasutamine garanteerib, et 95%-lise tõenäosusega jääb hinnaliikumine koridori $\mu \pm 2\sigma$ piiridesse, kus μ on tavaliselt viimase 20 perioodi keskmine. Hind puudutab koridori serva ainult väga tugevate kõikumiste ajal. Kui hind puudutab koridori alumist serva $\mu - 2\sigma$, siis on hind ekstreemselt madal ning varsti on tõenäoliselt oodata hinnatõusu. Kui hind puudutab koridori ülemist serva $\mu + 2\sigma$, on varsti oodata hinna alanemist. (Bollinger, 1992)

Erinevate kaupade tootmisel ei ole võimalik garanteerida, et kõikide eksemplaride parameetrid vastaksid etteantud nominaalväärtusele kuitahes täpselt. Tootmisprotsessi iseärasustest lähtudes esineb alati juhuslik kõikumine. Seepärast antakse toote tehnilises kirjelduses tihti parameetrite väärtuste lubatud piirmäärad. Kui tootja soovib, et 99,7% kõikidest eksemplaridest jääks antud piirmääradesse, siis annab ta kõikumiseks $\pm 3\sigma$, kus σ on selle parameetri standardhälve (vt joonis 5.29). Standardhälbe määramiseks mõõdab tootja tootmisliinilt juhuslikult võetud tooteid.

Näide 5.29. Patareide mõõtmete varieeruvus

Duracelli AAA patarei QU2400 tehnilises kirjelduses[a] on märgitud, et patarei pikkus jääb vahemikku 43,5 kuni 44,5 mm. Kui suur võib olla patareide pikkuse standardhälve, kui Duracell soovib garanteerida, et 99,7% kõikidest toodetud patareidest jääks sellesse vahemikku? Kui suur on pikkuse variatsioonikordaja?
Lubatud pikkusevahemiku võib anda kujul $44{,}0 \pm 0{,}5$ mm, kus 0,5 mm on vahemiku poollaius. Tõenäosusele 99,7% vastab poollaius 3σ (tabel 5.3). Järelikult $3\sigma = 0{,}5$ ning $\sigma = 0{,}5/3 \approx 0{,}167$ mm. Variatsioonikordaja on $0{,}167/44 \approx 0{,}38\%$.

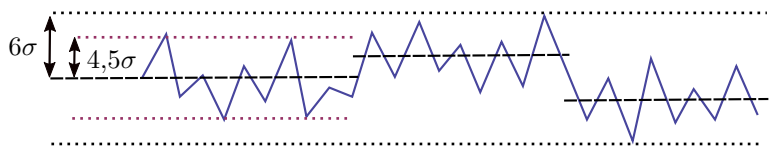
[a] Duracelli veebileht `http://ww2.duracell.com` *Technical Library*

- Kui tootmisprotsess on selline, et toote parameetrid varieeruvad palju, on standardhälve suur ja ka lubatud vahemik suur. Sellisel juhul ei ole toode konkurentsivõimeline.
- Kui tehnilises kirjelduses anda lubatud vahemik väiksem kui $\pm 3\sigma$, siis on tõenäosus, et mõni toode jääb vahemikust välja, suur. Näiteks $\pm 2\sigma$ korral 4,6% ehk ligikaudu viis toodet sajast. See aga tähendab võimalikke pretensioone.
- Et garanteerida väike kõikumine ja samal ajal lubatud piirmäärade suur usaldusväärsus, on ainukeseks võimaluseks vähendada tootmisprotsessis esinevat varieerumist (standardhälvet).

Näide 5.30. Kvaliteedijuhtimise Kuus Sigmat

1980-ndatel hakkasid USA autotootjad ja samuti telekommunikatsiooniseadmete tootjad tugevalt alla jääma Jaapani konkurentidele ja kandsid suurt majanduslikku kahju. Ameerika ostjad tajusid, et jaapanlaste valmistatud tooted olid parema kvaliteediga. USA autotootja Fordi läbiviidud defektimäärade uurimus näitas, et kui nende toodetud autode defektimäär kõikus 2600–6000 tk/mln, siis jaapanlaste autodel oli see 4 tk/mln. USA firma Motorola toodetud telefonidel oli defektimäär ligikaudu 2600 tk/mln, jaapanlaste toodetud seadmetel 3 tk/mln. 1986. aastal käivitas Motorola programmi, et viia kõigi oma toodete ja teenuste vead tasemele 3 tk/mln. Selle programmi nimetuseks oli Kuus Sigmat (*Six Sigma*). 1988. aastal sai Motorola selle eest USA kvaliteediauhinna *Malcolm Baldrige National Quality Award*.
Ettevõte, mille toodete defektimäär on väiksem kui 3 tk/mln kohta (täpsemalt 3,4 tk/mln), on kvaliteedijuhtimises saavuta-

nud taseme Kuus Sigmat. Kust tuleb termin „Kuus Sigmat“ ja miks just 3,4 tk/mln?
Arvutused näitavad, et kui toote parameetrid alluvad normaaljaotusele, siis ühele poole keskväärtusest kaugemale kui 6σ jäävate toodete osakaal on ligikaudu 10^{-9}, mis teeb üks toode miljardi kohta (defektiks loetakse seda, kui eemaldumine keskväärtusest on ühele poole, n-ö halvemas suunas). 3,4 tk/mln kohta vastab hoopis kaugusele $4{,}5\sigma$ (vt ülesanne 5.59). Kvaliteedijuhtimisel kasutatava statistilise protsessiohje korral eristatakse toodete parameetrite pikaajalist ja lühiajalist varieerumist. Lühiajaline varieerumine on näiteks päeva toodangu parameetrite varieerumine ja pikaajaline varieerumine kuu toodangu parameetrite varieerumine. Päeva jooksul jääb keskmine samaks, kuid kuu aja jooksul esineb ka mõningane keskmise kõikumine (vt joonis 5.30). On tähele pandud, et kui lühiajaline varieerumine jääb piiridesse $4{,}5\sigma$, siis pikaajaline varieerumine on piirides 6σ.



Joonis 5.30. Toodete parameetrite lühiajaline (kolm perioodi) ja pikaajaline varieerumine

Seega, Kuue Sigmaga määratud 3,4 defekti miljoni kohta vastab tegelikult lühiajalisel varieerumisel määratud kaugusele $4{,}5\sigma$. (Taghizadegan, 2006)

Tabel 5.4. Defektimäär erinevate sigma tasemete korral kvaliteedijuhtimises

Sigma tase	Lühiajaline varieerumine	Tõenäosus, et vastab normile	Tõenäosus, et ei vasta normile	Defektimäär tk/mln
3	$1{,}5\sigma$	93,3%	6,7%	66 807
4	$2{,}5\sigma$	99,38%	0,62%	6 210
5	$3{,}5\sigma$	99,977%	0,023%	233
6	$4{,}5\sigma$	99,99966%	0,00034%	3,4

Ka teenindusettevõtete kvaliteeti saab selle järgi hinnata. Kui näiteks McDonalds teenindab päevas miljon klienti ja ainult kolm neist ei ole teenindusega rahul, siis on McDonaldsi teeninduskultuur Kuue Sigma tasemel. Tänapäeval on Kuus Sigmat meetodite kogum ettevõtte või organisatsiooni protsesside

pidevaks parandamiseks. Probleemide kindlaksmääramisel kasutatakse statistilist analüüsi. Seda võib vaadelda ka kui juhtimisstrateegiat, mille eesmärgiks on ettevõttes maailmaklassi kvaliteedi saavutamine.

Alapeatükis 3.6 tutvusime standardiseeritud skaala mõistega. Kui normaaljaotusele alluv juhuslik suurus teisendada standardiseeritud skaalale, saame standardiseeritud normaaljaotuse.

Standardiseeritud normaaljaotus

Kui juhuslik suurus X allub normaaljaotusele keskväärtusega μ ja standardhälbega σ, siis selle standardiseeritud väärtused

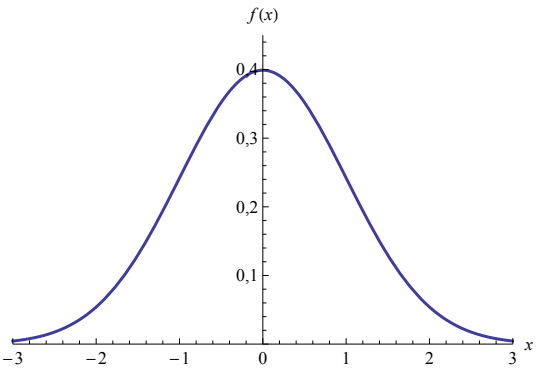
$$z = \frac{x - \mu}{\sigma} \tag{5.83}$$

alluvad **standardiseeritud normaaljaotusele** keskväärtusega 0 ja standardhälbega 1. Standardiseeritud normaaljaotuse tähis on N(0, 1).

Valemi (5.74) põhjal on standardiseeritud normaaljaotuse jaotustihedus:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \tag{5.84}$$

Vastav graafik on toodud joonisel 5.31.



Joonis 5.31. Standardiseeritud normaaljaotuse jaotuskõver

Tabelarvutuses on standardiseeritud normaaljaotuse tihedus- ja jaotusfunktsiooni leidmiseks funktsioon **NORM.S.DIST**.

Standardiseeritud normaaljaotust on mugav kasutada erandlike väärtuste määramisel. Kui erandlikeks väärtusteks lugeda need, mille

esinemise tõenäosus on väiksem kui 0,3%, siis vastavalt tabelile 5.3 on need väärtused keskväärtusest kaugemal kui 3σ:

$$|x-\mu| > 3\sigma$$
$$\frac{|x-\mu|}{\sigma} > 3.$$

Arvestades, et standardhälve σ on alati positiivne, $\sigma = |\sigma|$, võime kirjutada

$$\left|\frac{x-\mu}{\sigma}\right| > 3.$$

Võrreldes viimast võrratust valemiga (5.83), võime kirjutada tingimuse erandlike väärtuste määramiseks.

Normaaljaotusele alluva suuruse X **erandlike väärtuste** määramiseks kasutatakse tingimust

$$|z| > 3, \tag{5.85}$$

kus standardiseeritud väärtus z on leitud valemist (5.83).

Erandlike väärtuste tingimus

Näide 5.31. Erandlikud tulemused matemaatikaolümpiaadil

Joonisel 5.19 oli toodud koolinoorte matemaatikavõistluse „Känguru 2002“ juunioride tulemuste jaotus, mille keskväärtus $\mu = 61{,}4$ ja standardhälve $\sigma = 19{,}4$. Leida erandlikud väärtused eeldusel, et tulemuste jaotus allub normaaljaotusele.
Tabelarvutuses on valemi (5.83) põhjal leitud iga osavõtja tulemuse z skoor. Erandlikud tulemused olid esimesel 17 osavõtjal (120 punkti ja rohkem ($z > 3$)).

N05Jaotused N5.31

5.11. Ülesanded

5.1. Juhusliku suuruse X väärtuste hulk on $\{10, 11, 12, 13, 14\}$. Jaotusseadus on antud tabeli kujul: *Jaotusseadus*

x_i	10	11	12	13	14
p_i	0,1	0,2	0,4	0,2	0,1

1. Kontrollida jaotusseaduse normeerimistingimust.

2. Kui suur on tõenäosus, et juhuslik suurus X omandab väärtuse 10 või 11?
3. Kui suur on tõenäosus, et juhuslik suurus X omandab väärtuse, mis on suurem kui 12?
4. Kui suur on tõenäosus, et juhuslik suurus X omandab väärtuse, mis on suurem kui 14?

VASTUS lk 664.

5.2. Haiglas jälgiti operatsiooniruumide kasutamist 20 päeva jooksul. Kolmel päeval kasutati vaid 1 operatsiooniruumi, viiel päeval 2 ruumi, kaheksal päeval 3 ruumi ja vaid neljal päeval kasutati kõiki 4 ruumi. Olgu X mingil suvalisel päeval kasutuses olevate operatsiooniruumide arv.

1. Kirjutada välja juhusliku suuruse X jaotusseadus.
2. Konstrueerida vastav jaotuspolügoon.
3. Näidata, et jaotusseadus rahuldab normeerimistingimust.

VASTUS lk 664.

5.3. Leibkonna eelarve uuringus osalenud 9080 peres oli leibkonna liikmete arvu jaotus järgmine (*Leibkonna eelarve uuring* 2012):

Liikmete arv	Leibkondade arv
1	886
2	2 466
3	1 917
4	2 060
5	1 751

Olgu X juhuslik suurus, mis kirjeldab leibkonna liikmete arvu juhuslikult väljavalitud leibkonnas.

1. Panna kirja juhusliku suuruse X jaotusseadus.
2. Kui suur on tõenäosus, et juhuslikult väljavalitud leibkonnas on kolm liiget?
3. Kui suur on tõenäosus, et juhuslikult väljavalitud leibkonnas on rohkem kui kolm liiget?

VASTUS lk 664.

Jaotusfunktsioon

5.4. 2014. aasta mais koostas Eesti Töötukassa analüüsiosakond „Ettevõtluse alustamise toetuse mõjuanalüüsi“. Analüüsi eesmärgiks oli muu hulgas välja selgitada, kas töötukassa makstav ettevõtluse alustamise toetus on mõjus meede töötute tööle aitamiseks. Vaatluse all olid 2009. aasta 1. maist kuni 2011. aasta lõpuni avalduse esitanud ja hiljemalt 2011. aasta lõpuks väljamakse saanud inimesed ning nende asutatud ettevõtted. Tabelis on toodud toetust saanud ettevõtete jagunemine töötajate arvu järgi kolmanda majandusaasta aruannete põhjal, aruande esitas 420 ettevõtet. (Villsaar jt, 2014)

Töötajate arv	Protsent ettevõtetest
0	47%
1	39%
2	6%
3	2%
4	2%
5 või rohkem	4%

Leida
a) töötajate arvu jaotusfunktsiooni väärtused;
b) kui suurel osal ettevõtetest oli töötajate arv kuni kaks;
c) kui suurel osal ettevõtetest oli töötajate arv suurem kui kaks.

Vastus lk 665.

5.5. On hinnatud kahe erineva projekti võimalikku kasumit. Kumb projekt valida ja miks? *Oodatav väärtus*

Projekt A		Projekt B	
Tõenäosus	Kasum, €	Tõenäosus	Kasum, €
0,6	4 000	0,2	2 000
0,4	8 000	0,3	2 500
		0,3	4 000
		0,1	8 000
		0,1	12 000

Vastus lk 665.

5.6. Transpordiettevõte soovib osaleda vallasisese reisijateveo riigihankel. Hankel osalemiseks tuleb juurde osta üks buss. Valida on kahe variandi vahel: suur buss või väike buss. Väikese bussi korral on ühe reisi kulud küll väiksemad, aga kui reisijate arv on suur, siis tuleb teha tihedam graafik. See tähendab rohkem reise ning lisaks tõusevad ka tööjõukulud. Tabelis on prognoositud talitluskulud (kütusel- ja tööjõukulud) kummagi bussi korral sõltuvalt reisijate arvust. Kumb buss tuleks osta?

Reisijate arv aastas	Tõenäosus	Talitluskulud aastas, tuh eurot	
		Väike buss	Suur buss
Väike	0,2	110	170
Keskmine	0,5	180	200
Suur	0,3	260	220

Vastus lk 665.

5.7. Ettevõtja Kaarel tegeleb päikeseprillide maaletoomisega. Päikeseprillide müük toimub põhiliselt suvel, aga juba jaanuaris on tal vaja tellida kaup järgmiseks suveks. Loomulikult sõltub müük sellest, milline suvi tuleb: kas päikesepaisteline või vihmane. Tellimisel tuleb

Kaarlil valida, kas orienteeruda päiksepaistelisele suvele ja tellida suur kogus 100 tuhande euro eest või on oodata vihmast suve ja tuleks tellida väike kogus 40 tuhande euro eest. Kaarel teab, et kui tuleb päikeseline suvi, siis õnnestub tal suur kogus maha müüa ja planeeritav tulu on siis 120 tuhat eurot. Vihmase suve korral õnnestub maha müüa väike kogus ning planeeritav tulu on 48 tuhat eurot. Eelmisest suvest kaupa alles ei ole, s.t müüa saab ainult seda, mida ta jaanuaris tellib.

1. Milline variant tuleks Kaarlil valida, kui päikeselise ja vihmase suve tõenäosus on ühesugune?
2. Millistes piirides võivad tõenäosused varieeruda, et valikut ei peaks muutma?

VASTUS lk 665.

5.8. Ettevõte on viimase kolme aasta jooksul saanud erinevatel kuudel kasumit järgmiselt: 8 kuud 20 tuhat eurot, 16 kuud 30 tuhat eurot, 10 kuud 40 tuhat eurot ja 2 kuud 50 tuhat eurot. Leida oodatav kasum kuus. VASTUS lk 665.

5.9. Jäätisemüüja läbimüük sõltub ilmast. 20% päevadest on soe ilm, 30% külm ja ülejäänud päevadel keskmine. Sooja ilma korral on keskmine päevatulu 220 eurot, keskmise ilma korral 130 eurot ja külma ilma korral 40 eurot. Keskmine kulu päevas on 80 eurot. Leida oodatav kasum päevas. VASTUS lk 665.

5.10. Kaubitseja müüb turul piima. Piima ostab ta talunikelt hinnaga 30 senti liiter ja müüb hinnaga 45 senti liiter. 50 liitrit õnnestub tal maha müüa 10%-l päevadest, 100 liitrit müüb ta 40%-l, 200 liitrit 30%-l ja 300 liitrit õnnestub tal maha müüa 20%-l päevadest. Müümata jäänud piima ostab päeva lõpul temalt ära kohalik talunik hinnaga 7 senti liiter. Ühel päeval ostis kaupmees kauplemiseks 200 liitrit piima. Leida oodatav kasum sellel päeval. VASTUS lk 665.

5.11. Ettevõte, mis tegeleb torude paigaldamisega maasse, sõlmis lepingu 500 meetri toru panekuks. Kuivade ilmade korral pannakse päevas 20 meetrit toru, vihmaste ilmade korral töö efektiivsus langeb 60%. Vihmaste ilmade esinemise tõenäosus on 0,3. Leida oodatav päevade arv, mis kulub lepingu täitmiseks. VASTUS lk 665.

5.12. Hotellipidajad teavad, et mingi osa broneeringu teinud klientidest kohale ei ilmu ja seetõttu jääb osa tube tühjaks. Nende tubade eest tulu ei saada. Nimetame neid kliente, kes kohale ei ilmu, mitteilmujateks. Et vähendada mitteilmumisest tingitud tulude vähenemist, kasutavad hotellipidajad hotelli üle broneerimist (*overbooking*) (Bitran ja Gilbert, 1996; Toh ja Dekay, 2002). Kui hotell on ülebroneeritud ja kohale ilmub rohkem külastajaid, kui hotellis on kohti, siis need, kellele kohti ei jätku, paigutatakse mõnda teise hotelli. Sellise tulude optimeerimisega suurendas näiteks rahvusvaheline hotellikett Marriott

International (üle 535 000 hotelli) 1991. aastal oma tulu 25–35 miljoni dollari võrra (Baker ja Collier, 1999).

Ühes väikeses perehotellis on kolm tuba. Nädalavahetus maksab selles hotellis 100 eurot. Kasutades eelmiste perioodide andmeid, on hotellipidaja leidnud, kui suure tõenäosusega on mitteilmujate arv null kuni kolm (vt tabelit). Tõenäosus, et mitteilmujaid on neli või rohkem, on praktiliselt null.

Mitteilmujate arv	0	1	2	3
Tõenäosus	0,4	0,3	0,2	0,1

Kui hotell ülebroneerida ja osale saabujatest kohti ei jätku, on need võimalik paigutada lähedal asuvasse suurde hotelli, kus nädalavahetus maksab 150 €. Kuna broneeringu teinud kliendilt rohkem raha küsida ei tohi, siis vahe 50 € maksab kinni hotellipidaja, mille võrra väheneb tema tulu.

1. Kui suur on hotellipidaja oodatav tulu, kui ta väljastab kolm broneeringut?
2. Mitu broneeringut tuleks väljastada, et oodatav tulu oleks maksimaalne?

VASTUS lk 665.

5.13. Juhuslik suurus X allub pidevale ühtlasele jaotusele vahemikus $[20, 60]$. *Pidev ühtlane jaotus*

1. Panna kirja vastav jaotustihedus ja jaotusfunktsioon.
2. Leida järgmised tõenäosused:
   a) $P(X < 25)$;
   b) $P(X > 50)$;
   c) $P(25 < X < 50)$;
   d) $P(45 < X < 70)$.

VASTUS lk 665.

5.14. Teatud toote hind võib erinevatel müüjatel olla vahemikus 250–300 eurot. Eeldame, et hind allub selles vahemikus pidevale ühtlasele jaotusele. Leida hinna keskväärtus, mediaan ja standardhälve. Kui suur on tõenäosus, et hind on suurem kui 280 eurot? VASTUS lk 665.

5.15. 2013. aastal Eesti Statistikaameti läbiviidud Eesti sotsiaaluuringus küsitleti 15 053 isikut. Nendest 1646 valisid oma vanusevahemikuks variandi „25–34 aastat“. (*Eesti sotsiaaluuring* 2013) Mitme inimese vanus võis olla vahemikus 25–30 aastat? VASTUS lk 665.

5.16. Näidata, et jaotustihedusega (5.41) ristkülikjaotuse keskväärtus ja mediaan on mõlemad kohas $(a + b)/2$.

5.17. Ajakirjas Business Journal ilmus 23. oktoobril 2014 intervjuu *Binoomjaotus*

raamatu „Entrepreneurial StrengthsFinder“ ühe autori, Ph.D. Sangeeta Badaliga. Intervjuus mainiti, et ligikaudu 50% USA-s loodavatest ettevõtetest ei tegutse kauem kui viis aastat (Robison, 2014). Kui suur on tõenäosus, et viiest loodud ettevõttest üks ei tegutse kauem kui viis aastat? VASTUS lk 665.

5.18. 2012. aasta novembris tehtud Swedbanki uuring näitas, et Eesti suurimate kaubanduskettide kliendid eelistasid kaupade eest tasuda pangakaardiga. Näiteks Prisma Peremarketites oli kaardimaksete osakaal kõikidest ostutehingutest 63% (Inselberg, 2012). Leida tõenäosus, et kümnest Prisma Peremarketi ostjast

a) mitte ükski ei maksa kaardiga;
b) kõik maksavad kaardiga;
c) täpselt pooled maksavad kaardiga;
d) rohkem kui pooled maksavad kaardiga.
e) Kui suure tõenäosusega 50 ostjast rohkem kui pooled maksavad kaardiga?

VASTUS lk 665.

5.19. Tõenäosus, et tööpink vajab nädala jooksul seadistamist, on 0,2. Kui tsehhis on kuus tööpinki, leida tõenäosus, et nädala jooksul

a) ei vaja ükski tööpink seadistamist;
b) seadistamist vajab üks tööpink;
c) seadistamist vajab kaks tööpinki;
d) seadistamist vajab rohkem kui kaks tööpinki.

VASTUS lk 665.

5.20. Toetudes perioodilistele leibkonnauuringutele, avaldab Eesti Statistikaamet andmeid leibkonnaliikme kuu sissetuleku kohta tuludetsiilidena. Leida, kui suur on tõenäosus, et juhuslikult väljavalitud 10 leibkonna hulgas

a) ei ole ühtegi, kus sissetulek leibkonnaliikme kohta oleks väiksem kui esimene tuludetsiil;
b) on täpselt kaks peret, kus sissetulek leibkonnaliikme kohta on suurem kui kaheksas tuludetsiil.

VASTUS lk 665.

5.21. Tootmiseks vajaminevate detailide vastuvõtmiseks on kehtestatud järgmine süsteem: igast kastist valitakse juhuslikult välja 20 detaili. Kui nende hulgas on rohkem kui kaks defektset, siis seda kasti vastu ei võeta. Kui suur osa kastidest tagastatakse, kui on teada, et ligikaudu 5% detailidest on defektsed? VASTUS lk 665.

5.22. Tarnija saadetud tootepartii kõikne kontrollimine on suurte partiide korral väga töömahukas. Seepärast valitakse igast partiist juhuslikult mõned tooted ning testitakse ainult neid. Seda nimetatakse

proovivalimiks. Vastuvõtja poolt määratakse kindlaks partii aktsepteerimiseks lubatud defektsete toodete arv valimis m. Kui defektimäär p on defekti esinemise tõenäosus üksikul tootel ja n proovivalimi maht, siis suurte partiide korral defektsete toodete arv proovivalimis allub binoomjaotusele $B(n,p)$[5]. Tõenäosust $P(X \leq m)$ nimetatakse aktsepteerimise määraks. Aktsepteerimise määra sõltuvust defektimäärast kirjeldab nn OC-kõver (*Operating Characteristic Curve*). Selle kõvera korral on proovivalimi maht n ning lubatud defektsete toodete arv m fikseeritud. (Allen, 2006, ptk 10.6)

Olgu proovivalimis lubatud defektsete toodete arv $m = 1$. Konstrueerida ühes ja samas teljestikus kaks OC-kõverat $n = 25$ ja $n = 50$ korral. Defektimäär muutugu vahemikus 0 kuni 0,25. VASTUS lk 665.

5.23. Nafta puurimisega tegelev ettevõte kavatseb rajada neli uut puurauku. Tõenäosus, et puuraugust hakkab naftat saama, on 0,4, ja sellist puurauku nimetatakse õnnestunud puurauguks. Iga puuraugu rajamine maksab 200 tuhat dollarit. Nafta leidmisel on puuraugust saadav tulu 600 tuhat dollarit.

1. Kui suur on tõenäosus, et vähemalt üks puurauk õnnestub?
2. Kui suur on õnnestunud puuraukude arvu keskväärtus ja standardhälve?
3. Kui suur on ettevõtte oodatav kasum?
4. Arvestades kõiki võimalusi, kumb on suurema tõenäosusega, kas kasum või kahjum?

VASTUS lk 665.

5.24. Võlakiri on väärtpaber, mis näitab, et võlakirja väljalaskja on kohustatud kokkulepitud perioodi jooksul tagasi maksma võlakirja nimiväärtuse ja intressid. Krediidirisk on oht, et võlakirja väljalaskja ei täida oma kohustusi (muutub maksejõuetuks) ning võlakiri võib muutuda väärtusetuks. Krediidiriski saab vähendada, ostes mitmeid erinevaid võlakirju, s.t moodustades võlakirjade portfelli.

Üheks meetodiks, mida reitinguagentuur Moody's kasutab investeeringu oodatava kahju hindamisel, on nn binomiaalse laiendamise tehnika (*Binomial Expansion Technique*, BET) (Cifuentes ja O'Connor, 1996). Selle korral eeldatakse, et kõikide portfelli kuuluvate võlakirjade krediidirisk p on ühesugune ja sõltumatu ülejäänutest. Sõltumatuse eeldus kehtib siis, kui portfellis on võlakirjad ettevõtetelt, mis tegutsevad erinevatel tegevusaladel. Tõenäosus, et n võlakirjast koosneva

[5]Seda juhul, kui proovivalimi maht n on oluliselt väiksem partii suurusest N ning valimisse mittesattunud detailide arv on praktiliselt sama, mis partii suurus, s.t $N - n \approx N$. Väiksemate partiide korral, kui valimisse mittesattunud detailide arv märgatavalt väheneb, tuleb arvestada ka partii suurust N ning siis kasutatakse hüpergeomeetrilist jaotust.

portfelli korral krediidirisk realiseerub parajasti m võlakirja korral, leitakse binoomjaotuse valemist (5.49).

Üksiku võlakirja krediidirisk on 5% ning portfelli kuulub 20 üheaastast võlakirja.

1. Leida tõenäosus, et krediidirisk realiseerub
   a) vähemalt ühel võlakirjal;
   b) kõikidel portfelli kuuluvatel võlakirjadel.

Olgu ühe võlakirja nimiväärtus 50 eurot ning igalt võlakirjalt makstakse intressi 11% nimiväärtusest. Eeldame, et nende võlakirjade korral, mille väljalaskja muutus maksejõuetuks, ei saada intressi ega saada tagasi ka võlakirja ostmisel makstud raha (nimiväärtust).

2. Kui suur on tõenäosus, et võlakirjadelt saadav tulu ületab võlakirjade ostmisel tehtud kulu, s.t investeeringu kasum on positiivne?
3. Kui suur on oodatav kasum?
4. Kui suur on oodatav kasum siis, kui ostetakse samadel tingimustel (nimiväärtus 50 eurot, intress 11% nimiväärtusest) 20 ühe ja sama ettevõtte võlakirja, mille krediidirisk on 5%?

VASTUS lk 666.

5.25. IT-teenuse pakkuja kasutab servereid, kus maksimaalne samal ajal sisseloginud ehk konkureerivate kasutajate arv on ühe serveri kohta 50. Teenusepakkuja klientide koguarv on 720.

1. Kui palju peab teenusepakkujal servereid olema, et maksimaalne konkureerivate kasutajate arv ei ületaks serveripargi jõudlust?

Teenusepakkuja teab, et kõik tema kliendid ei ole pidevalt sisse loginud ja soovib vähendada serverite muretsemiseks tehtavaid kulutusi. Logifailide põhjal on teada, et keskmiselt on konkureerivate kasutajate arv pool klientide koguarvust ning eeldatakse, et iga kasutaja logib sisse sõltumatult teistest kasutajatest.

2. Kui palju peab servereid olema, et konkureerivate kasutajate arv ei ületaks serveripargi jõudlust 99,99% tõenäosusega?
3. Kui suur on eelmises punktis leitud serverite arvu korral tõenäosus, et konkureerivate kasutajate arv ületab serveripargi jõudlust?

VASTUS lk 666.

5.26. Ettevõttes toodetakse boilereid. Iga boileri korral tuleb teha kaheksa keevisõmblust. Peale boileri valmimist kontrollitakse kõik keevisõmblused kvaliteediinspektori poolt üle. Kui mõnel boileril avastatakse rohkem kui üks defektne keevisõmblus, informeeritakse töödejuhatajat keevitajast, kes selle boileri valmistas.

1. Keevitaja Peetri tehtud keevisõmblustest on 5% defektsed. Kui suurel osal tema valmistatud boileritest on defektseid õmblusi rohkem kui üks?

2. Poole aasta jooksul keevitab Peeter iga nädal 30 boilerit. Keskmiselt mitu korda nädalas tema nimi töödejuhatajale edastatakse?
3. Kui suur on tõenäosus, et Peetri nimi edastatakse töödejuhatajale rohkem kui kahel korral nädalas?

Vastus lk 666.

5.27. Katkendliku nõudluse (*lumpy demand*) korral nõudluse perioodid vahelduvad nõudluse puudumise või olulise vähenemise perioodidega. Sellise nõudluse korral on varude optimaalse taseme leidmiseks sobiv kasutada binoomjaotust (Pham, 2006). Eeldatakse, et iga kliendi tellimus sisaldab ainult üht eksemplari, s.t tellimuste summa võrdub kogunõudlusega vaadeldaval perioodil. Samuti eeldatakse, et kliendid on ühesugused, s.t tõenäosus, et neil vaadeldava perioodi jooksul tekib selle toote järgi vajadus, on ühesugune (Verganti, 1997). Tihti kasutatakse sellist lähenemist mitmesuguste varuosade (autod, lennukid, tööstusseadmed) optimaalse laovaru määramisel.

Piirkondlikus laos võtab ühe konkreetse varuosa laovarude täiendamine aega üks kuu alates tellimuse saatmisest. Järelikult tellimuse saatmise hetkel peab laovaru olema piisav rahuldamaks kuu aja nõudlust. Kliente, kes seda varuosa võivad laost vajada, on 50 ning statistika näitab, et kuu aja jooksul vajab seda keskmiselt 2% klientidest. Millise laovarude taseme juures tuleb vormistada uus tellimus, kui

a) see peab olema kuu nõudluse keskväärtusest kaheksa standardhälbe võrra suurem;
b) teenindustase, s.t tõenäosus, et seda varuosa jätkub kuuks ajaks, ei tohi olla väiksem kui 95%?

Viimasel juhul võib arvutusteks kasutada tabelarvutuse funktsiooni BINOM.INV. Vastus lk 666.

5.28. Ülesandes 5.12 oli hotellipidaja leidnud tõenäosused, et toa broneerinud klientidest ei ilmu kohale 0, 1, 2 või 3 klienti. Need tõenäosused ei sõltunud sellest, mitu broneeringut hotellipidaja väljastas.

Toetudes 12 hotellis läbiviidud uuringu andmetele (Toh ja Dekay, 2002), võtta üksiku mitteilmumise tõenäosuseks 3,7%. Eeldada, et mitteilmumised on üksteisest sõltumatud ja alluvad binoomjaotusele. Leida ülesandes 5.12 kirjeldatud hotellipidaja oodatav tulu, kui ta väljastab 3, 4 või 5 broneeringut. Vastus lk 666.

5.29. Lennuettevõtjad kasutavad tihti lennukite ülebroneerimist, s.t pileteid müüakse teadlikult rohkem, kui lennukis kohti on. Ülebroneerimise põhjuseks tuuakse asjaolu, et praktikas teatud osa reisijatest ei kinnita õigeaegselt oma broneeringut, ei saabu õigeks ajaks lennujaama või ei teavita lennuettevõtjat sellest, et nad on otsustanud valitud kuupäeval mitte lennata. Inglise keeles kasutatakse selliste broneeringute nimetamiseks terminit *no-show*. Kui lennuettevõtja kasutab ülebroneeringut, aga lennule ilmuvad kõik pileti broneerinud reisijad, on

lennuettevõtja sunnitud osa reisijaid lennult maha jätma. Vastavalt Euroopa Liidus kehtivale lennureisijate õigusi kaitsvale määrusele on lennust mahajätmise korral reisijal õigus saada rahalist hüvitist.

Sobiva ülebroneerimiste arvu Y määramiseks kasutatakse lennunduses erinevaid mudeleid. Lihtsama mudeli korral eeldatakse, et iga broneeringu kinnitamine on sõltumatu ülejäänud broneeringute kinnitamisest, s.t ignoreeritakse mitmekesi reisimist. Sellisel juhul, kui reisile on tehtud n broneeringut (sh ülebroneeringud), siis kinnitatud broneeringute arv m allub binoomjaotusele. Sobiva ülebroneerimiste arvu leidmiseks võib kasutada kas etteantud teenindustaset või oodatavat kasumit. (Taylor, 2007, ptk 19.14.1)

Lennukis Embraer-190 on 98 kohta. Olgu üksiku broneeringu kinnitamata jätmise tõenäosus 10%.

1. Milline võib olla maksimaalne ülebroneeringute arv, kui tõenäosus, et vähemalt üks reisija tuleb lennust maha jätta, ei tohi ületada 1%?
2. Teatud lennuliini pilet maksab 200 eurot, mis tuleb reisi broneerimisel kohe maksta. Lennult mahajäetud reisijale peab lennuettevõtja tagasi maksma piletiraha ning vastavalt Euroopa Liidu seadustele tuleb maksta ka hüvitist 250 eurot (kui lennu pikkus on kuni 1500 km). Need kokku on lennuettevõtja kulud mahajätmise korral. Eeldame, et muid kulusid, näiteks reisija hotelli paigutamine, ei esine. Milline peaks olema ülebroneerimiste arv, et sellest saadav oodatav kasum oleks maksimaalne?

VASTUS lk 666.

5.30. Tuletada valem variatsioonikordaja arvutamiseks binoomjaotusele alluva juhusliku suuruse korral. Kuidas variatsioonikordaja muutub, kui katsete arv n suureneb? VASTUS lk 666.

Poissoni jaotus

5.31. 1960-ndatel hakkas jõudsasti arenema arvutite tootmine. Shane M. Greenstein ja James B. Wade analüüsisid arvutiturgu aastatel 1968–1982 ja leidsid, et keskmiselt ilmus arvutiturule 0,37 uut toodet aastas ühe tootja kohta ning uute toodete arv aastas allus Poissoni jaotusele (Greenstein ja Wade, 1998). Leida tõenäosus, et mõni tootja tõi aastas turule

a) vähem kui ühe uue arvuti;
b) rohkem kui kaks uut arvutit.

VASTUS lk 666.

5.32. 2014. aastal viidi Harjumaal Viimsi vallas läbi teede ja tänavate liiklusuuring (Ess, 2014). Muuhulgas mõõdeti ka liiklusvoogusid Viimsi valla suurematel sissesõitudel. Vaatlused näitasid, et suvel tööpäeviti kella 8:00 ja 8:59 vahel sõidab Muuga tee kaudu valda sisse keskmiselt 82 autot tunnis ja vallast välja keskmiselt 53 autot tunnis.

1. Leida tõenäosused, et ühe minuti jooksul sõidab Muuga tee kaudu valda sisse m autot, kus $m = \{0, 1, 2, 3, 4, 5\}$. Konstrueerida jaotusseadust illustreeriv diagramm.
2. Leida tõenäosus, et ühe minuti jooksul sõidab valda sisse rohkem kui viis autot.
3. Leida tõenäosus, et ühe minuti jooksul ei sõida valda sisse ega vallast välja ühtegi autot.

VASTUS lk 666.

5.33. Politsei infotelefon 612 3000 töötab tööpäeviti hommikul kaheksast õhtul kuueni. Infotelefonile helistatakse keskmiselt 13 500 korda kuus ja kõnesid võtab vastu 10 inimest (Tartu Postimees, 11. märts 2014). Leida tõenäosus, et

a) minutis helistab politsei infotelefonile rohkem kui üks inimene;
b) 5 minuti jooksul helistab üle viie inimese.

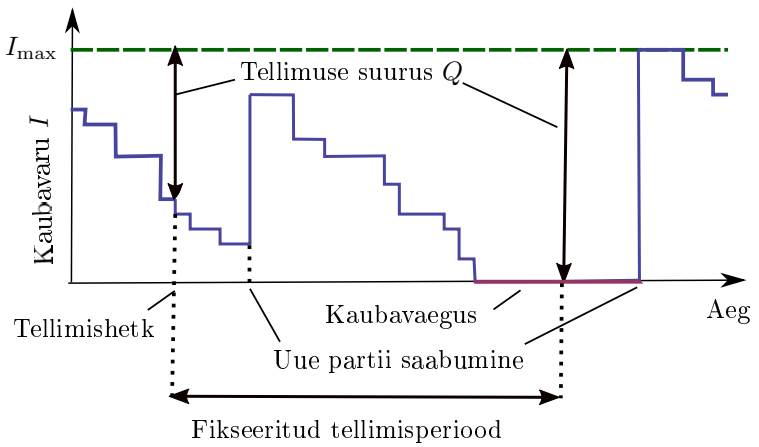
Tööpäevade arvuks kuus võtta 21. VASTUS lk 666.

5.34. Lennupiletite reserveerimise telefonile tuleb keskmiselt 48 kõnet tunnis. Kõnesid võtab vastu üks klienditeenindaja.

1. Kui suur on tõenäosus, et 5 minuti jooksul tuleb 3 kõnet?
2. Kui suur on tõenäosus, et 15 minuti jooksul tuleb 10 kõnet?
3. Kui suure tõenäosusega saab teenindaja puhata 3 minutit, kui momendil pole ühtegi pooleliolevat kõnet?

VASTUS lk 666.

5.35. Telefonipoes kasutatakse laovarude täiendamiseks fikseeritud tellimusperioodiga (*fixed-time period*) mudelit (vt joonis 5.32).



Joonis 5.32. Fikseeritud tellimusperioodiga kaubavarude juhtimine. Murdjoon kujutab müügist tingitud kaubavaru vähenemist

Iga kuu algul saadetakse tarnijatele uued tellimused, kus tellitava kauba kogus Q on maksimaalse kaubavaru I_{max} ning olemasoleva kaubavaru I vahe: $Q = I_{max} - I$. Igale kaubaartiklile on määratud oma

maksimaalne kaubavaru, mis sõltub selle kauba nõudlusest ning peab tagama kauba olemasolu kuu aja jooksul. Kuna nõudlus on juhuslik suurus, siis mõnikord võib juhtuda, et kaup lõpeb enne otsa, kui uus partii on kohale jõudnud.

Philipsi juhtmeta lauatelefoni Aloris 5100 jaoks on määratud maksimaalne kaubavaru $I_{max} = 10\,\text{tk}$. Ühes kuus ostetakse seda telefoni keskmiselt 8 tk.

1. Kui suur on tõenäosus, et nõudlus ületab pakkumise?
2. Mitmel kuul aastas võivad aparaadid otsa saada, enne kui uus partii kohtale jõuab?
3. Kui suur peaks olema maksimaalse kaubavaru minimaalne väärtus, et kaubavaeguse tõenäosus oleks väiksem kui 5%?

VASTUS lk 666.

5.36. G. Swartzman analüüsis ühes Michigani (USA osariik) haiglas patsientide haiglasse saabumise protsessi (Swartzman, 1970). Nelja nädala jooksul registreeriti 2359 patsiendi saabumine. Patsiendid võisid saabuda eelregistreerimisega ja eelregistreerimiseta. Eelregistreerimiseta patsiendid jagunesid nelja suurde kategooriasse, kelle saabumise kohta peeti eraldi arvestust. Analüüs näitas, et kõikides nendes kategooriates allus tunnis saabunud patsientide arv Poissoni jaotusele. Kuna patsientide saabumise sagedus oli päeva lõikes üldiselt erinev, jaotati ööpäev mitmeks osaks, mille kestel oli saabumise sagedus ligikaudu ühesugune. Tabelis on toodud keskmine tunni aja jooksul saabunud patsientide arv nendes neljas kategoorias hommikul kell 8:00–10:00.

Patsiendi kategooria	Patsienti tunnis
Erakorralist abi vajavad	3,12
Arsti suunamiskirjaga haiglasse	0,458
Ambulatoorsed patsiendid (röntgen, laboratoorium)	3,99
Ambulatoorsed patsiendid (muu)	2,99

1. Koostada diagramm, millel on kella 8 ja 10 vahel tunnis saabunud erakorralist abi vajavate patsientide arvu tõenäosusjaotus. Tunnis saabuvate patsientide arv võtta vahemikus 0 kuni 11.
2. Kui suur on tõenäosus, et kella 8 ja 10 vahel saabub parajasti kolm patsienti igast kategooriast?

Sõltumatute Poissoni jaotusele alluvate juhuslike suuruste $X_i \sim \text{Pois}(\lambda_i)$ summa on ka Poissoni jaotusega, mille parameeter $\lambda = \sum \lambda_i$. Järelikult tunnis haiglasse saabuvate eelregistreerimiseta patsientide koguarv allub Poissoni jaotusele, mille keskväärtus vahemikus 8:00–10:00 on tabelis toodud kategooriate keskväärtuste summa.

3. Leida, mitu eelregistreerimata patsienti saabub keskmiselt tunnis kella 8 ja 10 vahel.

4. Koostada diagramm, millel on kella 8 ja 10 vahel tunnis saabunud eelregistreerimiseta patsientide arvu tõenäosusjaotus. Tunnis saabuvate patsientide arv võtta vahemikus 0 kuni 25.

VASTUS lk 666.

5.37. Ajakirjas Journal of the American Statistical Association ilmunud artiklis analüüsis J. D. Davis mitmesuguste juhuslike sündmuste toimumist ja võrdles neid Poissoni jaotusega. Näiteks tehases, kus töötas üle 3000 töötaja, unustas iga päev teatud arv töötajaid maha oma personalikaardi, millega sai tehaseväravast sisse. 33 päeva jooksul toimunud vaatluse käigus selgus, et keskmiselt unustati päevas maha 27,79 personalikaarti. (Davis, 1952)

Leida, mitmel päeval on oodata tabelis märgitud arvu kaartide unustamist, s.t täita tabeli teine rida. VASTUS lk 666.

Päevas unustatud kaartide arv	0–23	24–26	27–29	30–32	33 ja rohkem
Oodatav päevade arv					

5.38. Alapeatüki 5.8 (Poissoni jaotus) alguses rääkisime, et Poissoni jaotus on binoomjaotuse piirjuht katsete arvu n lähenemisel lõpmatusele (vt ka lisa A.3). Näidata, et küllalt suure n ja väikese p korral annab Poissoni jaotus parameetriga $\lambda = np$ ligikaudu samad tulemused, mis binoomjaotus. Selleks koostada ühes ja samas teljestikus kaks graafikut, kus horisontaalteljel on m vahemikus 0 kuni 15 ning vertikaalteljel tõenäosus $P(X = m)$, mis on leitud

1) binoomjaotusest, kus $n = 50$ ja $p = 0{,}1$;
2) sama keskväärtusega Poissoni jaotusest.

VASTUS lk 666.

5.39. Joonisel 5.13 on näha, et kui Poissoni jaotuse keskväärtus $\lambda = 5$, on tõenäosused $P(X = 4)$ ja $P(X = 5)$ võrdsed. Näidata, et see on Poissoni jaotuse üldine omadus: kui jaotuse keskväärtus λ on täisarv, siis $P(X = \lambda - 1) = P(X = \lambda)$.

Eksponentjaotus

5.40. On teada, et sagedane töö katkestamine vähendab tootlikkust. Ajakirjas Business Journal ilmus 8. juunil 2006 intervjuu California Ülikooli professori Gloria Markiga, kes tutvustas oma uuringut tööaja kasutamisest. Uuringus osalesid juhtivtöötajad, finantsanalüütikud, projektijuhid, insenerid, tarkvaraarendajad. Uuring näitas, et kontoris töötavat inimest katkestatakse tema tegevuses keskmiselt iga 12 minuti tagant. Seejuures olid välja jäetud vähemtähtsad katkestamised, mis kestsid alla kahe minuti. (Robison, 2006)

1. Panna kirja kahe järjestikuse katkestamise vahelise aja jaotustihedus ja jaotusfunktsioon, kui aega mõõdetakse minutites.

2. Kui suur on tõenäosus, et töötajat katkestatakse järgmise 10 minuti jooksul?
3. Kui suur on tõenäosus, et töötajat ei katkestata 30 minuti jooksul?

VASTUS lk 667.

5.41. Elektri võrguteenuse pakkuja kodulehel on kirjas, et 70% elektririketest likvideeritakse kahe tunni jooksul. Eeldades, et rikke likvideerimise aeg allub eksponentjaotusele, leida,

a) milline on keskmine rikke likvideerimise aeg;
b) kui suur on tõenäosus, et rikke likvideerimiseks kulub rohkem kui viis tundi.

VASTUS lk 667.

5.42. Aktsiaturgude käitumist uurides ei analüüsita mitte ainult hindade ja tulumäärade muutumist, vaid ka kauplemise sagedust. Kuna hind kujuneb ostu-müügitehingutel, siis mida sagedamini need toimuvad, seda kiiremini hind muutub. Arvestades, et erinevate turul osalejate vahel sooritatud tehingud toimuvad sõltumatult, on alust eeldada, et kahe järjestikuse tehingu vaheline aeg allub eksponentsiaalsele jaotusele.

Ajakirjas Finance Letters 2005. aastal ilmunud artiklis analüüsisid autorid New Yorgi börsil 1999. aasta oktoobris tehtud 800 000 tehingu toimumisaega. Aktsiad, mille tehinguid analüüsiti, kuulusid Dow Jonesi tööstuse keskmisse indeksisse (*Dow Jones Industrial Average Index*). Näiteks General Motorsi aktsia korral oli keskmine aeg kahe järjestikuse tehingu vahel hommikupoole (9:00–10:59) 24,6 sekundit ja keskpäeval (11:00–13:59) 36,6 sekundit (Scalas jt, 2005). Nii hommiku kui ka keskpäeva jaoks leida tõenäosus, et General Motorsi aktsiatega ei tehta ühe minuti jooksul ühtegi tehingut. Kasutada selleks nii eksponentjaotust kui ka Poissoni jaotust. VASTUS lk 667.

5.43. Autoraadiote tootja on kindlaks teinud, et raadio keskmine riketeta tööaeg on 17 000 päeva (Majeske ja Herrin, 1998). Kui garantiiajaks on üks aasta, siis mitu protsenti toodangust vajab garantiiremonti? VASTUS lk 667.

5.44. Arvutitootja on garantiiremondiks planeerinud ressursse nii, et neid jätkub kuni 3%-le kõigist müüdud arvutitest. Kui pikk peaks olema arvuti garantiiaeg, kui keskmiselt esineb 0,02 riket aastas? VASTUS lk 667.

5.45. Michigani haigla erakorralise meditsiini osakonnas on öisel ajal (kell 2:00–8:00) kahe järjestikuse patsiendi saabumise vaheline keskmine aeg 39,40 minutit (Swartzman, 1970). Leida tõenäosus, et

a) kahe järjestikuse patsiendi saabumise vahele jääb rohkem kui üks tund;

b) kahe järjestikuse patsiendi saabumise vahele jääb rohkem kui kaks tundi;
c) järgmise patsiendi saabumiseni jääb vähemalt üks tund, kui eelmise saabumisest on möödunud juba üle tunni aja.

VASTUS lk 667.

5.46. Teenindussaalis pole parajasti ühtegi klienti ning klienditeendindaja palub osakonnajuhatajalt luba viieks minutiks lahkuda. Osakonnajuhataja luba ei anna. Ta põhjendab keeldumist sellega, et kuna tükk aega ei ole ühtegi klienti sisenenud, siis tõenäosus, et järgmise viie minuti jooksul keegi tuleb, on väga suur. Kas osakonnajuhataja põhjendus on loogiline? VASTUS lk 667.

5.47. Poissoni jaotus ja eksponentjaotus on omavahel seotud ning mõningate probleemide korral võib kasutada kas üht või teist jaotust, tulemus on sama.

Olgu meil tegemist Poissoni protsessiga, kus keskmine sündmuste arv ajaühikus on λ.

1. Kasutades Poissoni valemit (5.62), leida tõenäosus, et sündmuste arv ajaühikus on 0.
2. Poissoni jaotuse keskväärtus λ on võrdne vastava eksponentjaotuse parameetriga λ. Kui sündmuste arv ajaühikus on 0, siis kahe järjestikuse sündmuse vaheline ajavahemik on suurem kui valitud ajaühik. Kasutades eksponentjaotuse jaotusfunktsiooni (5.67), leida tõenäosus, et kahe järjestikuse sündmuse vaheline aeg on suurem kui üks.

VASTUS lk 667.

Normaaljaotus

5.48. Exir on üks Iraani juhtivaid farmaatsiatehaseid, kus toodetakse vähemalt 250 erinevat toodet. Tootmisprotsessis läbivad ravimid mitu tootmisliini ning lõpus on üheks oluliseks liiniks pulberpakendite täiteliin. Tootmisprotsessi optimeerimiseks ja kitsaskohtade avastamiseks koguti ühe aasta jooksul (2011 märts kuni 2012 märts) andmeid tootmisliinide koormuse kohta. Selgus näiteks, et pulberpakendite täiteliini päevatoodang allub normaaljaotusele keskväärtusega 44 600 ja standardhälbega 1540 toodet (Eskandari, Babolmorad ja Farrokhnia, 2013). Leida tõenäosus, et liini päevatoodang on

a) väiksem kui 43 000 tk;
b) suurem kui 46 000 tk;
c) 43 000 ja 46 000 vahel.

VASTUS lk 667.

5.49. Joonisel 5.18 oli toodud Xeroxi aktsia tulumäär päevas 8. veebruar kuni 9. august 2013. Tulumäära keskväärtus oli 0,2% ja standardhälve 1,5%. Eeldades, et tulumäär allub normaaljaotusele, leida

tõenäosus, et Xeroxi aktsia tulumäär päevas on negatiivne. VASTUS lk 667.

5.50. Hanois (Vietnam) on põhilisteks transpordivahenditeks mootorrattad, motorollerid ja mopeedid, mille osakaal on üle 80% kõikidest mootorsõidukitest. Linnas viidi läbi uuring, kus mõõdeti mootorrataste kiirust erineva laiuse ja liiklustihedusega tänavatel. Mootorrataste alla liigitusid kõik kaherattalised mootorsõidukid. Analüüs näitas, et nende kiirus allub normaaljaotusele, mille parameetrid on erinevatel tänavatel erinevad. Näiteks ühel tänaval oli mootorrataste keskmine kiirus 32,3 km/h standardhälbega 5,7 km/h, teisel tänaval aga 32,7 km/h standardhälbega 5,2 km/h. (Minh, Sano ja Matsumoto, 2005)

1. Püstitada hüpotees, kummal tänaval on kiiremini kui 40 km/h sõitvate mootorrataste osakaal suurem.
2. Kontrollida oma hüpoteesi vastavate arvutustega.

VASTUS lk 667.

5.51. Ülesandes 5.36 analüüsiti eelregistreerimiseta patsientide saabumist ühes Michigani haiglas, mis vastas Poissoni protsessile. Lisaks sellele saabus haiglasse ka eelregistreerimisega ambulatoorseid patsiente, kellele oli määratud kindel vastuvõtuaeg. Enamik neist patsientidest jõudis kohale mõni minut enne vastuvõtuaega, kuid oli ka neid, kes hilinesid. Patsientide saabumisaja erinevus vastuvõtuajast allus normaaljaotusele. Näiteks füsioteraapia kabinetti tulnud patsientide saabumisaja erinevus vastuvõtuajast allus normaaljaotusele keskväärtusega -2 minutit ja standardhälbega 12 minutit, kus keskväärtuse negatiivne väärtus tähendab varem tulemist. (Swartzman, 1970)

1. Kui suur osa patsiente saabus füsioteraapia kabinetti rohkem kui 10 minutit enne vastuvõtuaega?
2. Kui suur osa patsiente hilines füsioteraapia kabinetti rohkem kui 5 minutit?
3. Oletame, et füsioteraapia kabinetis kulub iga patsiendi peale 30 minutit ning vastuvõtuajad on parajasti sellise intervalliga. Kabineti ukse taga on oma aega ootavate patsientide jaoks üks tool. Parajasti kutsutakse kabinetti uus patsient ning see tool vabaneb. Kui suur on tõenäosus, et tool hõivatakse kohe järgmise patsiendi poolt?
4. Kui suur on tõenäosus, et see tool on järgmised 15 minutit tühi?

VASTUS lk 667.

5.52. Ühiskondliku transpordi sõiduplaanide koostamisel ning ümberistumiseks kuluva aja planeerimisel tuleb arvestada jalakäijate liikumiskiirusega. Jalakäija liikumiskiirus sõltub mitmetest individuaalsetest karakteristikutest, samuti ilmast, teeoludest ja muudest välistest tingimustest. Erinevate uuringute põhjal allub jalakäijate kiirus

normaaljaotusele keskväärtusega 1,34 m/s ja standardhälbega 0,37 m/s (Daamen ja Hoogendoorn, 2007).

Ümberistumiseks ühelt rongilt teisele peab läbima 260 meetrit. Aeg ühe rongi saabumise ning teise väljumise vahel on neli minutit.

1. Kui suur osa reisijatest jõuab selle ajaga ümber istuda, kui eeldada, et kõik liiguvad neile omase kiirusega ning ei jookse?
2. Milline peaks olema kahe rongi vaheline aeg, et vähemalt 95% reisijatest jõuaks ümber istuda?

VASTUS lk 667.

5.53. Leida normaaljaotuse keskväärtusele vastav jaotustiheduse väärtus, kui standardhälve

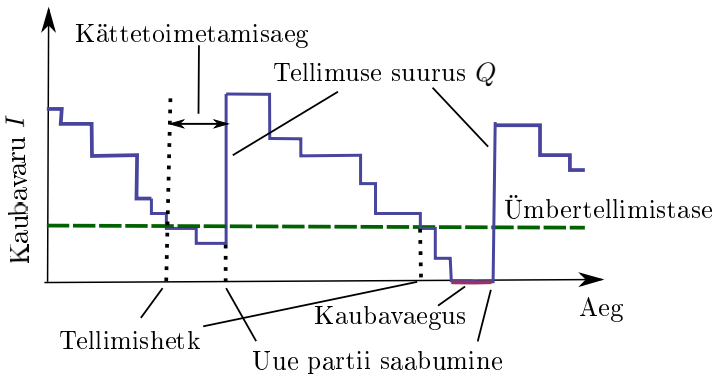
a) $\sigma = 1$;

b) $\sigma = 4$.

VASTUS lk 667.

5.54. Joogi kogus karastusjookide tootmisliinilt tulevates pudelites varieerub veidi ja see varieerumine allub normaaljaotusele. Milline võib olla maksimaalne standardhälve, kui vähemalt 99,7% pudelites peab joogi kogus olema vahemikus 495 kuni 505 milliliitrit? VASTUS lk 667.

5.55. Ehitusmaterjalide poes kasutatakse kaubavarude juhtimiseks fikseeritud tellimuskogusega (*fixed order quantity*) mudelit. Sellisel juhul tehakse uus tellimus siis, kui varud on langenud teatud tasemeni, mida nimetatakse ümbertellimistasemeks (joonis 5.33). Arvestada tuleb kättetoimetamisajaga ning ümbertellimistase leitakse iga kaubaartikli jaoks nii, et varudest jätkuks kuni uue partii saabumiseni. Tellimuse suurus on iga kord ühesugune. Tingituna nõudluse juhuslikkusest võib siiski mõnikord tekkida olukord, kus kaup lõpeb enne otsa, kui uus partii kohale jõuab.



Joonis 5.33. Ülesande 5.55 juurde: fikseeritud tellimuskogusega kaubavarude juhtimine. Murdjoon kujutab müügist tingitud kaubavaru vähenemist

Ühe konkreetse kaubaartikli korral on kättetoimetamisaeg üks nädal. Ostujuht teab, et seda kaupa ostetakse nädalas keskmiselt 750 tk

standardhälbega 210 tk ning nõudlus allub normaaljaotusele. Kui suur peab olema selle kauabaartikli ümbertellimistase, et kaubavaeguse tõenäosus poleks suurem kui 5%? VASTUS lk 667.

5.56. Ajakirjas The Accounting Review 1983. aastal ilmunud artiklis analüüsiti erinevate ettevõtete rahandussuhtarvude kõikumist. Vaatluse all oli periood 1950–1979 ning ettevõtete arv kõikus vahemikus 346 aastal 1950 kuni 1243 aastal 1978. Artikli autorid leidsid, et kui erindid eemaldada, siis vaadeldud rahandussuhtarvud alluvad normaaljaotusele. Näiteks ettevõtte likviidsust iseloomustav lühiajaliste kohustuste kattekordaja (käibevara jagatud lühiajaliste kohustustega) allus normaaljaotusele, mille keskväärtus oli 0,65 ja dispersioon 0,011. (Frecka ja Hopwood, 1983) Leida selle suhtarvu jaoks keskväärtuse suhtes sümmeetriline vahemik, millesse langeks 75% ettevõtte vastav näitaja. VASTUS lk 667.

5.57. Elektroonikakomponente tootvas ettevõttes jagatakse takistid nelja kvaliteediklassi vastavalt sellele, mitu protsenti on kõrvalekalle nominaalväärtusest. Tabelis on toodud vastavad määrad ning 100 oomise nominaalväärtusega takistite hulgihind. Kvaliteedikontroll on näidanud, et 100 Ω takistite tootmisel vastab toodangu keskväärtus nominaalväärtusele ja standardhälve on 3 Ω. Kui suur on oodatav tulu, kui toodetakse 10 000 takistit nominaalväärtusega 100 Ω? VASTUS lk 667.

Lubatud kõrvalekalle nominaalväärtusest	Kvaliteediklass	Tükihind, €
Kuni 1%	Eriti kõrge kvaliteediga	0,016
1% kuni 2%	Kõrge kvaliteediga	0,012
2% kuni 5%	Standardkvaliteediga	0,005
5% kuni 10%	Madala kvaliteediga	0,001
Rohkem kui 10%	Praak	

5.58. Peeter elab bussipeatusest 2 km kaugusel ning normaalse tempoga kõndides kulub tal koju jõudmiseks 20 minutit. Peeter teab, et ta kõnnib paljudest jalakäijatest kiiremini. Kasutades ülesandes 5.52 toodud andmeid jalakäijate liikumiskiiruse kohta, leida, mitmendasse kümnendikku Peeter oma kõndimiskiirusega kuulub. VASTUS lk 667.

5.59. Leida, kui suur osa normaaljaotusele alluva suuruse väärtustest jääb keskväärtusest paremal pool keskväärtusest kaugemale kui

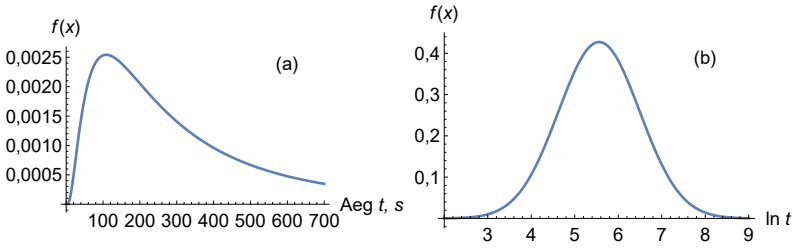
a) 3σ;
b) $4{,}5\sigma$;
c) 6σ.

Kõikidel juhtudel leida ka vastav defektimäär, s.t mitu tükki miljoni kohta. Nõuanne: kasutada võib standardiseeritud normaaljaotuse jaotusfunktsiooni, mis tabelarvutuses leitakse NORM.S.DIST abil. VASTUS lk 667.

5.60. Telefonikõne pikkus on juhuslik suurus, mille jaotuse uurimine aitab planeerida näiteks ettevõtte infotelefonile vastavate klienditeenindajate arvu. On tähele pandud, et mida kauem on kõne kestnud, seda tõenäosem on, et see kestab veel. Seda nähtust seletatakse psühhofüüsikast tuntud Weberi-Fechneri seadusega: mida intensiivsem on stiimul, seda aeglasemalt kasvab tajumulje. Näiteks kui kõne on kestnud üks minut, siis järgmist minutit tajutakse sama pikana kui kümmet sekundit siis, kui kõne on kestnud 30 sekundit. Sellest tulenevalt ei allu normaaljaotusele mitte telefonikõne pikkus, vaid kõne pikkuse logaritm. (Bolotin, 1994)

Kui juhusliku suuruse X naturaallogaritm allub normaaljaotusele, $\ln X \sim \mathrm{N}(\mu, \sigma)$, siis juhuslik suurus ise allub logaritmilisele normaaljaotusele (vt ka joonis 5.34). Kui on teada logaritmilise normaaljaotuse keskväärtus μ_{LN} ja standardhälve σ_{LN}, siis vastava normaaljaotuse keskväärtus μ ning standardhälve σ avalduvad nende kaudu järgmiselt:

$$\mu = \ln\left(\frac{\mu_{\mathrm{LN}}}{\sqrt{1+\left(\frac{\sigma_{\mathrm{LN}}}{\mu_{\mathrm{LN}}}\right)^2}}\right), \quad \sigma = \sqrt{\ln\left(1+\left(\frac{\sigma_{\mathrm{LN}}}{\mu_{\mathrm{LN}}}\right)^2\right)}. \tag{5.86}$$



Joonis 5.34. Kõne pikkuse logaritmiline normaaljaotus (a) ja vastav normaaljaotus (b)

Optimeerimaks klienditeenindusele kuluvaid ressursse, analüüsiti ühe panga kõnekeskuse tööd 12 kuu vältel. Selle aja jooksul tehti panka 450 000 kõnet. Registreeriti kõnede pikkus sekundites, mis allus logaritmilisele normaaljaotusele. Kõned jaotati nelja suurde gruppi: internetipanga abi, potentsiaalsed kliendid, aktsiatega kauplemine ja muud.

Keskmiselt kõige pikemad olid kõned, kus sooviti abi internetipanga kasutamise kohta. Nende kõnede pikkuse keskmine oli 401 sekundit ja standardhälve 473 sekundit. (L. Brown jt, 2005)

1. Leida selle normaaljaotuse keskväärtus ning standardhälve, millele allub internetipanga kasutamise kohta tehtud kõnede pikkuse naturaallogaritm.
2. Kui suur on tõenäosus, et kõne, kus soovitakse abi internetipanga kasutamise kohta, kestab
   a) alla viie minuti;
   b) rohkem kui 10 minutit.
3. Leida tõenäosus, et kõne kestab veel vähemalt viis minutit, kui see on juba kestnud vähemalt üks minut.
4. Leida tõenäosus, et kõne kestab veel vähemalt viis minutit, kui see on juba kestnud vähemalt viis minutit. Võrrelda seda tõenäosust eelmises punktis leitud tõenäosusega.

VASTUS lk 667.

Erinevad jaotusseadused

5.61. Ühel teelõigul toimub keskmiselt üks avarii nädalas. Leida, mitmel nädalal aastas ei tohiks toimuda ühtegi avariid ja mitmel nädalal aastas on oodata kolme avarii toimumist. VASTUS lk 667.

5.62. Toodangu kvaliteedi kontrollimisel võetakse tootmisliinilt juhuslikult 3 toodet. Kui neist üks või rohkem on defektsed, võetakse liinilt juhuslikult veel 3 toodet. Kui ka selles partiis on vähemalt üks defektne toode, siis tootmisliin peatatakse. Kui on teada, et defektse toote tõenäosus on 0,05, leida tõenäosus, et

a) tuleb võtta teine proovipartii;
b) liin peatatakse.

VASTUS lk 667.

5.63. Tipptundidel toimub linnas keskmiselt kaks avariid tunnis. Hommikul kestab tipptund 1 tund 30 minutit ja õhtul 2 tundi. Leida tõenäosus, et

a) hommikusel tipptunnil ei toimu ühtegi avariid;
b) õhtusel tipptunnil toimub 2 avariid;
c) hommikusel tipptunnil toimub 4 või rohkem avariid.
d) Kui suure tõenäosusega ei toimu ühtegi avariid ei hommikusel ega ka õhtusel tipptunnil?

VASTUS lk 667.

5.64. Viis raamatupidajat tegid 10 päeva jooksul kokku 88 005 raamatupidamiskannet. Analüüs näitas, et nendest 218 olid vigased (Davis, 1952).

1. Leida vea tõenäosus üksiku raamatupidamiskande korral.
2. Audiitor kontrollib juhuslikult väljavalitud 100 kannet. Kui suur on tõenäosus, et ta leiab vähemalt ühe vigase kande?

Vastus lk 667.

5.65. Keskmiselt sõidab auto nädalas läbi 800 km standardhälbega 90 km. Leida tõenäosus, et auto sõidab nädalas läbi 900–1000 km. Mitmel nädalal aastas peaks läbisõit jääma sellesse vahemikku? Nädalate arv aastas on 52. Vastus lk 667.

5.66. Valikvastustega testis on 20 küsimust ja igal küsimusel on kaks vastusevarianti, millest üks on õige. Kui üliõpilane pole testiks õppinud ja kõikide küsimuste korral valib vastusevariandi juhuslikult, siis kui suure tõenäosusega vastab ta õigesti rohkem kui pooltele küsimustele? Kui suur on see tõenäosus aga siis, kui igal küsimusel on kolm vastusevarianti ja ainult üks neist on õige? Vastus lk 667.

5.67. Volli elab tänavas, mida läbib keskmiselt kolm autot tunnis. Kui suure tõenäosusega saab Volli vähemalt poolt tundi vaikust nautida? Vastus lk 667.

5.68. 2013. aastal oli Eestis suurettevõtteid (töötajate arv 250 ja enam) 0,16% kõikidest ettevõtetest. Kui konverentsil on osalejaid 120 Eesti ettevõttest, siis kui suure tõenäosusega on konverentsil esindaja vähemalt ühest suurettevõttest? Eeldada, et konverentsil osalemise soov ei sõltu ettevõtte suurusest. Vastus lk 667.

5.69. Ajakiri Consumer Reports viis läbi uuringu, milles küsitleti 16 tuhandet ajakirja lugejat, kes olid aastatel 2007–2011 ostnud kööki uue pliidi. Selgus, et tootja Whirlpool pliidi soetanutest pidid 5% selle välja vahetama või laskma parandada. Tootja Jenn-Air pliidi soetanute hulgas oli selliseid 12%. (CR Buying Guide, 2013)

1. Kui suure tõenäosusega on juhuslikult valitud 20 Whirlpooli pliidi soetanu hulgas täpselt kaks ostjat, kes on pidanud seda parandama või selle välja vahetama?
2. Kui suure tõenäosusega on juhuslikult valitud 20 Jenn-Airi pliidi soetanu hulgas täpselt kaks ostjat, kes on pidanud seda parandama või selle välja vahetama?
3. Kui suure tõenäosusega on juhuslikult valitud 20 Jenn-Airi pliidi soetanu hulgas vähemalt üks, kes on pidanud seda parandama või selle välja vahetama?

Vastus lk 668.

5.70. Tuletõrjebrigaad A saab keskmiselt viis väljakutset päevas. Ühe tulekahju kustutamiseks arvestada keskmiselt üks tund.

1. Kui suure tõenäosusega saavad brigaadi liikmed puhata rohkem kui kaks tundi järjest?
2. Kui suure tõenäosusega tuleb väljakutsele saata brigaad teisest linnaservast, sest brigaad A on veel eelmist tulekahju kustutamas?

Vastus lk 668.

ÜL05Jaotused

Järgmiste ülesannete andmed on failis ÜL05Jaotused

A.5.1. Kui töötajaga juhtub tööõnnetus ja tööandja on selle eest vastutav, siis peab ta korvama tööõnnetuse tagajärjel tekkinud kahju. Tööandja vastutuskindlustus on ette nähtud juriidiliste isikute vastutuse kindlustamiseks tema alluvuses töötavate inimeste ohutuse eest. Tabelis on toodud tööandja vastutuskindlustuse lepingute, kahjujuhtumite ja maksete andmed, mis on võetud Finantsinspektsiooni avaldatud kindlustusseltside koondandmetest 2014. aasta I kuni IV kvartali kohta[6].

1. Leida keskmine kahjujuhtumi tõenäosus kvartalis.
2. Leida keskmise hüvitise suurus ühe kahjunõude kohta.

Olgu kindlustusseltsis sõlmitud 500 tööandja vastutuskindlustuse lepingut. Järgmiste arvutuste tegemisel kasutada punktides 1. ja 2. leitud tulemusi.

3. Leida selle kindlustusliigi maksete pealt kvartalis kogutava tulu miinimumväärtus, mis võrdub oodatava kuluga kvartalis.
4. Leida minimaalse omakapitali suurus, nii et rahalisi ressursse (tulu pluss omakapital) jätkuks kahjuhüvitiste väljamaksmiseks tõenäosusega 99,5%.

VASTUS lk 668.

A.5.2. Tabelis on toodud katkestustega väljalülitumiste arv Eleringi hallatavas elektrivõrgus aastatel 2006–2014 (Võrguteenuste . . .).

1. Leida katkestuste arvu keskväärtus ja dispersioon aastas.
2. Kas katkestuste arv aastas allub Poissoni jaotusele?
3. Mitmel aastal 20st võib katkestuste arv aastas olla suurem kui 30?

VASTUS lk 668.

A.5.3. AS ÕLI müüb aastas keskmiselt 144 000 liitrit diiselmootoriõli. Analüüs on näidanud, et varude täiendamisel on optimaalne tellimuskogus 40 vaati, s.o 8000 liitrit. Tellimuse saatmisest uue kaubakoguse kättesaamiseni kulub 1 nädal. Varude suurust on võimalik pidevalt jälgida ja uus tellimus tuleb teha, kui varude kogus on vähenenud teatud tasemeni, mida nimetatakse ümbertellimushetkeks. Ostujuht otsustab leida sellise ümbertellimushetke, et tõenäosusega 95% jätkuks allesjäänud kogusest uue partii saabumiseni, s.o üheks nädalaks. Selleks on tal kasutada eelmise aasta müügimahud nädalate lõikes (vt tabe-

[6] Finantsinspektsioon `https://www.fi.ee/index.php?id=3271`, kahjukindlustuse lepingud, preemiad ja nõuded.

lit). Ostujuht eeldab, et nõudlus allub normaaljaotusele. Millise varude taseme juures tuleb teha uus tellimus? VASTUS lk 668.

A.5.4. 2007. aastal pakkus SEB pank investeerimishoiust, mille tulusus sõltus USA dollari käitumisest euro suhtes. Aluseks oli EUR/USD valuutakurss, millele oli leitud ülemine ja alumine barjäär. Kui dollari kurss ei puudutanud vaatlusperioodi jooksul kordagi ülemist ega alumist barjääri, siis maksti hoiuse lõpptähtpäeval maksimaalset intressi. Kui aga valuutakurss puudutas ülemist või alumist barjääri, siis intressi ei makstud. Ülemise ja alumise barjääri määras pank.

Tabelis on toodud dollari kurss euro suhtes perioodil 2.02.–24.02.2015. Eeldades, et kursimuutused alluvad normaaljaotusele, leida alumine ja ülemine barjäär, nii et nende vahele jäämise tõenäosus oleks 90% ja vahemik oleks sümmeetriline keskväärtuse suhtes. VASTUS lk 668.

A.5.5. Vetropack Straza D.D. on Horvaatias asuv klaasitehas, mis toodab muuhulgas ka klaaspudeleid. Pudelid peavad olema kerged, läbipaistvad, odavad ja samal ajal vastupidavad temperatuurile ning mehaanilistele mõjudele. Klaaspudelite tootmise tehnoloogia on keerukas ning pudelite kvaliteedi tagamiseks tuleb tootmisprotsessi pidevalt täiustada.

Teatud tüüpi pudelite nominaalmass on 190 grammi ning lubatud kõrvalekalle nominaalist mõlemale poole 1 gramm. Kõik need pudelid, mille mass erineb nominaalmassist rohkem kui 1 g võrra, tunnistatakse praagiks. Tootmisliinilt tulnud pudelite mass allub normaaljaotusele. Tabelis on liinilt L621 võetud 80 pudeli mass. (Kovačec, Pilipović ja Štefanić, 2010)

1. Leida pudelite massi aritmeetiline keskmine ja standardhälve.
2. Kasutades eelmises punktis leitud parameetritega normaaljaotust, leida, kui suur osa tootmisliinilt L621 tulevast toodangust on praak.

VASTUS lk 668.

Peatükk 6

Valikuuringud

2012. aasta algul viidi Eestis läbi rahvaloendus, mille käigus küsitleti kõiki Eesti elanikke. Küsitlejaid oli 2000 ja nad külastasid ligikaudu 500 tuhandet isikut (65% täitsid ankeedi veebipõhiselt) (Tiit, 2014). Rahvaloenduse planeeritud kulud olid 18,9 miljonit eurot.

Aastatel 2011–2012 viidi Eestis läbi valikuuring „Tean ja oskan“, mille käigus küsitleti 13 tuhandet isikut. Uuringu eelarve oli 1,6 miljonit eurot (Raudvere, 2011).

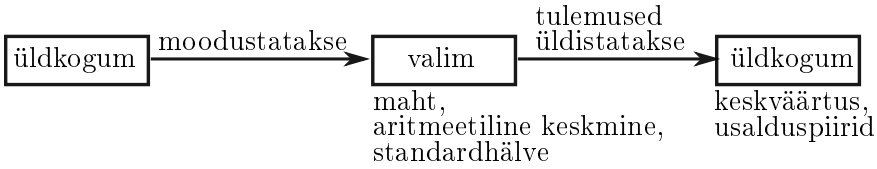
Kõikse vaatluse suured kulud on üks põhjustest, miks kasutatakse valikuuringuid.

6.1. Kogum, valim ja valikumeetodid

Statistiline uuring võib olla

- kõikne: uuritakse läbi kogu üldkogum;
- valikuline: uuritakse läbi üldkogumit esindav osa valim.

Valikuuringu **eemärgiks** on valimi põhjal järelduste tegemine **üldkogumi** kohta. See tähendab, et valim on vahend üldkogumi analüüsimiseks.



Valikuuringu **kasutamise põhjused**:

- väiksem maksumus;
- suurem kiirus, tulemused saab kiiremini kätte;
- suurem paindlikkus;
- laiem rakendatavus — võimaldab kasutada spetsiaalset aparatuuri või spetsiaalselt ettevalmistatud intervjueerijaid;

- suurem täpsus andmete kogumisel — väiksema töömahu tõttu võimalik kasutada kõrgema kvalifikatsiooniga tööjõudu, suurema kontrolliga vähendada andmete kogumisel ja töötlemisel tekkivaid vigu;
- objekti testimine võib teatud juhtudel rikkuda objekti ning testimiseks tuleb kasutada proovivalimit:
  - peale vedrumadratsi vastupidavuse testimist ei saa seda madratsit enam müüa;
  - peale toiduainetest proovide võtmist ei saa neid enam müüa;
  - klaasi tugevuse testimisel klaas puruneb.

Järgnevalt mõningad mõisted, mida valikuuringutes kasutatakse.
Objekt on indiviid, organisatsioon, nähtus, mida mõõdetakse, vaadeldakse või küsitletakse.
Üldkogum (*population*) on objektide hulk, mille kohta soovitakse vastavalt probleemülesandele saada informatsiooni.
Osakogum (*subpopulation*) on üldkogumi alamhulk, mis on fikseeritud mõne tausttunnuse või uuritava tunnuse väärtuste järgi ja mille kohta soovitakse esitada eraldi hinnanguid. Näiteks kui üldkogumiks on Eesti elanikud, siis osakogumiteks võivad olla mehed ja naised või linna- ja maaelanikud.
Loend, freim (*sampling frame*) on vahend pääsemiseks üldkogumi objektide juurde. Selleks võib olla loetelu, nimekiri, register, andmebaas.
Valim (*sample*) on üldkogumi osahulk, mis määratakse kindlate valikumeetoditega. Objektide võtmine valimisse toimub loendi abil.
Kadu (*loss*) on valimi osa, mis mingil põhjusel jääb uuringust kõrvale (ei saada kätte, ei vasta).
Kao määr (*loss rate*) on kao osakaal valimis.
Vastamismäär (*response rate*) on vastanute osakaal valimis.

Erinevate uuringute korral võib vastamismäär olla erinev. 2008. aastal ilmus ajakirjas Human Relations artikkel, mille autorid olid läbi töötanud 1607 uuringu aruannet, mis avaldati akadeemilistes ajakirjades aastatel 2000–2005. Need uuringud kokku hõlmasid 100 tuhandet organisatsiooni ja 400 tuhandet isikut. Keskmine vastamismäär isikuküsitluste korral oli 52,7% standardhälbega 20,4%. Keskmine vastamismäär organisatsioonide korral oli 35,7% standardhälbega 18,8%. (Baruch ja Holtom, 2008)

Tabelis 6.1 on Eesti Statistikaameti ettevõtete aastastatistika üldkogum, valim ja vastanute arv kolmel erineval aastal[1]. Eesti 2014. aasta tööjõu-uuringus oli keskmine vastamismäär erinevates maakondades

[1] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel EM026: ettevõtete aastastatistika üldkogum, valim ja vastanud.

74,8%, kõige väiksem Hiiu maakonnas (58,3%) ja kõige suurem Saare maakonnas (86,7%)[2].

Tabel 6.1. Eesti Statistikaameti ettevõtete aastastatistika üldkogum, valim ja vastanud

Aasta	Kogum	Valim	Vastanud	Valimi osatähtsus kogumis, %	Vastamis-määr, %
2005	44 094	12 613	10 305	28,6	81,7
2010	61 293	11 643	9 231	19,0	79,3
2012	69 694	12 037	8 844	17,3	73,5

Valikumeetodid võivad olla tõenäosuslikud ehk juhuslikud ning empiirilised.

Tõenäosusliku valikumeetodi (*probability/random sampling*) korral

- on kõigile üldkogumi objektidele tagatud teadaolev valimisse kaasamise tõenäosus;
- on võimalik leida valimi põhjal tehtud hinnangute usaldusvahemikku ja teame tõenäosust, et meid huvitava parameetri väärtus kogumis langeb sellesse vahemikku.

Empiirilise valiku (*empirical sampling*) korral ei ole objektide valimisse sattumise tõenäosused teada. Eesmärgiks on saada valim, mille struktuur langeb kokku üldkogumi struktuuriga.

Objekti juhuslikku valikut saab teostada mitmel viisil.

Tõmbeviisi valik. Ühe juhusliku katse tulemus otsustab, millised kogumi objektid valimisse võetakse. Arvutis kasutatakse juhuarvude generaatoril põhinevat valimi moodustamist (*sampling*), mis objektikoodide üldkogumist väljastab etteantud mahuga juhuvalimi. Selline võimalus on olemas kõigis statistikapakettides ja samuti tabelarvutuses. Vajalik on kogumisse kuuluvate objektide loetelu ehk loend.

Loeteluviisi valik. Kogumi iga objekti korral sooritatakse katse, mille tulemus otsustab, kas see objekt võetakse valimisse või mitte. Kasutatakse jälle juhuarvude generaatorit. Kui soovime, et valimisse sattumise tõenäosus oleks iga objekti korral 10%, siis lastakse juhuarvude generaatoril genereerida juhuslik arv lõigul $[0, 1]$. Kui genereeritud arv on lõigul $[0, 0{,}1]$, siis vastav objekt võetakse valimisse. Seda meetodit saab kasutada ka siis, kui loend puudub. Näiteks võime seda kasutada iga poodi astunud isiku korral või tänavaküsitluses.

Tõenäosuslikud valikumeetodid

Järgnevalt on kirjeldatud tuntumaid **tõenäosuslikke valikumeetodeid**.

Lihtne juhuvalik (*random sampling*). Kasutatakse loeteluviisi vali-

[2] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel TT57: Eesti tööjõu-uuringu vastamismäär.

kut, kõigil objektidel on ühesugune valimisse sattumise tõenäosus.
Süstemaatiline valik (*systematic random sampling*). Objektide valik loendist toimub fikseeritud sammuga ning esimene objekt määratakse juhuslikult.
Kihtvalik (*stratisfied random sampling*). Üldkogum jagatakse osadeks ehk kihtideks. Kihtide määramise aluseks on tunnused, mis tõenäoliselt mõjutavad mõõdetavat tunnust (näiteks vanus, sugu). Igas kihis rakendatakse valimi saamiseks mingit tõenäosuslikku valikumeetodit. Kihi valimilt saadakse hinnangud kihi parameetrite jaoks. Kihtide parameetrite põhjal saadakse hinnangud üldkogumi parameetritele.
Klastervalik (*cluster sampling*). Üldkogum koosneb objektigruppidest ehk klastritest. Toimub juhuslik klastrite valik ning iga klastri parameetrid leitakse selle klastri kõigi objektide põhjal. Klastrite parameetrite alusel leitakse üldkogumi parameetrite hinnangud.
Kaheastmeline valik (*two-stage cluster sampling*). Esimesel astmel klastrite juhuslik valik, teisel astmel klastritest objektide juhuslik valik.

Näide 6.1. Noorsoouuring KISS 1999

Noorsoouuring KISS uurib noorte seksuaalset küpsemist ja sellega seonduvaid riske. KISS on lühend sõnadest küpsemine, inimsuhted, sõbrad, seksuaalsus. 1999. aasta uuringu juhuvalimi koostamisel lähtuti eesti ja vene õppekeelega üldhariduskoolide üheksandate klasside loendist. Kasutati kihtvalimit (eesti ja vene koolid), milles tehti klastervalik õpilaste arvu järgi: väikesed, keskmised ja suured klassid. Klastrites sooritati juhuvalik. (Papp, Part ja Tõrik, 2001)

Empiirilise valiku meetodid

Tuntumad **empiirilise valiku** meetodid on järgmised.
Kvootide meetod (*quota sampling*). Antakse ette kvoodid, kui palju objekte tuleb vaadeldavatest tunnusrühmadest valida. Kvoodid määratakse vastavalt üldkogumi struktuurile.
Tasakaalustatud valik (*balanced sampling*). Analoogne kvootide meetodiga, kuid üldkogumi ja valimi võrdlemiseks ei kasutata mitte tausttunnuse väärtuse sagedusi, vaid näiteks keskmisi.
Ekspertvaliku (*expert sampling*) korral on valik täiesti subjektiivne. Näiteks nimetab ekspert kümme tänavat, mis on tema arvates uuritava linna jaoks tüüpilised, ning andmed korjatakse nendelt tänavatelt.
Sobivusvalimi (*availability sampling*) korral valitakse välja objektid, mis mingil põhjusel on sobivad. Näiteks küsitletakse inimesi, keda isiklikult tuntakse, või postitatakse Facebooki viide veebiküsitlusele ja sellele reageerivad vaid postitaja sõbrad. Mõnikord nimetatakse sellist valikuviisi ka **mugavusvalimiks** (*convenience sampling*).

Spontaanne valim (*voluntary sample*) tekib siis, kui inimesed ise vabatahtlikult otsustavad küsitluses osaleda. Tänapäeval on laialt levinud mitmesuguste küsitluste paigutamine veebilehtedele või vaatajate arvamuse registreerimine telesaadetes. Valim moodustub nendest, kes satuvad sellele veebilehele või vaatavad telesaadet ning kes soovivad küsimus(t)ele vastata. Tavaliselt on neil sügavam huvi vastava teema vastu.

Näide 6.2. Välisturistide alkoholi ostumahu uuring ja ekspertvalik

2011. aasta novembrist kuni detsembrini viis TNS Emor läbi uuringu, mille eesmärk oli kaardistada välisturistide ostukäitumist alkohoolsete jookide puhul. Uuringu käigus viidi läbi intervjuud poodide või alkoholiosakondade juhatajatega. Kasutati ekspertvalikut: valimisse valiti spetsialiseerunud turismipiirkonna kauplused ja jaeketid, mida turistid sageli külastavad. (Voog, Sarv ja Männaste, 2012)

Valimi maht

Väga tihti küsitakse, kui suur peab olema valim. Ühest vastust sellele pole. Valimi maht sõltub uuritava tunnuse varieerumisest kogumis.

Tabel 6.2. Valimi suurus mõningate uuringute korral

Uuring	Üldkogum	Üldkogumi maht, tuh	Valimi maht, tuh	Valimi osakaal kogumist, %
Erakondade reiting (Emor, perioodiline)	18-aastased ja vanemad	1 068	1	0,09
Tean ja oskan (ESA[a], 2011)	16–65-aastased isikud	860	13	1,5
Sotsiaaluuring (ESA, 2013)	Leibkonnad	600	7	1,2
Noorsoouuring KISS (1999)	9. kl õpilased	12	1,2	10
EUROSTUDENT IV (Praxis, 2010)	Üliõpilased	65	8	12
Ettevõtete aastastatistika (ESA, 2013)	Ettevõtted	73	11,5	17
Töölepingu seaduse uuring (Praxis, 2013)	Ettevõtted	17	4,4	26

[a] ESA — Eesti Statistikaamet

Olgu meil näiteks kastis kuulid, mille diameeter on ühesugune, s.t varieerumine puudub. Mitu kuuli peame kastist võtma, et saaksime

määrata kuulide diameetri keskväärtust? Piisab ühest kuulist, s.t valimi maht on antud juhul 1. Kui aga kastis on erineva diameetriga kuulid, peab valimi maht olema suurem. Valimi mahu määramist vaatame lähemalt alapeatükis 6.7. Arvestada tuleb ka prognoositava vastamismääraga. Tabelis 6.2 on toodud valimi maht ja selle võrdlus üldkogumi mahuga mõningate uuringute korral.

Valimi esinduslikkus

Valimi koostamisel tuleb arvestada seda, et valim peab olema **esinduslik** ehk **representatiivne**, s.t see peab olema väike mudel üldkogumist ning valimi struktuur erinevate tausttunnuste (inimese sugu, vanusegrupp, ettevõtte tegevusala, ettevõtte suurus jms) suhtes peab olema sama, mis üldkogumis. Olulised on need tausttunnused, mis mõjutavad uuritavaid tunnuseid.

Näide 6.3. Valimiste võitja prognoosimine USA presidendivalimistel

Ajakiri Literary Digest oli USA presidendivalimiste võitjat õigesti prognoosinud alates 1916. aastast. Prognoosimiseks viidi läbi küsitlusi, kus valimi maht oli *ca* 2,4 miljonit inimest. 1936. aasta valimiste võitjaks prognoosis ajakiri vabariiklast Alfred Landonit. Tegelikult võitis valimised demokraat Franklin Roosevelt. Tema võitu oli õigesti prognoosinud George Gallup, kuigi Gallupi kasutatud valim 50 tuhat inimest oli oluliselt väiksem. (Squire, 1988)

Miks siis *Literary Digest* tookord valesti prognoosis? Esmakordselt kasutati uuringu läbiviimiseks telefoniküsitlust. Aga 1936. aastal oli USA-s telefon vaid 25%-l elanikest, jõukamatel. Presidendikandidaat Alfred Landon ei lubanud suuri muutusi ja jõukamad olid sellega rahul. Kuid vaesem elanikkond ei olnud tolleaegse USA majandusega rahul ning pooldasid Roosevelti, kes lubas suuri muutusi. Kuigi Literary Digesti valim oli suur, ei olnud see 1936. aastal esinduslik, ei vastanud USA elanike varanduslikule struktuurile.

Valikunihe

Valikunihe (*sampling bias*) tekib siis, kui valimi moodustamisel on kogumi teatud tüüpi objektidel suurem võimalus valimisse sattuda kui teistel. Näitest 6.3 on näha, et valesti koostatud suur valim annab halvema tulemuse kui õigesti koostatud väike valim. Valimi suurus ei korva väärast koostamisest tingitud nihet. Suur nihkes valim on halvem kui väike nihkes valim, sest see lisab tulemustele võltsi usaldusväärsust.

Lähemalt võib valikuuringute kohta lugeda Tartu Ülikoolis välja antud õpikust „Tõenäosuslik valikuuring“, mille autoriteks on Imbi Traat ja Jaano Inno (1997).

6.2. Punkthinnang ja vahemikhinnang

Statistilise uuringu eesmärk on teha järeldusi üldkogumi kohta. Näiteks: tunnuse X keskväärtus on uuritavas üldkogumis μ. Kui kasutame kõikset uuringut, saame selle keskväärtuse leida. Valikuuringu korral saame seda keskväärtust vaid **hinnata** (*estimate*). Selleks kasutatakse spetsiaalseid hindamisreegleid, mille väljatöötamine on ühe matemaatilise statistika osa — hinnangute teooria — ülesanne.

Hinnata võib uuritava tunnuse erinevaid parameetreid (keskväärtus, dispersioon). Hinnatakse nende parameetrite tegelikke väärtusi s.t nende väärtusi üldkogumis. Parameetri hindamise tulemusena saadakse arv, mis on **punkthinnang**. Seda võib nimetada ka lihtsalt hinnanguks (*estimator*), kuid termin „punkthinnang“ rõhutab erinevust vahemikhinnangust. Vahemikhinnang on vahemik, kuhu üldkogumi väärtus teatud tõenäosusega langeb.

Punkthinnang

Valimi põhjal leitud näitaja on üldkogumi vastava näitaja **punkthinnang** (*point estimate*).

Hinnang leitakse valimi põhjal, aga valim on võetud juhuslikult. Oletame, et soovime leida Eesti elanike keskmist kuupalka. Küsitleme juhuvalimit, kuhu kuulub 1000 inimest. Tulemuseks saame, et nende 1000 inimese keskmine palk on 950 eurot. Seejärel moodustame teise juhuvalimi, kus on samuti 1000 inimest. Kuna valimisse sattumine on juhuslik, siis teise valimisse kuuluvad teised isikud ja nende keskmine palk võib olla 1020 eurot. Kolmanda juhuvalimiga saame uue keskmise. Näitena on tabelis 6.3 toodud viie juhuvalimi keskväärtus üldkogumist, mille keskväärtus on $\mu = 942{,}8$. Kõikide valimite maht on $n = 50$.

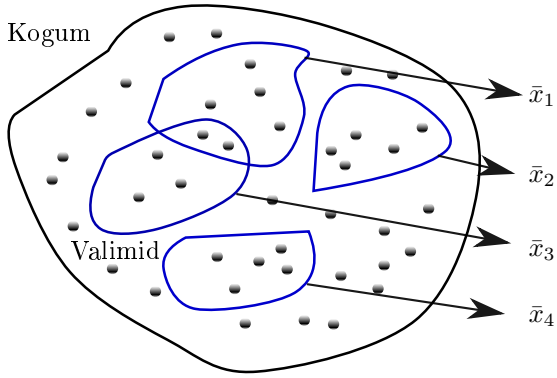
Tabel 6.3. Viie juhuvalimi keskmised, kui üldkogumi keskväärtus $\mu = 942{,}8$. Kõikide valimite maht $n = 50$

Valimi nr	1	2	3	4	5
Valimi keskmine	981,6	938,0	898,8	968,0	935,8

Punkthinnang on juhuslik suurus.

Iga konkreetse valimi põhjal leitud hinnang võib erineda tegelikust väärtusest. Kui teeme palju valimeid, võime leida hinnangute keskväärtuse. Olgu näiteks a hinnatava parameetri väärtus üldkogumis, $\hat{a}$ selle punkthinnang ja $E[\hat{a}]$ hinnangute keskväärtus. **Hinnangu nihe** (*bias*)

Hinnangu nihe



Joonis 6.1. Valimid on võetud juhuslikult, järelikult valimi keskmine on juhuslik suurus, mis võib nelja valimi korral omandada väärtusi $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$ ja $\bar{x}_4$

on hinnangute keskväärtuse ja hinnatava parameetri tegeliku väärtuse vahe:

$$b = E[\hat{a}] - a. \tag{6.1}$$

Kuna ühe konkreetse valimi põhjal leitud üldkogumi parameetri punkthinnang on juhuslik suurus, on otstarbekas kasutada vahemikhinnangut. Tõenäosus, et parameetri väärtus üldkogumis võrdub valimi põhjal saadud punkthinnanguga, on praktiliselt null. Nullist erinev on vahemikku sattumise tõenäosus.

Vahemikhinnang

Vahemikhinnang on valimi põhjal määratud vahemik, mis katab parameetri tegeliku väärtuse etteantud (küllalt suure) tõenäosusega.

Kõige sagedamini kasutatakse vahemikhindamisel **usaldusvahemikku** (*confidence interval*), mille korral on määratud ka vahemikku sattumise tõenäosus ehk **usaldatavus**.

Usaldusvahemik, usaldatavus

Parameetri a **usaldusvahemikuks usaldatavusega** β nimetatakse vahemikku $(\hat{a}-\Delta a; \hat{a}+\Delta a)$, mis katab parameetri väärtuse a tõenäosusega β:

$$\beta = P(|\hat{a} - a| < \Delta a), \tag{6.2}$$

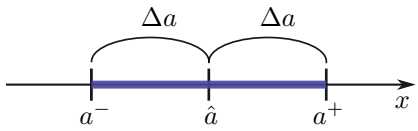
kus $\hat{a}$ on parameetri a punkthinnang.

Usaldusvahemiku otspunktideks on alumine usalduspiir a^- ja ülemine usalduspiir a^+ (vt ka joonis 6.2):

$$a^- = \hat{a} - \Delta a,$$
$$a^+ = \hat{a} + \Delta a,$$

kus $\hat{a}$ on valimi põhjal leitud punkthinnang. Suurust Δa nimetame usaldusvahemiku **poollaiuseks**.

Usaldusvahemiku poollaius



Joonis 6.2. Usalduspiirid a^- ja a^+, usaldusvahemik ja selle poollaius Δa

Mõningatel juhtudel võib usaldusvahemik olla punkthinnangu suhtes ebasümmeetriline (näiteks mediaani usaldusvahemik, vt alapeatükk 6.10). Siis ei saa seda iseloomustada poollaiusega, vaid antakse lihtsalt alumine ja ülemine usalduspiir.

6.3. Üldkogumi keskväärtuse, dispersiooni ja standardhälbe punkthinnangud

Edaspidi tuleb eristada valimi keskmist $\bar{x}$, mida me teame, ning üldkogumi keskväärtust, mida me ei tea. Eristamiseks tähistame üldkogumi keskväärtust tähega μ. Sama probleem on dispersiooni ja standardhälbega. Üldkogumi dispersioon on σ^2 ja standardhälve σ. Valimi dispersioon on s^2 ja standardhälve s (vt ka tabel 6.4).

Olgu meil valim mahuga n ja elementidega x_i. Üldkogumi **keskväärtuse** μ punkthinnanguks on valimi aritmeetiline keskmine ehk lihtsalt **valimi keskmine** (*sample mean*):

Valimi keskmine, dispersioon ja standardhälve

$$\bar{x} = \frac{1}{n}\sum x_i. \tag{6.3}$$

Üldkogumi **dispersiooni** punkthinnang on

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}, \tag{6.4}$$

mida nimetatakse **valimi dispersiooniks** (*sample variance*). Üldkogumi **standardhälbe** punkthinnang on

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}, \tag{6.5}$$

mida nimetatakse **valimi standardhälbeks** (*sample standard deviation*).

Valimi dispersiooni avaldis erineb üldkogumi dispersiooni avaldisest (3.3) nimetaja poolest: nimetajas on $n-1$. Kui me kasutaksime kogumi dispersiooni hindamiseks hälvete ruutude aritmeetilist keskmist

$$\frac{\sum (x_i - \bar{x})^2}{n},$$

saaksime nihkega hinnangu (vt lisa A.7). Nihe tuleneb sellest, et me ei tea üldkogumi keskväärtust μ ja kasutame arvutamisel selle hinnangut ehk valimi keskmist $\bar{x}$. Suurte valimite ($n > 100$) korral on σ^2 ja s^2 erinevus tühine, aga väikeste valimite korral erinevad need oluliselt.

Tabel 6.4. Näitajate tähistus kogumi ja valimi korral

	Maht	Keskmine	Dispersioon	Standardhälve
Üldkogum	N	μ	σ^2	σ
Valim	n	$\bar{x}$	s^2	s

Tabelarvutuses on valimi dispersiooni leidmiseks funktsioon **VAR.S** ja valimi standardhälbe leidmiseks **STDEV.S**.

6.4. Valimi keskmise valimjaotus

Tabelis 6.3 oli toodud viie erineva valimi keskmised. Kuna valimi keskmine on juhuslik suurus, võime rääkida valimite keskmiste jaotusest. Seda nimetatakse **valimi keskmise valimjaotuseks** (*sample distribution of sample mean*), sest see saadakse suure arvu ühesuguse mahuga valimite moodustamisel. Muidugi võib uurida mitte ainult aritmeetilise keskmise, vaid ka näiteks standardhälbe või mediaani valimjaotust.

Valimjaotus

Valimjaotus (*sampling distribution*) on mingi statistiku kui juhusliku suuruse jaotus, mille väärtused saadakse suurest arvust juhuvalimitest mahuga n.

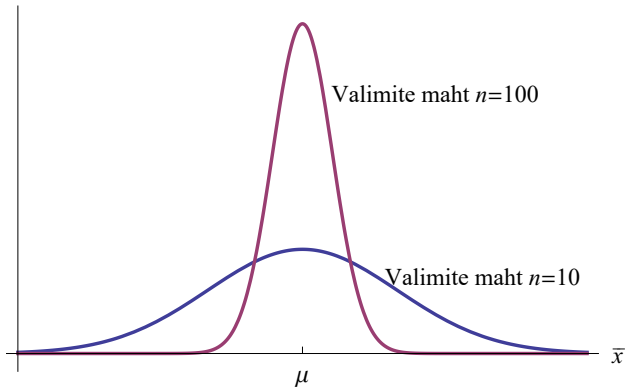
Milline on valimi keskmise valimjaotus? Üldiselt sõltub see valikumeetodist, valimi suurusest ning uuritava tunnuse jaotusest kogumis. Järgnevalt analüüsime valimjaotust, mis tekib lihtsa juhuvaliku korral.

Tsentraalne piirteoreem
Küllalt suure valimi mahu n korral alluvad valimite keskmised $\bar{x}$ normaaljaotusele keskväärtusega μ ja standardhälbega $\sigma/\sqrt{n}$, kus σ on kogumi standardhälve:

$$\bar{x} \sim \mathrm{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right). \tag{6.6}$$

Tsentraalne piirteoreem

Tentraalsest piirteoreemist järeldub, et kui kogumi standardhälve on σ, siis valimi keskmise valimjaotuse standardhälve on $\sigma/\sqrt{n}$. Mida suurem on valimite maht n, seda vähem valimite keskmised hajuvad. Joonisel 6.3 on toodud kaks valimi keskmise valimjaotust ühe ja sama üldkogumi korral. Valimjaotustel on ühesugune keskväärtus, kuid erinev standardhälve, sest valimite mahud on erinevad. Näeme, kui oluline on valimi maht. Väikese valimi korral tõenäosus, et valimi keskmine erineb kogumi keskväärtusest väga palju, võib olla üpris suur.



Joonis 6.3. Valimi keskmise valimjaotused, kui valimite maht on 100 ja 10. Valimid on võetud ühest ja samast kogumist, mille keskväärtus on μ

Pole oluline, millisele jaotusele allub tunnuse väärtus kogumis. Valimite keskmised alluvad alati normaaljaotusele, kui valimid on vaid küllalt suured. Praktikas võib lugeda küllalt suurteks selliseid valimeid, kus valimi maht $n \geq 30$.

Alapeatükis 5.10 nägime, et normaaljaotuse $N(\mu, \sigma)$ korral sõltub vahemikku $\mu \pm k\sigma$ jäämise tõenäosus standardhälbe σ kordajast k (valemid (5.80)–(5.82)). Näiteks 68,3% kõikidest väärtustest jääb keskväärtusest mitte kaugemale kui ühekordne standardhälve (vt ka tabel 5.3). Tsentraalse piirteoreemi järgi alluvad valimite keskmised normaaljaotusele $N(\mu, \sigma/\sqrt{n})$, kus n on valimi maht. Seega, tõenäosusega 68,3%

jääb ühe konkreetse valimi keskmine vahemikku $\mu \pm \sigma/\sqrt{n}$:

$$P\left(\mu - \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + \frac{\sigma}{\sqrt{n}}\right) \approx 0{,}683. \tag{6.7}$$

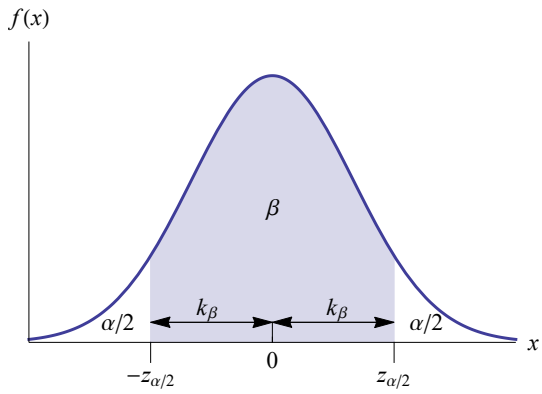
Tõenäosusega 99,7% jääb mingi konkreetse valimi keskmine vahemikku $\mu \pm 3\sigma/\sqrt{n}$:

$$P\left(\mu - 3\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 3\frac{\sigma}{\sqrt{n}}\right) \approx 0{,}997. \tag{6.8}$$

Üldiselt ei pea vahemiku laiust määrav standardhälbe kordaja olema täisarv. Valemid (6.7) ja (6.8) võime üldisemalt kirja panna järgmiselt:

$$P\left(\mu - k_\beta\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + k_\beta\frac{\sigma}{\sqrt{n}}\right) = \beta, \tag{6.9}$$

kus k_β on **tõenäosuskordaja**. Valemis (6.7) $k_\beta = 1$ ja valemis (6.8) $k_\beta = 3$. Sellisel juhul on tõenäosus β ligikaudne arv. Tavaliselt võetakse aga tõenäosus β täpselt ette ning leitakse sellele vastav tõenäosuskordaja k_β.



Joonis 6.4. Tõenäosuskordaja k_β leitakse standardiseeritud normaaljaotusest $N(0,1)$ ja see võrdub täiendkvantiiliga $z_{\alpha/2}$

Normaaljaotuse korral leitakse tõenäosuskordaja väärtus standardiseeritud normaaljaotusest $N(0,1)$. Jooniselt 6.4 on näha, et see võrdub standardiseeritud normaaljaotuse täiendkvantiiliga $z_{\alpha/2}$, kus $\alpha = 1-\beta$.

Keskväärtuse valimjaotuse standardhälbe $\sigma/\sqrt{n}$ leidmisel on probleem selles, et enamasti me ei tea kogumi standardhälvet σ. Valimi põhjal saame leida vaid selle hinnangu s, mis on valimi standardälve (6.5). Sellisel juhul on keskväärtuse valimjaotuse dispersiooni hinnang

$$s^2(\bar{x}) = \frac{s^2}{n} \tag{6.10}$$

ning valimjaotuse standardhälbe hinnang ruutjuur sellest.

Tsentraalse piirteoreemi korral eeldatakse, et kogumi maht N on lõpmata suur. Järelikult valem (6.10) kehtib lõpmatu üldkogumi korral. Praktikas on kogumi maht lõplik suurus[3]. Valimi moodustamisel kogumi maht väheneb iga objekti valimisel, kui me seda objekti tagasi ei pane. Sellist valikumeetodit nimetatakse ka **tagasipanekuta** ehk kordumisteta valikuks. Tagasipanekuta valiku korral kehtib keskväärtuse valimjaotuse dispersiooni jaoks aga valem

$$s^2(\bar{x}) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right). \tag{6.11}$$

Võrreldes valemeid (6.10) ja (6.11), näeme, et kogumi mahu arvestamisel tuleb tagasipanekuta valiku korral dispersiooni (6.10) korrutada avaldisega

$$1 - \frac{n}{N}, \tag{6.12}$$

kus n on valimi maht ja N kogumi maht. Kui kogumi maht on valimi mahuga võrreldes väga suur, siis n/N läheneb nullile, parandusliige (6.12) läheneb ühele ning valem (6.10) valemile (6.11). Järelikult valimi mahuga võrreldes lõpmatult suure kogumi korral võime kasutada valemit (6.10).

Standardviga

Keskväärtuse valimjaotuse standardhälbe hinnang on **standardviga** (*standard error*), mis lõpmatult suure kogumi korral

$$se = \frac{s}{\sqrt{n}}, \tag{6.13}$$

kus s on valimi standardhälve ja n valimi maht.
Lõpliku kogumi korral, mille maht N on teada, on keskväärtuse valimjaotuse standardhälbe hinnang

$$se = \frac{s}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}. \tag{6.14}$$

Nagu eespool nägime, siis see, kas kasutada standardvea leidmiseks valemit (6.13) või korrigeeritud valemit (6.14), sõltub valimi ja kogumi mahtude suhtest n/N. Kui see suhe on väike, siis standardvea parandus on ligikaudu üks ning kasutame valemit (6.13). Mitmed autorid (D. Anderson, Sweeney ja Willams, 1999; McClave, Benson ja Sincich, 2005) soovitavad korrigeeritud valemit (6.14) kasutada siis, kui $n/N > 0{,}05$

[3]Lõpliku kogumi korral kehtib valem (6.10), kui me kasutame tagasipanekuga ehk kordumistega valikut. Praktikas meid tagasipanekuga valik ei huvita: üht ja sama isikut ei intervjueerita mitu korda.

(vt ka tabel 6.5). Kuna parandus on väiksem kui 1, siis korrigeerimata valemi (6.13) kasutamisel suure suhte n/N korral tuleb standardviga suurem, s.t me ülehindame standardviga. Valemit (6.13) tuleb paratamatult kasutada siis, kui me kogumi mahtu N ei tea.

Tabel 6.5. Standardvea paranduse sõltuvus valimi ja kogumi mahtude suhtest

$\frac{n}{N}$	0,5	0,25	0,1	0,05	0,01
$\sqrt{1-\frac{n}{N}}$	0,71	0,87	0,95	0,97	0,99

Toome veel kokkuvõtlikult ära, mis vahe on valimi standardhälbel ja standardveal:

- valimi standardhälve iseloomustab üksikute väärtuste hajumist;
- valimi standardviga iseloomustab valimite keskmiste hajumist.

Valimi standardhälve on null siis, kui hajumine puudub ja kõik väärtused on valimis ühesugused. Valimi standardviga läheneb nullile siis, kui valimi maht suureneb ja läheneb kogumi mahule N. Kui valimi maht saab võrdseks kogumi mahuga, on standardviga (6.14) null.

6.5. Keskväärtuse usalduspiirid suure valimi korral

Eelmises alapeatükis analüüsisime valimi keskmise valimjaotust ning saime teada, kuhu võib valimi keskmine $\bar{x}$ sattuda kogumi keskväärtuse μ suhtes (valem (6.9)). Praktikas on meil aga olukord vastupidine: meil on valimi keskmine $\bar{x}$ teada ja me peame leidma kogumi keskväärtuse usaldusvahemiku. Vaatame uuesti seoses (6.9) esinevat vahemikku valimi keskmise $\bar{x}$ jaoks. Lihtsuse mõttes võtame kasutusele tähistuse $\Delta x = k_\beta(s/\sqrt{n})$:

$$\mu - \Delta x \leq \bar{x} \leq \mu + \Delta x. \tag{6.15}$$

Vasakpoolsest võrratusest saame:

$$\begin{aligned} \mu - \Delta x &\leq \bar{x} \\ \mu &\leq \bar{x} + \Delta x. \end{aligned} \tag{6.16}$$

Avaldise (6.15) parempoolsest võrratusest saame:

$$\begin{aligned} \bar{x} &\leq \mu + \Delta x \\ \bar{x} - \Delta x &\leq \mu. \end{aligned} \tag{6.17}$$

Võrratused (6.16) ja (6.17) võib kokku võtta:

$$\bar{x} - \Delta x \leq \mu \leq \bar{x} + \Delta x. \tag{6.18}$$

Kui võrratused (6.15) määravad ära vahemiku, kuhu langeb valimi keskmine $\bar{x}$, siis avaldises (6.18) on meil kirja pandud vahemik, milles asub kogumi keskväärtus μ. Seega saime **kogumi keskväärtuse vahemikhinnangu**. See ongi meie eesmärk: ühe valimi alusel määrata kogumi keskväärtuse usalduspiirid.

Valemi (6.9) põhjal saame kirja panna selle vahemiku seose usaldatavusega β:

$$P\left(\bar{x} - k_\beta \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + k_\beta \frac{s}{\sqrt{n}}\right) = \beta. \tag{6.19}$$

Keskväärtuse usalduspiirid, suur valim

Suure ($n \geq 30$) valimi korral on **kogumi keskväärtuse usalduspiirid** usaldatavusega β:

$$\bar{x} \pm \Delta x, \tag{6.20}$$

$$\Delta x = z_{\alpha/2} \frac{s}{\sqrt{n}}, \tag{6.21}$$

kus $\bar{x}$ on valimi keskmine, s valimi standardhälve, n valimi maht ning tõenäosuskordaja $z_{\alpha/2}$ on standardiseeritud normaaljaotuse täiendkvantiil, kus vea tõenäosus on $\alpha = 1 - \beta$. Suurust Δx võib nimetada usaldusvahemiku **poollaiuseks**.
Lõpliku kogumi mahu N korral on usaldusvahemiku poollaius

$$\Delta x = z_{\alpha/2} \frac{s}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}. \tag{6.22}$$

Valem (6.21) kehtib lihtsa juhuvaliku korral ning arvestab ainult valimi juhuslikkusest tingitud viga. Muid võimalikke vigu nagu näiteks loendiviga, kaoviga, mõõtmisviga (vt alapeatükk 6.12), arvesse võetud ei ole. Kuidas hinnata viga teistsuguste valikuviiside korral, võib lugeda Imbi Traadi ja Jaan Inno õpikust „Tõenäosuslik valikuuring“.

Kõige sagedamini kasutatatakse usaldatavuse β väärtust 0,95. Mõnikord võetakse ka sellest väiksem 0,9 või suurem 0,99. Standardiseeritud normaaljaotuse täiendkvantiilid $z_{\alpha/2}$ mõningate tõenäosuste korral on toodud tabelis 6.6.

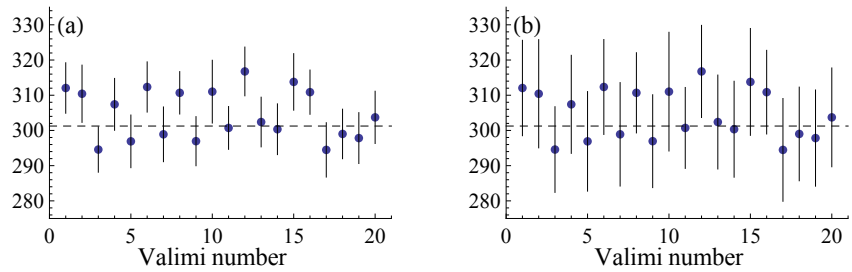
Tabel 6.6. Mõned standardiseeritud normaaljaotuse täiendkvantiilid

β	0,05	0,15	0,3	0,5	0,75	0,9	0,95	0,99
$\alpha/2$	0,475	0,425	0,35	0,25	0,125	0,05	0,025	0,005
$z_{\alpha/2}$	0,06	0,19	0,39	0,67	1,15	1,64	1,96	2,58

Suurema usaldatavuse korral

- on tõenäosus, et kogumi keskväärtus langeb usalduspiiridesse, suurem;
- on usaldusvahemik laiem.

Jooniselt 6.5 on näha, kuidas usaldatavuse suurenedes muutuvad usaldusvahemikud laiemaks. Seepärast ei olegi mõistlik kasutada väga suurt usaldatavust. Mis kasu on meil näiteks teadmisest, et Eesti elanike keskmine brutopalk jääb vahemikku 100 kuni 3000 eurot, kuigi selle tulemuse usaldatavus on väga kõrge.



Joonis 6.5. Kahekümne valimi ($n = 50$) usaldusvahemikud usaldatavuse 0,7 (a) ja 0,95 (b) korral. Punktid näitavad valimite keskmisi ning vertikaalsed lõigud usaldusvahemikke. Horisontaalne kriipsjoon vastab kogumi keskväärtusele. Valimid on joonistel (a) ja (b) samad, kuid usaldusvahemike arvutamisel on võetud erinev usaldatavus. Usaldatavuse 0,7 korral on vea tõenäosus $\alpha = 30\%$, mis 20 valimi puhul on 6. Joonisel (a) on tegelik valimite arv, mille usaldusvahemikesse kogumi keskväärtus ei jää, 8. Usaldatavuse 0,95 korral on vea tõenäosus $\alpha = 5\%$, mis 20 valimi puhul on 1. Joonisel (b) on tegelik valimite arv, millega määratud usaldusvahemikesse kogumi keskväärtus ei jää, 1 (valim nr 12)

Standardiseeritud normaaljaotuse täiendkvantiili leidmiseks tabelarvutuses tuleb kasutada funktsiooni **NORM.S.INV(*probability*)**, kus *probability* $= \alpha/2$. See funktsioon väljastab väärtuse $-z_{\alpha/2}$ (vt joonis 6.4). Positiivne väärtus on selle vastandväärtus. Mõned näited:

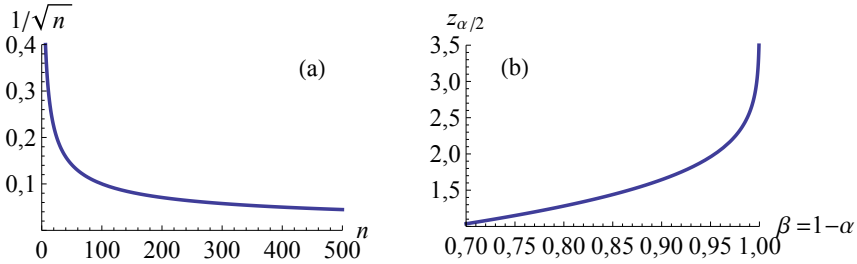
β	$\alpha/2$	$z_{\alpha/2}$
0,9	$(1-0{,}9)/2 = 0{,}05$	$-$NORM.S.INV(0,05) $\approx 1{,}64$
0,95	$(1-0{,}95)/2 = 0{,}025$	$-$NORM.S.INV(0,025) $\approx 1{,}96$

On olemas ka tabelarvutusfunktsioon usaldusvahemiku poollaiuse Δx leidmiseks, mis arvutab valemi (6.21) järgi, ja see on **CONFIDENCE.NORM**. Siis tuleb ette anda vea tõenäosus $\alpha = 1 - \beta$ (*Alpha*), valimi standardhälve s (*Standard_ dev*) ja valimi maht n (*Size*).

Valemist (6.21) on näha, et usaldusvahemiku poollaius Δx sõltub usaldatavusega β määratud tõenäosuskordajast $z_{\alpha/2}$, valimi standardhälbest s ja valimi mahust n. Valimi standardhälve sõltub sellest, kuidas uuritav tunnus kogumis hajub ja seda me mõjutada ei saa. Kui me

ettevõetud usaldatavuse korral soovime vähendada usaldusvahemiku poollaiust ehk määramatust, siis ainuke võimalus on suurendada valimi mahtu n. Kui soovime poollaiust Δx näiteks 10 korda vähendada, tuleb valimi mahtu $10^2 = 100$ korda suurendada. See aga suurendab valikvaatluse kulusid. Joonisel 6.6 (a) on esitatud kordaja $1/\sqrt{n}$ sõltuvus valimi mahust n. Näeme, et väikeste valimite korral väheneb usaldusvahemiku laius n kasvades kiiresti. Valimi mahu suurendamine 400-st 500-ni aga enam nii suurt efekti ei anna.

Joonis 6.6 (b) kirjeldab tõenäosuskordaja sõltuvust usaldatavusest. Kui usaldatavus β hakkab lähenema ühele, siis tõenäosuskordaja $z_{\alpha/2}$ väärtus ja koos sellega usaldusvahemiku laius kasvavad järsult.



Joonis 6.6. (a) Kordaja $1/\sqrt{n}$ sõltuvus valimi mahust n. (b) Tõenäosuskordaja $z_{\alpha/2}$ sõltuvus usaldatavusest β

Näide 6.4. Kulud toidule leibkonna liikme kohta


N06Valikvaatlused
N6.4

2012. aasta leibkonna eelarve uuringus küsitleti 9080 isikut. Ühes küsimuses paluti hinnata toiduainetele ja mittealkohoolsetele jookidele tehtavaid kulutusi aastas leibkonnaliikme kohta. Vastuste keskväärtus oli 915,3 eurot standardhälbega 541,72 eurot (*Leibkonna eelarve uuring* 2012). Kui suured on keskmised kulud toiduainetele ja mittealkohoolsetele jookidele aastas leibkonnaliikme kohta?

Paneme kirja lähteandmed:

valimi maht $n = 9080$;

valimi keskmine $\bar{x} = 915{,}30$;

valimi standardhälve $s = 541{,}72$.

Usaldatavuseks võtame $\beta = 0{,}95$, sellele vastav tõenäosuskordaja tabelist 6.6 (või tabelarvutuses $-$NORM.S.INV(0,025)) on $z_{0{,}025} = 1{,}96$. Valemist (6.21) leiame usaldusvahemiku poollaiuse:

$$\Delta x = 1{,}96 \cdot \frac{541{,}72}{\sqrt{9080}} \approx 11{,}1.$$

Vastus: usaldatavusega 0,95 oli keskmine kulu toiduainetele ja mittealkohoolsetele jookidele 915,3±11,1 eurot aastas pereliikme kohta. Alumist ja ülemist usalduspiiri kasutades võime kulude usaldusvahemiku anda kujul (904,2, 926,4) eurot.

Keskväärtuse usalduspiire on võimalik leida ka intervallitud variatsioonridade korral, kuid siis tuleb valimi keskmise leidmiseks kasutada kaalutud aritmeetilist keskmist (2.6). Valimi dispersiooni valemi saame kaalutud dispersiooni valemist (3.4), kui nimetajas võtame $\sum f_i$ asemel vabadusastmete arvu $n-1=\sum f_i-1$:

$$s^2=\frac{\sum(x_i-\bar{x})^2 f_1}{\sum f_i-1}. \tag{6.23}$$

Valimi standardhälve on nagu ikka ruutjuur valimi dispersioonist.

N06Valikvaatlused
N6.5

Näide 6.5. Noorte säästmisharjumused USA-s

2006. aastal viidi USA-s läbi uuring teismeliste noorte finantskäitumise kohta. 39 küsimusest koosnevale küsitlusele vastas 243 noort vanuses 14 kuni 19 aastat. Küsimusele, kui palju raha on säästetud, olid vastusevariantidena antud vahemikud. Vastuste jaotus on toodud tabelis. (Koonce jt, 2008)

Summa dollarites	Vastanute arv
≤ 100	75
101–200	30
201–300	28
301–400	17
401–500	13
501–1000	29
>1000	51

Leiame keskmise säästetud summa usaldusvahemiku. Selleks tuleb algul leida valimi keskmine ja standardhälve. Kuna andmed on esitatud sagedustabelina, tuleb kasutada kaalutud aritmeetilise keskmise ja dispersiooni valemeid. Kaalutud aritmeetilise keskmise leidmiseks lisame tabelisse veerud „Klassi keskmine x_i“ ja „Korrutised $f_i x_i$“. Dispersiooni leidmiseks lisame veerud „$(x_i-\bar{x})^2$“ ja „Korrutised $f_i(x_i-\bar{x})^2$“, mis saab täita peale keskmise $\bar{x}$ leidmist. Klasside keskmiste leidmisel tuletame meelde, et avatud klasside korral võetakse klassi laiuseks kõrvaloleva klassi laius.

Summa dollarites	Klassi keskmine x_i	Vastanute arv f_i	Korrutised $f_i x_i$	$(x_i - \bar{x})^2$	Korrutised $f_i(x_i - \bar{x})^2$
≤ 100	50,0	75	3 750,0	170 994,2	12 824 562
101–200	150,5	30	4 515,0	97 978,0	2 939 341
201–300	250,5	28	7 014,0	45 375,1	1 270 504
301–400	350,5	17	5 958,5	12 772,3	217 128
401–500	450,5	13	5 856,5	169,4	2 202
501–1000	750,5	29	21 764,5	82 360,7	2 388 461
> 1000	1 250,5	51	63 775,5	619 346,3	31 586 663
KOKKU		243	112 634		51 228 861

Vastavalt valemile (2.6) on aritmeetiline keskmine

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{112634}{243} = 463{,}51.$$

Dispersiooni leidmiseks arvestame sellega, et tegemist on valikvaatlusega ja me soovime leida kogumi dispersiooni hinnangut, järelikult valemis (3.4) peab nimetajas olema $n - 1 = \sum f_i - 1$:

$$s^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i - 1} = \frac{51228861}{243 - 1} = 211689{,}5.$$

Valimi standardhälve on siis

$$s = \sqrt{s^2} = \sqrt{211689{,}5} \approx 460{,}1$$

ja usaldusvahemiku poollaius

$$\Delta x = 1{,}96 \cdot \frac{s}{\sqrt{n}} = 1{,}96 \cdot \frac{460{,}1}{\sqrt{243}} \approx 57{,}8.$$

Vastus: usaldatavusega 0,95 on USA teismelised keskmiselt säästnud $463{,}5 \pm 57{,}8$ dollarit.
Omaette probleem on see, kas keskväärtus on siin sobiv statistik, iseloomustamaks USA noorte säästmisharjumusi. Kui me vaatame erineva vahemiku valinud vastajate arvusid f_i, siis näeme, et moodklassiks on „≤ 100“. Seega, kõige sagedamini säästetakse kuni 100 dollarit.

Kummal juhul on keskväärtuse jaoks saadud tulemus täpsem, kas näites 6.4 saadud tulemus $915{,}3 \pm 11{,}1$ eurot aastas pereliikme kohta või näites 6.5 saadud tulemus USA noorte säästmise kohta $463{,}5 \pm 57{,}8$ dollarit? Ilmselt ei saa me võrrelda usaldusvahemikke, sest tegemist on erinevate suurustega. Küll aga võime mõlemal juhul leida suhte

$\Delta x/\bar{x}$ ja võrrelda neid suhteid. Eesti elanike kulud toidule on määratud täpsusega

$$\frac{11{,}1}{915{,}3} \approx 0{,}0112 = 1{,}12\%.$$

USA noorte säästmine on leitud täpsusega

$$\frac{57{,}8}{463{,}5} \approx 0{,}125 = 12{,}5\%.$$

Esimene tulemus on oluliselt täpsem. Selle üheks põhjuseks on oluliselt suurem valimi maht.

Suhteline viga

Usaldusvahemiku määramise täpsust näitab **suhteline viga**

$$E = \frac{\Delta x}{\bar{x}}, \tag{6.24}$$

kus Δx on usaldusvahemiku poollaius ning $\bar{x}$ valimi keskmine.

6.6. Keskväärtuse usalduspiirid väikese valimi korral

t-jaotus

Väikeste valimite korral erineb valimi keskmise valimjaotus normaaljaotusest ja tsentraalne piirteoreem ei kehti. Sellisel juhul kasutatakse kogumi keskväärtuse usalduspiiride määramisel ***t*-jaotust** ehk Studenti jaotust. Jaotuse võttis kasutusele inglise matemaatik William Seally Gosset (1876–1937) oma töös, mille ta avaldas varjunime all Student.

Väikeste valimite korral alluvad valimite keskmiste standardiseeritud väärtused t-jaotusele

$$\frac{\bar{x} - \mu}{se} \sim t(\nu), \tag{6.25}$$

kus μ on üldkogumi keskväärtus, se valimi standardviga ning ν vabadusastmete arv:

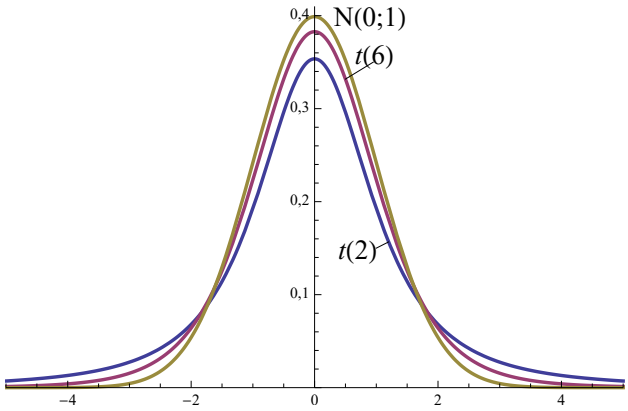
$$\nu = n - 1, \tag{6.26}$$

kus n on valimi maht.

Vabadusastmete arv

Üldiselt on **vabadusastmete arv** sõltumatute muutujate arv (vt ka lisa A.6). Kui me arvutame valimi standardhälvet ja leiame hälbed valimi keskmisest, on vabadusastmete arv ühe võrra väiksem valimi mahust n (vt lisa A.7).

Valimi mahu ja seega ka vabadusastmete arvu ν suurenedes läheneb $t(\nu)$-jaotus standardiseeritud normaaljaotusele $N(0, 1)$ (joonis 6.7). Seepärast võibki suurte valimite korral kasutada normaaljaotust.



Joonis 6.7. Kaks t-jaotuse kõverat vabadusastmete arvuga 2 ja 6 ning võrdluseks standardiseeritud normaaljaotuse $N(0,1)$ kõver

Väikese ($n < 30$) valimi korral on **kogumi keskväärtuse usaldusvahemik** usaldatavusega β:

$$\bar{x} \pm \Delta x, \tag{6.27}$$

$$\Delta x = t_{\alpha/2}(\nu)\frac{s}{\sqrt{n}}, \tag{6.28}$$

kus $\bar{x}$ on valimi keskmine, s valimi standardhälve, n valimi maht, $\nu = n-1$ vabadusastmete arv ning tõenäosuskordaja $t_{\alpha/2}(\nu)$ on $t(\nu)$-jaotuse täiendkvantiil, kus vea tõenäosus $\alpha = 1-\beta$.

Keskväärtuse usaldusvahemik, väike valim

t-jaotuse täiendkvantiili $t_{\alpha/2}(\nu)$ nimetatakse mõnikord ka Studenti koefitsiendiks ja selle väärtused on toodud lisas B.1.

t-jaotuse täiendkvantiili leidmiseks tabelarvutuses saab kasutada funktsiooni **T.INV**, kus *Probability* $= \alpha/2$ ja *Deg_freedom* $= \nu$. See funktsioon väljastab väärtuse $-t_{\alpha/2}(\nu)$ (nii nagu ka standardiseeritud normaaljaotuse korral, vt joonis 6.4). Positiivne väärtus on selle vastandarv. Lihtsam võimalus on kasutada funktsiooni **T.INV.2T**, kus *Probability* $= \alpha$. Need kaks funktsiooni on omavahel seotud järgmiselt: T.INV.2T(α; ν) = – T.INV($\alpha/2$; ν).

Usaldusvahemiku poollaiuse Δx leidmiseks on kõige mugavam kasutada funktsiooni **CONFIDENCE.T**, mis arvutab valemi (6.28) järgi. Siis tuleb ette anda vea tõenäosus $\alpha = 1-\beta$ (*Alpha*), valimi standardhälve s (*Standard_ dev*) ja valimi maht n (*Size*).

N06Valikvaatlused
N6.6

Näide 6.6. Kauba läbimüük

Kauba A nädalane läbimüük viies juhuslikult valitud kesklinna poes oli 16, 82, 29, 31 ja 55 tk. Leida selle kauba keskmine nädala läbimüük kesklinna poodides usaldatavusega 0,9 ja 0,95.
Paneme kirja andmed:
valimi maht $n = 5$;
valimi keskmine $\bar{x} = 42{,}6$;
valimi standardhälve $s = 26{,}14$.
Tõenäosuskordaja vabadusastmete arvu 4 ja usaldatavuse 0,9 korral $t_{0{,}05}(4) = 2{,}13$ ning usaldatavuse 0,95 korral on $t_{0{,}025}(4) = 2{,}78$ (tabelarvutuses T.INV.2T(0,1;4) ja T.INV.2T(0,05;4) või lisa B.1).
Usaldusvahemiku poollaius, kui usaldatavus on 0,9:

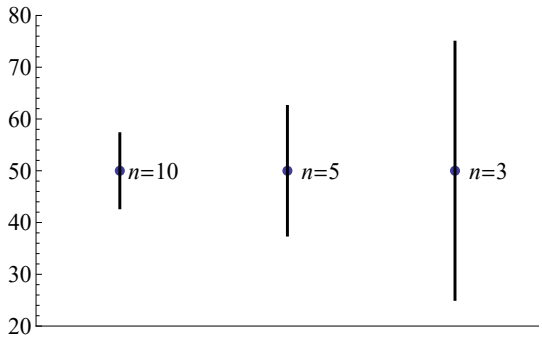
$$\Delta x = t_{\alpha/2}(\nu)\frac{s}{\sqrt{n}} = 2{,}13 \cdot \frac{26{,}14}{\sqrt{5}} \approx 24{,}9$$

ja usaldatavuse 0,95 korral

$$\Delta x = t_{\alpha/2}(\nu)\frac{s}{\sqrt{n}} = 2{,}78 \cdot \frac{26{,}14}{\sqrt{5}} \approx 32{,}5.$$

Vastus: usaldatavusega 0,9 on antud kauba keskmine läbimüük nädalas 42,6±24,9 tk ja usaldatavusega 0,95 on läbimüük 42,6± 32,5 tk nädalas.

Valimi mahu vähenedes suureneb usaldusvahemiku laius ja seega ka määramatus (joonis 6.8). See on tingitud kahest põhjusest. Esiteks väheneb valemi (6.28) nimetajas olev maht n, mille tulemusena murd suureneb. Teiseks suureneb tõenäosuskordaja $t_{\alpha/2}(\nu)$, sest väheneb vabadusastmete arv (vt joonis 6.7).



Joonis 6.8. Erineva mahuga n valimite korral leitud usaldusvahemikud. Valimite keskmised ja standardhälbed on ühesugused

Kuna suurte valimite korral läheneb t-jaotus standardiseeritud normaaljaotusele, siis täiendkvantiil $t_{\alpha/2}(\nu)$ läheneb normaaljaotuse täiendkvantiilile $z_{\alpha/2}$ (vt tabel 6.7). Seepärast võib ka suurte valimite korral kasutada t-jaotusest leitavat tõenäosuskordajat $t_{\alpha/2}(\nu)$.

Kasutades üldist tõenäosuskordaja tähistust k_β, võib kogumi keskväärtuse usaldusvahemiku poollaiuse valemi panna kirja kujul

$$\Delta x = k_\beta \frac{s}{\sqrt{n}}, \tag{6.29}$$

kus k_β on

- suure valimi ($n > 30$) korral standardiseeritud normaaljaotuse täiendkvantiil $z_{\alpha/2}$;
- väikese valimi korral $t(\nu)$-jaotuse täiendkvantiil $t_{\alpha/2}(\nu)$.

Tabel 6.7. $t(\nu)$-jaotuse täiendkvantiili võrdlus standardiseeritud normaaljaotuse täiendkvantiiliga 1,96, kui usaldatavus on 0,95. Viimases veerus on t-jaotuse täiendkvantiili suhteline erinevus standardiseeritud normaaljaotuse täiendkvantiilist

Valimi maht n	$t(\nu)$	Suhteline erinevus normaaljaotusest, %
3	4,30	120
4	3,18	62
5	2,78	42
10	2,26	15
20	2,09	7
30	2,05	4
50	2,01	3
100	1,98	1,2
200	1,97	0,6
500	1,965	0,2

6.7. Valimi mahu planeerimine

Hinnangu täpsuse määrab usaldusvahemiku laius. Fikseeritud usaldatavuse korral on usaldusvahemiku laius määratud valimi mahuga ja valimit moodustades on võimalik vahemiku laiust ette planeerida.

Näide 6.7. Keskmine läbimüük ja valimi mahu planeerimine

Näites 6.6 leiti viie juhuslikult väljavalitud poe põhjal keskmise läbimüügi usaldusvahemik. Usaldatavusega 0,95 oli see

$42{,}6 \pm 32{,}5$ tk nädalas. Valimi standardhälve oli $s = 26{,}14$ ja vastav tõenäosuskordaja 2,78.
Usaldusvahemiku poollaius 32,5 on küllalt suur ja moodustab 76,2% valimi keskmisest. Kui palju poode tuleks valimisse võtta, kui me tahame, et hinnang oleks täpsem ja usaldusvahemiku poollaius oleks väiksem kui 40% valimi keskmisest?
Usaldusvahemiku poollaiuse jaoks kehtib siis tingimus

$$\Delta x \leq 0{,}4 \cdot 42{,}6 = 17{,}04.$$

Valemi (6.28) põhjal võime kirjutada:

$$t_{\alpha/2}(\nu)\frac{s}{\sqrt{n}} \leq 17{,}04$$
$$t_{\alpha/2}(\nu)\frac{s}{17{,}04} \leq \sqrt{n}$$
$$\left(t_{\alpha/2}(\nu)\frac{s}{17{,}04}\right)^2 \leq n.$$

Paneme leitud avaldisse valimi standardhälbe s väärtuse ja usaldatavusele 0,95 vastava tõenäosuskordaja $t_{0{,}025}(4) = 2{,}78$:

$$n \geq \left(2{,}78 \cdot \frac{26{,}14}{17{,}04}\right)^2$$
$$n \geq 18{,}14.$$

Järelikult peab valimi maht olema 19. Valimi mahtu tuleb ümardada alati ülespoole, sest allapoole ümardades ei ole võrratus täidetud.
Kui tahame aga, et usaldusvahemiku laius oleks alla 20%, saame vajaliku valimi mahuks 73.

Valimi mahu planeerimine

Valimi mahu planeerimisel tuleb algul teha uuritavast üldkogumist **proovivalim**. Olgu proovivalimi maht n_0 ja standardhälve s_0. Kui soovime, et usaldusvahemiku poollaiuse jaoks kehtiks tingimus

$$\Delta x \leq d, \tag{6.30}$$

saame vajaliku valimi mahu n määrata tingimusest

$$n \geq \left(t_{\alpha/2}(\nu)\,\frac{s_0}{d}\right)^2. \tag{6.31}$$

Kuna aga tõenäosuskordaja $t_{\alpha/2}(\nu)$ väheneb vaatluste arvu n suurenemisel, toimub valimi mahu mõningane ülehindamine. Kui vaatluste

sooritamine on väga kallis, ei tohiks ülehindamist lubada ja siis kasutatakse mitmesammulist protseduuri, kusjuures valimi suurendamine toimub objektide lisamise teel eelmisele valimile.

Suurte valimite korral võetakse valemis (6.31) tõenäosuskordajaks standardiseeritud normaaljaotuse täiendkvantiil $z_{\alpha/2}$:

$$n \geq \left(z_{\alpha/2} \, \frac{s_0}{d} \right)^2 . \tag{6.32}$$

Mõnikord võib juhtuda, et planeeritud mahuga valimi põhjal leitud usalduspiirid on siiski laiemad ning soovitud tingimus (6.30) pole täidetud. See juhtub siis, kui planeeritud mahuga valimi standardhälve tuleb märgatavalt suurem proovivalimi standardhälbest s_0. Näiteks sattusid valimisse mõned ekstreemsed väärtused, mida väikeses proovivalimis ei esinenud.

6.8. Kaheväärtuselise tunnuse osakaalu usalduspiirid

Alapeatükis 3.9 leidsime valemi kaheväärtuselise tunnuse standardhälbe jaoks (3.26). Sellest lähtudes saame leida ühe väärtuse osakaalu usalduspiirid (*confidence interval for proportion*) kaheväärtuselise tunnuse korral.

Olgu kogumis positiivsete tulemuste osakaal p. Selle osakaalu punkthinnang on

$$\hat{p} = \frac{m}{n}, \tag{6.33}$$

kus n on valimi maht ja m positiivsete tulemuste arv valimis.

Kaheväärtuselise tunnuse osakaalu usaldusvahemik, suur valim

Suure valimi korral on kaheväärtuselise tunnuse osakaalu p usaldatavusega β määratud usaldusvahemik

$$\hat{p} \pm \Delta p, \tag{6.34}$$

$$\Delta p = z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \tag{6.35}$$

kus $\hat{p}$ on ühe väärtuse osakaal valimis, n valimi maht ning tõenäosuskordaja $z_{\alpha/2}$ on standardiseeritud normaaljaotuse täiendkvantiil, nii et $\alpha = 1 - \beta$.

Lõpliku kogumi mahu N korral on osakaalu usaldusvahemiku poollaius

$$\Delta p = z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \sqrt{1 - \frac{n}{N}}. \tag{6.36}$$

Nii nagu keskväärtuse usalduspiiride leidmisel, kui valimi maht on kogumi mahuga võrreldes suur, $n/N > 0{,}05$, tuleb kasutada korrigeeritud valemit (6.36))(vt ka tabel 6.5).

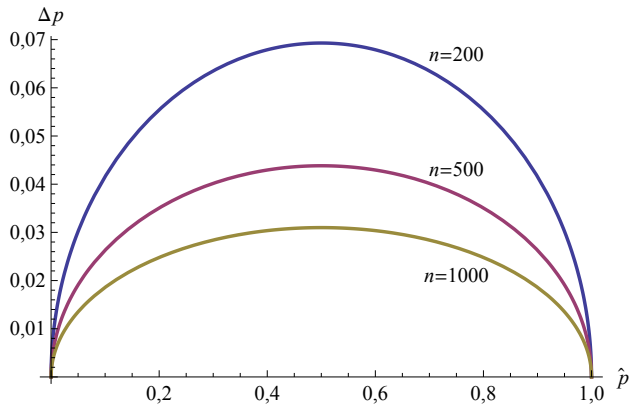
Näide 6.8. Valimistest osavõtt

2003. aasta riigikogu valimiste eel viidi Eestis läbi mitmesuguseid küsitlusi. Ühes Emori uurimuses küsitleti üle Eesti 500 valimisealist kodanikku, valima lubas minna neist 51% (Ideon, 2003). Kui suur protsent kõigist valimisealistest kodanikest pidi selle uuringu järgi valima minema?
Teeme vastavad arvutused. Valimi maht $n = 500$, valima minejate osakaal valimis $\hat{p} = 0{,}51$. Võttes usaldatavuseks $\beta = 0{,}95$, saame tõenäosuskordaja $z_{0{,}025} = 1{,}96$. Usaldusvahemiku poollaius

$$\Delta p = 1{,}96 \cdot \sqrt{\frac{0{,}51(1-0{,}51)}{500}} \approx 0{,}044.$$

Vastus: usaldatavusega 0,95 pidi selle uuringu põhjal valima minema 51,0% ± 4,4% ehk 46,6% kuni 54,4% kõigist valimisealistest kodanikest. Vabariigi valimiskomisjoni andmetel osales valimistel tegelikult 58% valimisõiguslikest kodanikest.

Valemist (6.35) võib leida, kuidas, usaldusvahemiku poollaius Δp sõltub osakaalust valimis fikseeritud valimi mahu n korral (joonis 6.9). Kõige laiem on usaldusvahemik osakaalu $\hat{p} = 0{,}5$ korral, sest siis on mittehomogeensus ehk hajuvus kõige suurem ja järelikult ka standardhälve kõige suurem. Matemaatiliselt näidati seda alapeatükis 3.9.



Joonis 6.9. Osakaalu usaldusvahemiku poollaiuse Δp sõltuvus vastava väärtuse osakaalust valimis $\hat{p}$ valimi mahu n erinevate väärtuste korral

Analoogselt alapeatükis 6.7 vaadeldud valimi mahu planeerimisega üldkogumi keskväärtuse hindamisel saab ka osakaalu hindamisel tule-

Valimi mahu planeerimine

tada tingimuse valimi mahu jaoks. Kui tahame, et osakaalu usaldusvahemiku poollaius Δp ei ületaks teatud väärtust D, s.t

$$\Delta p \leq D, \tag{6.37}$$

siis valemist (6.35) saame tingimuse valimi mahu jaoks:

$$n \geq z^2_{\alpha/2} \frac{\hat{p}_0(1-\hat{p}_0)}{D^2}, \tag{6.38}$$

kus $\hat{p}_0$ on eelvalimi põhjal hinnatud osakaal.

Valem (6.36) annab õiged usalduspiirid siis, kui valimi maht $n > 30$, kuid arvestada tuleb ka osakaalu $\hat{p}$. Kui osakaal on väga lähedal nullile või ühele (näiteks 0,001 või 0,999), peab valimi maht olema õigete usalduspiiride saamiseks väga suur. Kirjanduses soovitatakse enamasti väikese osakaalu korral tingimust $n\hat{p} \geq 5$ ning suure osakaalu korral $n(1-\hat{p}) \geq 5$ (L. D. Brown, Cai ja DasGupta, 2001). Arvestades seost (6.33), on $n\hat{p}$ võrdne positiivsete tulemuste arvuga m. Järelikult valemi (6.36) kasutamiseks peab valimis nii üht kui teist väärtust omavate elementide arv olema ≥ 5.

Suure valimi tingimus

Kaheväärtuselise tunnuse osakaalu usalduspiiride leidmiseks võib valemit (6.36) kasutada juhul, kui kumbagi väärtust omavate elementide arv on valimis ≥ 5, s.t kehtib tingimus

$$n \cdot \min(\hat{p}, 1-\hat{p}) \geq 5. \tag{6.39}$$

Kui tingimus 6.39 pole täidetud, ei ole osakaalude valimjaotus normaaljaotus ning usalduspiiride leidmiseks tuleb kasutada teistsugust lähenemist. Erinevad autorid on soovitanud mitmeid lähenemisviise, sagedasti leiab kasutamist Agresti-Coulli korrigeeritud intervall (Agresti ja Coull, 1998). Seda nimetatakse mõnikord ka „nelja lisamise reegliks“ (*„plus four“ method*), sest korrigeeritud osakaalu $\tilde{p}$ leidmiseks lisatakse neli tulemust, millest kaks on positiivsed:

$$m \to m+2,$$
$$n \to n+4.$$

Nende asenduste tegemisel valemis (6.35) saadakse osakaalu jaoks korrigeeritud usalduspiirid.

Korrigeeritud usalduspiirid

Korrigeeritud usalduspiirid osakaalu jaoks:

$$\tilde{p} \pm \Delta\tilde{p},$$
$$\Delta\tilde{p} = z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}, \tag{6.40}$$

kus valimi korrigeeritud osakaal

$$\tilde{p} = \frac{m+2}{n+4} \tag{6.41}$$

ja m on positiivsete tulemuste arv valimis.

Näide 6.9. Pettusealaste riskide uuring

2014. aasta juunis viis audiitorfirma Ernst & Young Eestis läbi pettusealaste riskide uuringu. Uuringu eesmärgiks oli saada teavet pettuste esinemise, juhtumitega kaasneva kahju ja riskide juhtimiseks rakendatavate meetmete kohta Eesti äriühingutes ja avaliku sektori asutustes. Veebipõhisele küsitlusele vastas 112 juhtimisfunktsiooniga isikut erinevatest organisatsioonidest. Küsimuse „Kas järgnev tegevus tundub Teile õigustatud, et majandussurutises toime tulla?“ vastusevariandi „Rahalised maksed (majandus)tegevuse jätkamiseks ja elavdamiseks“ korral valis kaks vastajat vastusevariandi „Jah“. (*Pettuseriskide alane uuring Eestis* 2014)
Leiame, kui suur osa juhtidest arvab, et rahalised maksed (majandus)tegevuse jätkamiseks ja elavdamiseks on majandussurutise korral õigustatud. Korrigeeritud osakaal valemist (6.41)

$$\tilde{p} = \frac{2+2}{112+4} = \frac{4}{116} \approx 0{,}0345.$$

Usaldusvahemiku poollaius usaldatavuse 0,95 korral valemist (6.40)

$$\Delta\tilde{p} = 1{,}96\sqrt{\frac{0{,}0345(1-0{,}0345)}{112+4}} \approx 0{,}0332.$$

Vastus: seda, et rahalised maksed (majandus)tegevuse jätkamiseks ja elavdamiseks on majandussurutise korral õigustatud, arvab $3{,}45\% \pm 3{,}32\%$ juhtidest.

6.9. Kolme ja enama väärtusega kvalitatiivse tunnuse osakaalude usalduspiirid

Küsitlustes kasutatakse tihti valikvastustega küsimusi, kus on rohkem kui kaks vastusevarianti. Siis ei ole tegemist kaheväärtuselise tunnusega ning vastuste osakaalude usalduspiiride leidmisel ei ole korrektne kasutada eelmises alapeatükis toodud valemeid.

Kvalitatiivne tunnus on kas nimiskaalas või järjestusskaalas mõõdetud tunnus. Nimiskaalas on näiteks järgmised tunnused:

- isiku sotsiaal-majanduslik seisund:
  - töötav,
  - töötu,
  - õpilane,
  - pensionär;
- eluruumi tüüp:
  - eramaja,
  - mitmepereelamu,
  - korter.

Järjestusskaalas on näiteks

- isiku haridustase:
  1) I taseme haridus (alghariduseta, alghariduse, põhihariduse);
  2) II taseme haridus (keskharidus, kutseõpe põhihariduse baasil);
  3) III taseme haridus (kutseõpe keskhariduse baasil, kõrgharidus, magister, doktor);
- toimetulek:
  1) suurte raskustega;
  2) mõningate raskustega;
  3) tuleb toime.

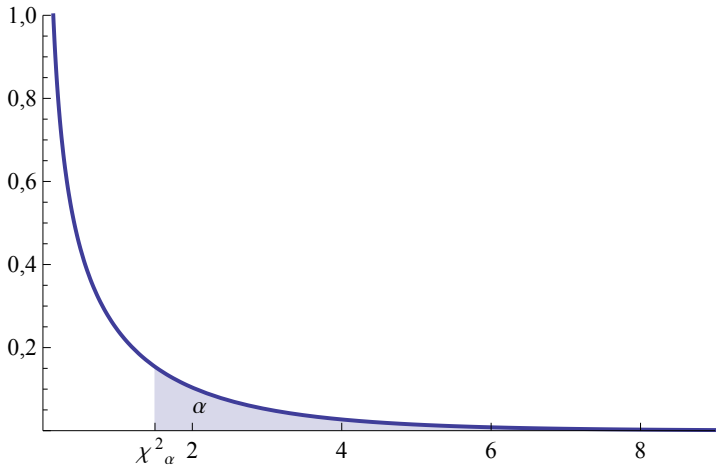
Kui valimis on mingi vastusevariandi esinemise osakaal $\hat{p}_i$, siis kuidas nüüd leida selle osakaalu usalduspiire üldkogumi jaoks? Erinevate autorite poolt on pakutud mitmeid meetodeid, millest mõned on arvutuslikult üpris mahukad (Fitzpatrick ja A. Scott, 1987; Sison ja Glaz, 1995; Wang, 2008). Toome siin ära ühe lihtsama meetodi, mille tuletas Leo Goodman (1965).

Eeldatakse, et valikvastuste hulgast saab valida **ühe ja ainult ühe** vastusevariandi, s.t peab kehtima tingimus

$$\sum_{i=1}^{k} \hat{p}_i = 1, \tag{6.42}$$

kus k on valikute arv ja $\hat{p}_i$ on i-nda valiku osakaal valimis. Järgnevad valemid ei sobi kasutamiseks juhul, kui vastaja võib valida mitu vastusevarianti.

Olgu meil kvalitatiivne tunnus, millel on k väärtust $m_1, m_2, \ldots, m_k$. Vastavate väärtuste osakaalud valimis on $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k$ ja kehtib tingimus (6.42). Siis i-nda väärtuse osakaalu $\hat{p}_i$ usaldusvahemiku poollaiuse leidmisel tuleb valemis (6.35) standardiseeritud normaaljaotuse täiendkvantiili $z_{\alpha/2}$ asemel kasutada ühe vabadusastmega **χ^2-jaotuse** täiendkvantiili. Ühe vabadusastmega χ^2-jaotuse jaotustiheduse graafik on toodud joonisel 6.10 (vt ka lisa A.5).



Joonis 6.10. Ühe vabadusastmega χ^2-jaotuse jaotustiheduse graafik ja α järku täiendkvantiil χ^2_α

Kvalitatiivse tunnuse osakaalu usaldusvahemik suure valimi korral

Kui usaldatavuseks on β ja vea tõenäosus $\alpha = 1 - \beta$, siis suure valimi korral on i-nda osakaalu p_i usaldusvahemiku poollaius

$$\Delta p_i = \sqrt{\chi^2_{\alpha/k}(1)}\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n}}, \tag{6.43}$$

kus $\chi^2_{\alpha/k}(1)$ on ühe vabadusastmega χ^2-jaotuse α/k järku täiendkvantiil, $\hat{p}_i$ on i-nda väärtuse osakaal valimis ning n valimi maht.

Valemist (6.43) leitud usalduspiiride korral üksiku osakaalu p_i jaoks tõenäosus, et usalduspiirid ei kata tegelikku väärtust, on α/k, kus k on väärtuste arv. Tõenäosus, et usalduspiirid ei kata tegelikku väärtust kas p_1 või p_2 või ... p_k korral, on α. Usaldatavus β on siis tõenäosus,

et **kõik** k usaldusvahemikku katavad vastavate osakaalude tegelikud väärtused.

Tabelarvutuses on χ^2-jaotuse täiendkvantiili leidmiseks funktsioon **CHISQ.INV.RT**. Usaldatavust β omava usaldusvahemiku jaoks vajaliku täiendkvantiili leidmisel on argument *Probability* $(1-\beta)/k$ ning vabadusastmete arv *Deg_freedom* on 1. Näiteks kui esitatud küsimusel on 4 vastusevarianti ning soovime leida osakaalude usalduspiire usaldatavusega $\beta = 0{,}95$, siis valemi (6.43) kasutamiseks tuleb leida

$$\chi^2_{\alpha/k}(1) = \chi^2_{0{,}05/4}(1) = \chi^2_{0{,}0125}(1) = \quad \text{CHISQ.INV.RT}(0{,}0125;1).$$

Näide 6.10. Hinnang majanduslikule olukorrale

N06Valikvaatlused
N6.10

2012. aasta leibkonna eelarve uuringus paluti hinnata, milline võiks olla vastaja leibkonna majanduslik olukord ühe aasta pärast. Kokku oli vastajaid 9080 ning vastused jagunesid järgmiselt (*Leibkonna eelarve uuring* 2012):

		Sagedus n_i	Osakaal $\hat{p}_i$
1	Märksa parem	277	3,05%
2	Mõnevõrra parem	1847	20,34%
3	Üldiselt samasugune	4421	48,69%
4	Mõnevõrra kehvem	1801	19,83%
5	Märksa kehvem	734	8,08%

Leiame osakaalude usalduspiirid usaldatavusega 0,95, s.t $\alpha = 0{,}05$. Vastusevariantide arv $k = 5$ ning $\alpha/k = 0{,}05/5 = 0{,}01$. χ^2-jaotuse täiendkvantiili $\chi^2_{0{,}01}(1)$ leiame tabelarvutuses: $\text{CHISQ.INV.RT}(0{,}01;1) = 6{,}6349$.
Vastusevariandi 1 jaoks saame valemist (6.43)

$$\Delta p_1 = \sqrt{6{,}6349} \cdot \sqrt{\frac{0{,}0305(1-0{,}0305)}{980}} \approx 0{,}0046.$$

Järelikult nende osakaal, kes aastal 2012 arvasid, et aasta pärast on nende majanduslik olukord märksa parem, oli $3{,}05\% \pm 0{,}46\%$ ehk vahemikus (2,59%; 3,52%). Niimoodi leiame usalduspiirid kõikide vastusevariantide jaoks.

		Osakaal $\hat{p}_i$	Δp_i	Alumine piir	Ülemine piir
1	Märksa parem	3,05%	0,46%	2,59%	3,52%
2	Mõnevõrra parem	20,34%	1,09%	19,25%	21,43%
3	Üldiselt samasugune	48,69%	1,35%	47,34%	50,04%
4	Mõnevõrra kehvem	19,83%	1,08%	18,76%	20,91%
5	Märksa kehvem	8,08%	0,74%	7,35%	8,82%

Sageli kasutatakse rohkem kui kahe väärtusega kvalitatiivse tunnuse osakaalude usalduspiiride leidmiseks siiski valemit (6.35), mis kehtib kaheväärtuselise tunnuse korral. Kuna valemid (6.43) ja (6.35) erinevad ainult tõenäosuskordaja poolest, siis kahel erineval meetodil leitud usalduspiiride võrdlemiseks piisab nende kordajate võrdlemisest. Tabelis 6.8 on võrreldud vastavaid tõenäosuskordajaid usaldatavuse 0,9 ja 0,95 korral. Esimeses veerus on kvalitatiivse tunnuse erinevate väärtuste arv k. Kui $k = 2$, on tõenäosuskordajaks $z_{\alpha/2}$, ja kui $k = 3, 4, \ldots$, on tõenäosuskordaja $\sqrt{\chi^2_{\alpha/k}(1)}$. Näeme, et normaaljaotuse täiendkvantiili $z_{\alpha/2}$ kasutamisel $k = 3, 4, \ldots$ korral on usalduspiirid kitsamad ja seega alahinnatud.

Tabel 6.8. Usaldusvahemiku tõenäosuskordajad kaheväärtuselise tunnuse ($z_{\alpha/2}$) ja rohkem kui kahe väärtusega kvalitatiivse tunnuse korral $\left(\sqrt{\chi^2_{\alpha/k}(1)}\right)$

k	$\beta = 0{,}9$	$\beta = 0{,}95$	
2	1,64	1,96	$z_{\alpha/2}$
3	2,13	2,39	
4	2,24	2,50	
5	2,33	2,58	
6	2,39	2,64	$\sqrt{\chi^2_{\alpha/k}(1)}$
7	2,45	2,69	
8	2,50	2,73	
9	2,54	2,77	
10	2,58	2,81	

Valemit (6.43) võib usalduspiiride leidmiseks kasutada suure valimi korral, kui valimis on i-nda väärtuse esinemissagedus $n_i > 100$. Nende väärtuste jaoks, mille korral see tingimus pole täidetud, tuleb usaldusvahemiku leidmiseks kasutada täpsemat, kuid keerulisemat valemit. Sellisel juhul on osakaalude $p_1, p_2, \ldots, p_k$ usaldusvahemikud

$$p_i^- \leq p_i \leq p_i^+, \tag{6.44}$$

kus alumine ja ülemine piir (Goodman, 1965)

$$p_i^- = \frac{\chi^2 + 2n_i - \sqrt{D}}{2(n + \chi^2)}, \tag{6.45}$$

$$p_i^+ = \frac{\chi^2 + 2n_i + \sqrt{D}}{2(n + \chi^2)} \tag{6.46}$$

ja suurus D leitakse valemist

$$D = \chi^2 \left(\chi^2 + 4(n - n_i)\frac{n_i}{n}\right). \tag{6.47}$$

Siin on χ^2 ühe vabadusastmega χ^2-jaotuse täiendkvantiil $\chi^2_{\alpha/k}(1)$, n_i on i-nda väärtuse esinemissagedus valimis ja n valimi maht. Valemid (6.45)–(6.47) on teatud ruutvõrrandi lahendid.

Näide 6.11. Varimajandus Bulgaarias

N06Valikvaatlused
N6.11

Aastatel 2007–2008 viidi Bulgaaria ettevõtete hulgas läbi uuring varimajanduse kohta, valimis oli 345 ettevõtet. Küsimusele, kui suure osa SKP-st moodustab Bulgaarias varimajandus, vastas 343 ettevõtet järgnevalt (Goev, 2009):

Vastusevariant	Vastajate arv	Osakaal
Varimajandus puudub	1	0,3%
Kuni 10%	4	1,2%
10% kuni 25%	69	20,1%
25% kuni 50%	178	51,9%
50% kuni 75%	74	21,6%
Rohkem kui 75%	17	5,0%

Kuna enamiku vastusevariantide korral on vastajate arv väiksem kui 100, kasutame osakaalude usalduspiiride leidmiseks täpseid valemeid (6.45)–(6.47). Võtame usaldatavuseks 0,95. Vastusevariantide arv $k = 6$ ning $\alpha/k = 0{,}05/6$. χ^2-jaotuse täiendkvantiili $\chi^2_{0{,}05/6}(1)$ leiame vastava tabelarvutusfunktsiooni abil: CHISQ.INV.RT$(0{,}05/6; 1) = 6{,}9604$. Arvutusteks kasutame tabelarvutust ja võrdluseks leiame usalduspiirid ka ligikaudse valemi (6.43) põhjal.

		Täpne valem		Ligikaudne valem	
Vastusevariant	Vastajate arv	Alumine piir	Ülemine piir	Alumine piir	Ülemine piir
Varimajandus puudub	1	0,03%	2,53%	−0,48%	1,06%
Kuni 10%	4	0,34%	3,94%	−0,36%	2,70%
10% kuni 25%	69	15,03%	26,40%	14,41%	25,83%
25% kuni 50%	178	44,81%	58,90%	44,78%	59,01%
50% kuni 75%	74	16,31%	27,97%	15,71%	27,43%
Rohkem kui 75%	17	2,66%	9,04%	1,86%	8,05%

Nagu eelmises näites toodud tabelist on näha, võib väikese vastajate arvu korral ligikaudsest valemist leitud usaldusvahemiku alumine piir olla negatiivne, kuid osakaal ei saa olla negatiivne. Negatiivsete väärtuste vältimiseks tulebki kasutada täpset valemit. Kui mingi vastusevariandi valijaid on rohkem kui 100, langevad täpse ja ligikaudse

valemi põhjal leitud usaldusvahemiku piirid ligikaudu kokku ning siis võib kasutada lihtsamat ligikaudset valemit (6.43).

6.10. Mediaani usalduspiirid

Kui juhusliku suuruse X jaotus üldkogumis on sümmeetriline, siis langevad keskväärtus ja mediaan kokku. Mediaani punkthinnanguks on sel juhul soovitatav kasutada valimi aritmeetilist keskmist, sest selle valimjaotuse hajumine on väiksem kui valimi mediaani valimjaotusel. Usalduspiirideks on siis keskväärtuse usalduspiirid.

Kui juhusliku suuruse jaotus üldkogumis ei ole sümmeetriline, siis valimi aritmeetiline keskmine ei sobi üldkogumi mediaani Me hinnanguks. Toome siin ära nn Thompsoni-Savuri protseduuri (Savur, 1937; W. R. Thompson, 1936), kuidas hinnata üldkogumi mediaani usalduspiire, kui juhusliku suuruse jaotus üldkogumis on suvaline jaotus.

Olgu meil valim mahuga n. Arvestades mediaani definitsiooni, võime valimi iga elemendi x_i jaoks kirjutada, et

$$P(x_i < Me) = 0{,}5;$$
$$P(x_i > Me) = 0{,}5.$$

Üldkogumi mediaanist väiksema väärtusega elementide arv valimis on n^- ja mediaanist suurema väärtusega elementide arv on n^+. Nii n^- kui ka n^+ alluvad ühesugusele binoomjaotusele $B(n, 0{,}5)$.

Olgu Me^- ja Me^+ vastavalt mediaani usaldusvahemiku alumine ning ülemine piir. Alumiseks piiriks Me^- võib olla kasvavalt järjestatud valimi esimene element x_1 või teine element x_2 jne. Ülemiseks piiriks võib olla järjestatud valimi viimane element x_n või eelviimane element x_{n-1} jne. Kui alumiseks usalduspiiriks on kasvavalt järjestatud valimi

- **esimene** element, siis tõenäosus, et üldkogumi mediaan on sellest väiksem, võrdub tõenäosusega, et valimis pole ühtegi üldkogumi mediaanist väiksemat elementi:

$$P(Me < Me^-) = P(n^- = 0) = P(n^- < 1);$$

- **teine** element, siis tõenäosus, et üldkogumi mediaan on sellest väiksem, võrdub tõenäosusega, et valimis pole ühtegi või on täpselt üks üldkogumi mediaanist väiksem element:

$$P(Me < Me^-) = P(n^- = 0) + P(n^- = 1) = P(n^- < 2).$$

Analoogselt, kui ülemiseks usalduspiiriks on kasvavalt järjestatud valimi

- **viimane** element, siis tõenäosus, et üldkogumi mediaan on sellest suurem, võrdub tõenäosusega, et valimis pole ühtegi üldkogumi mediaanist suuremat elementi:

$$P(Me > Me^+) = P(n^+ = 0) = P(n^- < 1);$$

- **eelviimane** element, siis tõenäosus, et üldkogumi mediaan on sellest suurem, võrdub tõenäosusega, et valimis pole ühtegi või on üks üldkogumi mediaanist suurem element:

$$P(Me > Me^+) = P(n^+ = 0) + P(n^+ = 1) = P(n^+ < 2).$$

See loogika on kokkuvõtlikult esitatud tabelis 6.9. Teises veerus on tõenäosused $P(Me < Me^-)$, kui alumiseks usalduspiiriks on valitud $\{x_1, x_2, \ldots, x_k\}$. Tabeli viimases veerus on tõenäosused $P(Me > Me^+)$, kui ülemiseks usalduspiiriks on valitud $\{x_n, x_{n-1}, \ldots, x_{n-k+1}\}$.

Tabel 6.9. Mediaani usalduspiiride valik ning piiridest välja jäämise tõenäosused. x_i on kasvavalt järjestatud valimi elemendid, n^- on kogumi mediaanist väiksemate valimi elementide arv ja n^+ mediaanist suuremate elementide arv

Me^-	$P(Me < Me^-)$	Me^+	$P(Me > Me^+)$
x_1	$P(n^- < 1)$	x_n	$P(n^+ < 1)$
x_2	$P(n^- < 2)$	x_{n-1}	$P(n^+ < 2)$
x_3	$P(n^- < 3)$	x_{n-2}	$P(n^+ < 3)$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_k	$P(n^- < k)$	x_{n-k+1}	$P(n^+ < k)$
$\ldots$	$\ldots$	$\ldots$	$\ldots$

Kuna n^- ja n^+ alluvad samale binoomjaotusele, siis mingi k korral

$$P(n^- < k) = P(n^+ < k). \tag{6.48}$$

Kui soovime leida mediaani usaldusvahemikku (Me^-, Me^+) usaldatavusega 0,95, siis vahemikust välja jäämise tõenäosus on $1 - 0{,}95 = 0{,}05$ ning

$$P(Me < Me^-) = 0{,}025, \;\; P(Me > Me^+) = 0{,}025. \tag{6.49}$$

Nüüd tuleb binoomjaotusest $B(n, 0{,}5)$ leida maksimaalne k väärtus, mille korral kehtib tingimus $P(n^- < k) < 0{,}25$. Arvestame sellega, et

$$P(n^- < k) = \sum_{m=0}^{k-1} P(n^- = m),$$

kus tõenäosused $P(n^- = m)$ leitakse Bernoulli valemist (5.49).

Tabelarvutuses saab k väärtuse leida funktsiooni BINOM.INV abil:

$$k = \text{BINOM.INV}(n; 0{,}5; \alpha/2),$$

kus n on valimi maht ja $\alpha = 1-\beta$. Vastava järjenumbriga elemendi leidmiseks tabelarvutuses tuleb kasutada funktsiooni **SMALL**$(Array;K)$, kus $Array$ on valimi andmete piirkond ning K otsitava elemendi järjenumber.

Mediaani usaldusvahemiku ülemisele piirile vastava elemendi järjenumber järjestatud valimis on (vt tabel 6.9)

$$n - k + 1. \tag{6.50}$$

Kui näiteks valimi maht $n = 10$, siis tabelist 6.10 näeme, et $P(n^- < 2) \approx 0{,}011$ ja $P(n^- < 3) \approx 0{,}055$, mis juba ületab väärtuse 0,025. Järelikult tuleb valida $k = 2$ ning mediaani alumine usalduspiir on kasvavalt järjestatud valimi teine element.

Tabel 6.10. Binoomjaotusest $B(n, 0{,}5)$ leitud tõenäosused, kui valimi maht $n = 10$ (tõenäosused on ümardatud)

k	$P(n^- < k)$
1	0,001
2	0,011
3	0,055
...	...

Mediaani ülemise usalduspiiri järjenumber

$$n - k + 1 = 10 - 2 + 1 = 9.$$

Järelikult on mediaani ülemine usalduspiir järjestatud valimi üheksas element. Näeme, et tingimuste (6.49) täpset täitmist pole võimalik tagada. Seepärast pole leitud vahemiku usaldatavus täpselt 0,95, vaid on sellest suurem:

$$\beta = 1 - 2 \cdot 0{,}011 = 0{,}978.$$

Olgu meil järjestatud valim, milles on 10 elementi {20, 27, 30, 35, 39, 43, 49, 55, 63, 70}. Mediaani alumiseks usalduspiiriks on teine element 27 ja ülemiseks usalduspiiriks üheksas element 63. Järelikult on mediaani usaldusvahemik $(27, 63)$ ja selle usaldatavus $\beta \approx 0{,}978$.

Tabelis 6.11 on esitatud mediaani usalduspiiride järjenumbrid mõningate erineva mahuga valimite jaoks. Tabeli viimases veerus on binoomjaotusest leitud vahemikku langemise tõenäosus ehk tegelik usaldatavus

$$\beta = P(x_k \le Me \le x_{n-k+1}) = 1 - \left(P(Me < x_k) + P(Me > x_{n-k+1})\right).$$

Tabel 6.11. Mediaani usalduspiiride järjenumbrid järjestatud valimis

N06Valikvaatlused T6.11

Valimi maht n	Me^- järjenumber k	Me^+ järjenumber $n-k+1$	Usaldatavus β
6	1	6	0,969
7	1	7	0,984
8	1	8	0,992
9	2	8	0,961
10	2	9	0,979
11	2	10	0,988
12	3	10	0,961
13	3	11	0,978

Arvestades võrdust (6.48), on tõenäosused $P(Me < x_k)$ ja $P(Me > x_{n-k+1})$ võrdsed, ning usaldatavus

$$\beta = 1 - 2 \cdot P(Me < x_k) = 1 - 2 \cdot P(n^- < k). \qquad (6.51)$$

Tõenäosuse $P(n^- < k)$ saab leida binoomjaotusest $B(n, 0{,}5)$.

Mediaani usalduspiiride leidmiseks tuleb kasutada niisiis järgmist protseduuri.

Mediaani usalduspiiride leidmine

1. Võtta ette usaldatavus β.
2. Kasutades tabelarvutusfunktsiooni BINOM.INV, leida alumise usalduspiiri järjenumber kasvavalt järjestatud valimis:

   $$k = \text{BINOM.INV}(n; 0{,}5; \alpha/2),$$

   kus n on valimi maht ja $\alpha = 1 - \beta$.
3. Leida ülemise usalduspiiri järjenumber $n - k + 1$.
4. Mediaani usalduspiiride Me^- ja Me^+ leidmiseks leida kasvavalt järjestatud valimist elemendid järjenumbritega k ja $n - k + 1$. Tabelarvutuses kasutada funktsiooni SMALL:

   $$Me^- = \text{SMALL}(Array; k),$$
   $$Me^+ = \text{SMALL}(Array; n - k + 1),$$

   kus *Array* on valimi andmete piirkond (andmed ei pea olema sorteeritud).
5. Kui valim on väike, siis tuleb leida usaldusvahemiku tegelik usaldatavus. Suurte valimite korral binoomjaotusest leitud tegelik usaldatavus läheneb ette võetud usaldatavusele. Tabelarvutuses funktsiooni BINOM.DIST kasutamisel tuleb arvestada sellega, et $\text{BINOM.DIST}(k; n; 0{,}5; 1) = P(X \le k)$, meie vajame aga tõenäosust $P(X < k)$. Tegelik usaldatavus valemist (6.51)

   $$\beta = P(Me^- \le Me \le Me^+) = 1 - 2 \cdot P(n^- < k) =$$
   $$= 1 - 2 \cdot \text{BINOM.DIST}(k - 1; n; 0{,}5; 1).$$

Mediaani punkthinnangut see protseduur leida ei võimalda, sest üldiselt sõltub mediaani punkthinnangu leidmise metoodika juhusliku suuruse jaotusest üldkogumis.

Näide 6.12. Kulud toidule ja mediaani usalduspiirid

N06Valikvaatlused
N6.12

Näites 6.4 leidsime leibkonnauuringu põhjal, kui palju kulutatakse keskmiselt toidukaupadele ja mittealkohoolsetele jookidele pereliikme kohta aastas. Keskväärtuse usalduspiirideks saime $915{,}3 \pm 11{,}1$ eurot aastas pereliikme kohta usaldatavusega 0,95. Leiame nüüd mediaani usalduspiirid. Valimi maht $n = 9080$.

1. Võtame usaldatavuseks 0,95.
2. Mediaani alumise usalduspiiri järjenumber

$$k = \text{BINOM.INV}(9080; 0{,}5; 0{,}05/2) = 4447.$$

3. Mediaani ülemise usalduspiiri järjenumber

$$n - k + 1 = 9080 - 4447 + 1 = 4634.$$

4. Vastavate järjenumbritega elemendid valimis

$$Me^- = \text{SMALL}(Array; 4447) = 814{,}82,$$
$$Me^+ = \text{SMALL}(Array; 4634) = 838{,}83,$$

   kus *Array* on valimi andmete piirkond.
5. Kuigi valim on suur, kasutame binoomjaotust tegeliku usaldatavuse leidmiseks:

$$\beta = 1 - 2 \cdot \text{BINOM.DIST}(4447 - 1; 9080; 0{,}5; 1) \approx 0{,}95.$$

Vastus: toidukaupadele ja mittealkohoolsetele jookidele tehtavate kulutuste mediaan jääb usaldatavusega 0,95 vahemikku (814,82, 838,83) eurot pereliikme kohta aastas. Kuna keskväärtus on mediaanist suurem, siis kulude jaotus on parempoolse asümmeetriaga: on üksikuid peresid, kelle kulud on ekstreemselt suured.

Suured valimid

Suurte valimite korral läheneb binoomjaotus normaaljaotusele ning siis kasutatakse alumise usalduspiiri järjenumbri leidmiseks ka valemit

$$k = 0{,}5\left(n - z_{\alpha/2}\sqrt{n}\right), \tag{6.52}$$

kus $z_{\alpha/2}$ on normaaljaotuse täiendkvantiil ja n valimi maht. Kuna järjenumber peab olema täisarv, ümardatakse arvutuse (6.52) tulemus täisarvuni. Ülemise usalduspiiri järjenumber leitakse valemist (6.50).

Lähtudes valemist (6.52), võib mediaani usaldusvahemiku piirid suure valimi korral anda ka valimi protsentiilide kaudu. Alumine piir on valimi protsentiil järguga

$$0{,}5\left(1-\frac{z_{\alpha/2}}{\sqrt{n}}\right) \tag{6.53}$$

ja ülemine piir on valimi protsentiil järguga

$$0{,}5\left(1+\frac{z_{\alpha/2}}{\sqrt{n}}\right)+\frac{1}{n}. \tag{6.54}$$

Arvestades, et usaldatavuse $\beta = 0{,}95$ korral $z_{\alpha/2} = 1{,}96$, saame leida mediaani usaldusvahemiku piirid valimi protsentiili järkude kaudu erineva mahuga valimite korral. Need on esitatud tabelis 6.12. Konkreetse valimi korral tuleb valimist leida protsentiilijärkudele vastavad protsentiilid. Näiteks kui valimi maht on 500, siis mediaani usaldusvahemiku alumine piir on valimi element, mis vastab protsentiilile 45,6%, ja ülemine piir protsentiilile 54,6% vastav element.

Tabel 6.12. Mediaani usaldusvahemiku piirid valimi protsentiilides

Valimi maht	Valimi protsentiili järk (ümardatud)	
	Alumine piir	Ülemine piir
100	40,2%	60,8%
200	43,1%	57,4%
500	45,6%	54,6%
1000	46,9%	53,2%
10000	49,0%	51,0%

6.11. Valimi kaalumine

Valimi kaalumist kasutatakse valimi esinduslikkuse parandamiseks. Kui valimis on mõne olulise tausttunnuse (inimese sugu, elukoht, vanus, ettevõtte suurus, tegevusvaldkond) proportsioonid oluliselt erinevad üldkogumis esinevatest proportsioonidest, võib valimi põhjal tehtud hinnang tulla nihkega. Selle vältimiseks viiakse läbi valimi kaalumine.

1. Arvestades tausttunnuse väärtuste esinemise osakaalusid üldkogumis, kaalutakse valim ümber, nii et kaalutud valimis on erinevate väärtuste osakaalud samad, mis kogumis.
2. Igale valimis esinevale objektile omistatakse vastav kaal sõltuvalt tema tausttunnuse väärtusest.
3. Valimi aritmeetilise keskmise leidmisel kasutatakse kaalutud aritmeetilist keskmist.

N06Valikvaatlused
N6.13

Näide 6.13. Ülikooli vilistlaste küsitlus ning valimi kaalumine

Ühe ülikooli vilistlaste hulgas viidi läbi küsitlus. Ülikoolis on kolm teaduskonda ning vastanute hulgas oli kõigi kolme teaduskonna vilistlasi. Kuid erinevate teaduskondade vilistlaste osakaalud valimis ei vastanud nende osakaaludele üldkogumis (kõigi vilistlaste hulgas). Seepärast tuli läbi viia valimi kaalumine.

1. etapp: kaalude leidmine.

Näiteks teaduskonna A vilistlasi on üldkogumis 50%, kuid valimis ainult 30%. Leiame kaalu, millega tuleb arvu 0,3 korrutada, et saada arv 0,5:

$$\frac{0{,}5}{0{,}3} \approx 1{,}667.$$

Niimoodi leitakse kaalud ka ülejäänud teaduskondade jaoks.

Teaduskond	Lõpetanute osakaal üldkogumis	Osakaal valimis	Kaalud	Korrigeeritud osakaalud valimis
A	50%	30%	1,667	50%
B	20%	50%	0,400	20%
C	30%	20%	1,500	30%

2. etapp: kaalude omistamine.

Lihtsuse mõttes vaatame valimit mahuga 10. Igale isikule omistatakse kaal vastavalt sellele, mis teaduskonna vilistlane ta on. Kaalude summa on kokku 10.

Isik	Teaduskond	Kaal
1	A	1,667
2	A	1,667
3	A	1,667
4	B	0,400
5	B	0,400
6	B	0,400
7	B	0,400
8	B	0,400
9	C	1,500
10	C	1,500

3. etapp: kaalutud aritmeetilise keskmise leidmine.

Oletame, et üheks küsimuseks oli vilistlase brutokuupalga suurus. Keskmise kuupalga leidmiseks kasutatakse kaalutud aritmeetilist keskmist. Selleks leitakse viimasesse veergu korrutised $f_i x_i$.

Isik	Teaduskond	Kuupalk x_i, eurot	Kaal f_i	$f_i x_i$
1	A	760	1,667	1266,7
2	A	800	1,667	1333,3
3	A	930	1,667	1550,0
4	B	850	0,400	340,0
5	B	1 200	0,400	480,0
6	B	1 100	0,400	440,0
7	B	950	0,400	380,0
8	B	800	0,400	320,0
9	C	1 050	1,500	1575,0
10	C	890	1,500	1335,0

Viimase veeru summa on 9020. Kaalutud aritmeetiline keskmine

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{9020}{10} = 902.$$

Vastus: valimi keskmine on 902 eurot. Kui kasutada kaalumata valimit, saaksime keskmiseks 933 eurot. Seda sellepärast, et teaduskonna B lõpetajaid, kelle kuupalk on üldiselt suurem, on valimis palju. Kuid nende osakaal kõigi vilistlaste hulgas on väike.

Kui valim vajab kaalumist mitme erineva tausttunnuse järgi, siis objektidele kaalude omistamisel korrutatakse omavahel erinevate tausttunnuste järgi saadud kaalud.

Näide 6.14. Valimi kaalumine kahe tausttunnuse järgi

N06Valikvaatlused
N6.14

Jätkame näites 6.13 toodud valimiga. Olgu teiseks tausttunnuseks sugu, mille osakaalud valimis samuti ei lange kokku osakaaludega üldkogumis.
1. Leiame soole vastavad kaalud.

Sugu	Lõpetanute osakaal üldkogumis	Osakaal valimis	Kaalud	Korrigeeritud osakaalud valimis
M	50%	40%	1,250	50%
N	50%	60%	0,833	50%

2. Nüüd omistatakse igale isikule kaal, mis on kahe kaalu korrutis: $f_i = f_{1i} f_{2i}$.

Isik	Teaduskond	Sugu	Kaal teaduskonna järgi f_{1i}	Kaal soo järgi f_{2i}	Kaalude korrutis $f_i = f_{1i}f_{2i}$
1	A	M	1,667	1,250	2,083
2	A	M	1,667	1,250	2,083
3	A	N	1,667	0,833	1,389
4	B	N	0,400	0,833	0,333
5	B	N	0,400	0,833	0,333
6	B	N	0,400	0,833	0,333
7	B	M	0,400	1,250	0,500
8	B	N	0,400	0,833	0,333
9	C	M	1,500	1,250	1,875
10	C	N	1,500	0,833	1,250

3. Kuupalga kaalutud aritmeetilise keskmise leidmisel kasutatakse kaaluna kahe kaalu korrutist.

Isik	Teaduskond	Sugu	Kuupalk x_i, eurot	Kaal f_i	$f_i x_i$
1	A	M	760	2,083	1583,33
2	A	M	800	2,083	1666,67
3	A	N	930	1,389	1291,67
4	B	N	850	0,333	283,33
5	B	N	1 200	0,333	400,00
6	B	N	1 100	0,333	366,67
7	B	M	950	0,500	475,00
8	B	N	800	0,333	266,67
9	C	M	1 050	1,875	1968,75
10	C	N	890	1,250	1112,50

Kuupalga kaalutud aritmeetiline keskmine

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{9414{,}58}{10{,}514} \approx 895{,}44.$$

Vastus: valimi keskmine peale kaalumist kahe tausttunnuse järgi on 895,44 eurot.

6.12. Vea komponendid

Olgu $\hat{a}$ valimi põhjal arvutatud hinnang üldkogumi parameetrile a. Vahet

$$u = \hat{a} - a$$

nimetatakse hinnangu $\hat{a}$ **veaks**. Vea suurust ei ole võimalik leida, sest parameetri tegelik väärtus a on teadmata. Teatud juhtudel on aga võimalik hinnata vea ülempiiri.

Praktilistes valikuuringutes mõjutavad hinnangu $\hat{a}$ viga valimi võtmise tõenäosusliku protseduuri kõrval ka teised faktorid. Tinglikult võib eristada järgmisi veakomponente:

- valikuviga;
- loendiviga;
- kaoviga;
- objektide asendamise viga;
- mõõtmisviga;
- töötlusviga.

Järgnevalt kirjeldame neid veakomponente lähemalt.

Valikuviga (*sampling error*) esineb alati hinnangutes, mis arvutatakse valimi põhjal ja seda tingib valikudisainist (valimi moodustamise protseduurist) põhjustatud varieeruvus hinnangutes. Valikuviga mõõdab hinnangu varieeruvust üle erinevate valimite, mida antud disainiga võib saada. Tõenäosuslike valikuuringute korral saab valikuviga hinnata. Enamasti valikuviga väheneb valimi mahu kasvades.

Loendiviga (*coverage error*) on põhjustatud ebakorrektsest loendist. Kui loend sisaldab üldkogumisse mittekuuluvaid objekte, siis on loend **ülekaetud**. Selliseid objekte on võimalik avastada ja loendist eemaldada. Kui loend ei sisalda kõiki üldkogumisse kuuluvaid objekte, on tegemist **alakaetusega**. Loendisse mittekuuluvad objektid ei saa kunagi valimisse sattuda. Sageli erinevad loendis mitteolevad objektid oma karakteristikute poolest neist objektidest, mis loendis on, ja sel juhul saame nihkega hinnangu. Ebakorrektsest loendist põhjustatud vea suurust ei suudeta tavaliselt mõõta. Loendivigadele tuleb juhtida tähelepanu uuringu kirjelduses.

Kaoviga (*nonresponse error*) on põhjustatud uuringus esinevast kaost, kui mingitel põhjustel ei saada andmeid kõigi valimisse sattunud objektide kohta. Võib esineda objekti kadu (nt inimene ei ela märgitud aadressil) või tunnuse väärtuse kadu (mõnel objektil puudub uuritava tunnuse väärtus, nt vastamisest keeldumine). Kadu iseloomustab nii küsitluse korraldust kui ka vastajate hoiakuid. Kao määr ulatub sageli 30%–40%-ni. Suur kao määr põhjustab nihet ja suurt dispersiooni. Kao kompenseerimiseks on kasutusel mitmeid meetodeid (omistusmeetodid, kaalumismeetodid).

Objektide asendamise viga (*substitution error*) on tingitud valimisse sattunud objekti asendamisest mõne teise kättesaadava objektiga. Näiteks kui antud aadressil ei ela mitte valimisse sattunud pere, vaid mõni teine pere, küsitleb intervjueerija viimast.

Mõõtmisviga (*measurement error*) tekib uuritava tunnuse mõõtmisel. Mõõtmisvigu võib olla mitmeid.

- **Mõõtmisvahendi viga**. Näiteks aja mõõtmisel kella viga. Küsimustiku korral on mõõtmisvahendiks küsimus ja mõõtmisviga tuleneb sellest, et küsimus ei mõõda täpselt seda, mida me tahame, et ta mõõdaks. Oluline on küsimuste oskuslik sõnastamine, soovitav on lasta seda teha spetsialistidel. Peab olema kindel, et küsija ja vastaja saavad küsimustest ühtmoodi aru. Tuleb vältida mitmetähenduslikke ja ebamääraseid küsimusi, samuti suunavaid küsimusi. Näiteks kui küsida inimestelt, kas nad kasutavad mingit toodet, kalduvad inimesed vastama „jah“ ka siis, kui nad seda toodet ei kasuta. See on mugavam ja tuleneb inimeste loomulikust hoiakust valmistada küsijale heameelt. Sageli esitatakse mitu erinevalt sõnastatud küsimust, mis peaksid mõõtma üht ja sama, ning analüüsimisel kasutatakse vastavate vastuste keskmist või summat.
- **Mõõtmisolukorra viga**. Eri objektide mõõtmise korral võib mõõtmisolukord olla erinev ja see võib avaldada mõju mõõteväärtustele.
- **Intervjueerija viga**. Küsimus esitatakse valesti, vastused märgitakse üles valesti. Intervjueerija vigu ning valesti antud vastuseid saab kontrollida mitmesuguste loogiliste seoste abil. Intervjueerija mõju saab uurida statistilise analüüsi vahenditega, võrreldes erinevate intervjueerijate poolt kogutud andmeid.

Töötlusviga tekib andmete kodeerimisel, sisestamisel, analüüsimisel. Vigade vähendamiseks tuleb kodeerimine ja sisestamine teha võimalikult lihtsaks, varustada loogiliste seoste ja väärtusvahemike kontrolliga. Programmis Excel on näiteks valik *Validation*. Eraldi tuleb kodeerida tunnuse väärtuse puudumine, et eristada seda sisestamata jäänud väärtusest. Puuduva väärtuse kood tuleb valida selline, mis vastusena ei sobi. Näiteks:

- kahe vastusevariandi korral „ei” – „0“, „jah“ – „1“ ja puuduv vastus „9“;
- kuu sissetuleku registreerimisel ei tohi puuduvat vastust kodeerida arvuga „0“. Siis näiteks „−1“.

Näide 6.15. Loendiviga ja kadu leibkonna eelarve uuringus

Leibkonna eelarve uuringus on loendiks rahvastikuregister. 2010. aasta uuringus oli üldkogumiks tavaleibkondades elavad kõik 2010. aasta 1. jaanuari seisuga vähemalt 15-aastased Eesti alalised elanikud, v.a pikka aega (vähemalt aasta) institutsioonides viibijad. Valimis oli 7803 isikut. Järgmises tabelis on esitatud

loendivead ja nende suurused (Tikva ja Arnik, 2012). Suurima osatähtsusega põhjus oli pikaajaline viibimine välismaal.

Loendiviga	Arv	% valimist	% loendiveast
Küsitletav surnud	85	1,1	21,8
Küsitletav viidud institutsiooni (hooldekodusse, vanglasse)	41	0,5	10,5
Küsitletav viibib välismaal vähemalt aasta	264	3,4	67,7
KOKKU	390	5,0	100,0

Kadu oli kokku 3781 isikut. Kao põhjused on jagatud kolme kategooriasse: kontakti puudumine, keeldumine ja muud põhjused. Üle poole kaost tuli kontakti puudumisest. Järgnevas tabelis on suurema osatähtsusega kadude põhjused kõigis kolmes kategoorias.

Kao põhjus	Arv	Osatähtsus kaos, %
Kontakti puudumine	1951	51,6
majja pääsenuna ei õnnestu küsitletavat tabada, kuna teda pole kodus	896	23,7
antud aadressil ei ela valimisikut	508	13,4
ei pääse majja/trepikotta	177	4,7
valimiisik ei viibi uuringuperioodil elukohas	135	3,6
valimiisik elab ajutiselt mujal	113	3,0
.....	....	.....
Keeldumine	1391	36,8
küsitletav keeldub (kategooriliselt) vastamast	882	36,8
küsitletav keeldub ajanappuse tõttu	259	6,9
küsitletaval puudub usaldus, on kahtlus andmete konfidentsiaalsuse tagamise suhtes	110	2,9
küsitletav on pettunud riigis, statistikas või uuringute kasulikkuses	55	1,5
.....	....	.....
Muud põhjused	439	11,6
küsitletav ei viibi kokkulepitud ajal ja kohas või väldib kontakti	110	2,9
kõrge vanus, ei saa ise hakkama, ei ole seetõttu võimeline uuringus osalema	103	2,7
.....	....	.....

Näide 6.16. Mõõtmisvead terviseuuringutes

Aastal 2012 Tervise Arengu Instituudi korraldatud Eesti täiskasvanud rahvastiku tervisekäitumise uuringu järgi on 19% täiskasvanutest rasvunud. Valimi suurus oli 5000 ja küsitlus viidi läbi posti teel. Valimisse sattunud isikutel paluti kirja panna oma kehamassiindeks. (Tekkel ja Veideman, 2013)
2014. aastal Tartu Ülikoolis kaitstud doktoritöö tulemusena selgus aga, et rasvunud on 32% täiskasvanutest (Eglit, 2014). Valimi maht oli 495 ja doktoritöö autor Triin Eglit külastas inimesi ise ning mõõtis nende kehamassiindeksit.
Miks nii suur erinevus?
Posti teel läbiviidud uuringu tulemusi võisid mõjutada mitmed vead.

1. Ülekaalulised inimesed on vähem altid vastama küsimusele nende kehamassiindeksi kohta. See võis põhjustada kaovea.
2. Inimesed ise kalduvad näitama väiksemat kehamassiindeksit, kui see tegelikult on. See on tahtlik mõõtmisviga.
3. Vanemad inimesed ei pruugi oma kaalu ja pikkust täpselt teada. Ka see on mõõtmisviga.

Valikuviga, mis on põhjustatud valimi kasutamisest, tuleb hinnata. Ülejäänud vigu tuleks uuringu planeerimisel ja korraldamisel vältida või minimeerida.

6.13. Ülesanded

Kogumi keskväärtuse usaldusvahemik, suur valim

6.1. Ajakirjas International Journal of Consumer Studies 2005. aastal ilmunud artiklis uuriti Austraalia majapidamiste finantskäitumist (Worthington, 2006). Valimisse kuulus 3268 majapidamist. Eluaseme keskmine väärtus neil majapidamistel oli 1469 tuhat Austraalia dollarit standardhälbega 1321 tuhat dollarit. Kui suur on Austraalia majapidamiste keskmine eluaseme väärtus usaldatavusega 0,9 ja usaldatavusega 0,95? VASTUS lk 668.

6.2. Ajakirjas Industrial and Labor Relations Review 1993. aastal ilmunud artiklis analüüsiti riikliku koolitusprogrammi mõju USA Michigani osariigis (Holzer jt, 1993). Selle programmi käigus eraldatakse ettevõtetele toetust töötajate koolitamiseks. Valimisse kuulus 390 ettevõtet, milles keskmine koolitusaeg töötaja kohta aastas oli 14,97 tundi

standardhälbega 25,71 tundi. Mitu tundi aastas koolitatakse töötajaid Michigani osariigi ettevõttes keskmiselt? VASTUS lk 668.

6.3. Valimi standardhälbe 53 korral saadi keskväärtuse usalduspiirid (242, 254) usaldatavusega 0,95. Kui suur oli valimi maht? VASTUS lk 668.

6.4. Kasutades valimit mahuga 250, saadi üldkogumi keskväärtuse usalduspiirideks $35{,}6 \pm 1{,}25$ usaldatavusega 0,9. Kui suur oli valimi standardhälve? VASTUS lk 668.

6.5. Valimi maht oli 500, keskmine 660 ja standardhälve 220. Keskväärtuse usalduspiirideks saadi $660{,}0 \pm 25{,}4$. Kui suur on usaldatavus? VASTUS lk 668.

6.6. Usaldatavusega 0,9 saadi usaldusvahemiku poollaiuseks 10,55. Kui suur oleks usaldusvahemiku poollaius usaldatavuse 0,99 korral, kui on teada, et valimi maht oli suur? VASTUS lk 668.

6.7. Mitu korda tuleb valimi mahtu suurendada, kui soovime usaldusvahemiku poollaiust vähendada 5 korda? VASTUS lk 668.

6.8. Mitu protsenti väheneb usaldusvahemiku poollaius, kui valimi mahtu suurendada 50-lt 100-ni? 200-lt 250-ni? VASTUS lk 668.

6.9. Kasutades tabelarvutuses funktsiooni NORM.S.INV, leida standardiseeritud normaaljaotusest usaldusvahemiku tõenäosuskordaja $z_{\alpha/2}$ usaldatavuse 0,95, 0,99 ja 0,999 korral. VASTUS lk 668.

6.10. Valimi maht oli 300 ja standardhälve 45. Keskväärtuse usaldusvahemiku poollaiuseks saadi 5,64. Kui suurt usaldatavust kasutati? Näpunäide: algul arvutada $z_{\alpha/2}$ ja siis leida usaldatavus β, kasutades tabelarvutuses standardiseeritud normaaljaotuse jaotusfunktsiooni leidmiseks funktsiooni NORM.S.DIST. VASTUS lk 668.

Lõplik kogum

6.11. Ettevõttes töötab 200 töötajat. Juhuslikult välja valitud 50 töötaja korral mõõdeti konkreetse tööülesande sooritamiseks kulunud aeg. Keskmiseks ajaks kujunes 13 minutit standardhälbega 5,7 minutit. Leida, millistesse piiridesse jääb keskmine selle ülesande sooritamiseks kulunud aeg kõikidel töötajatel. Usaldatavuseks võtta 95%. Arvutused teha kahel juhul:

a) eeldada, et kogumi maht on valimi mahuga võrreldes väga suur;
b) kasutada usalduspiiride leidmiseks lõpliku kogumi mahuga valemit.

VASTUS lk 668.

6.12. Eesti Statistikaamet kogub ettevõtetelt mitmesugust majandusstatistikat. Alla 20 töötajaga ettevõtetest tehakse stratifitseeritud

(alamgruppidesse jaotatud) lihtne juhuslik valik hõivatute arvu ja tegevusala järgi. Tabelis on toodud kogumi ja valimi maht kolmel tegevusalal aastal 2012[4]. Leida standardvea parandus, mida tuleb arvestada lõpliku kogumi mahu korral. VASTUS lk 668.

Tegevusala	Kogum	Valim
Jäätmetöötlus ja -kõrvaldus	32	22
Ehitus	8780	1218
Jaemüük posti või Interneti teel	623	42

Väike valim

6.13. USA-s mõõdetakse alates 1991. aastast erinevate keemiliste elementide kontsentratsiooni müügil olevates toiduainetes. Analüüse kogutakse üldise toitumise uuringu (*Total Diet Study*) käigus igal aastal neli korda ning kokku analüüsitakse 280 erinevat müügil olevat toidukaupa. Analüüside aruanded avaldatakse USA toidu- ja ravimiameti (FDA) kodulehel. Vastavalt aastatel 2006–2011 läbiviidud mõõtmistele oli elavhõbeda sisaldus mõningates kalades järgmine (Total Diet Study, 2014):

a) küpsetatud lõhe: valimi maht 9, keskmine 0,022 mg/kg, standardhälve 0,011 mg/kg;
b) kuivatatud tuun: valimi maht 9, keskmine 0,142 mg/kg, standardhälve 0,153 mg/kg.

Leida elavhõbeda sisalduse keskväärtuse usalduspiirid mõlema kalatoote korral. VASTUS lk 668.

6.14. Valimi maht oli 10, keskmine 42,5 ja standardhälve 12,1. Keskväärtuse usaldusvahemikuks saadi $42{,}5 \pm 7{,}01$. Millise usaldatavusega see vahemik on? VASTUS lk 668.

Valimi mahu planeerimine

6.15. Näites 6.5 leiti, kui palju USA noored keskmiselt säästavad. Valimi maht oli 243, keskmine 463,5 dollarit ja standardhälve 460,1 dollarit. Usaldusvahemikuks saadi $463{,}5 \pm 57{,}8$ dollarit usaldatavusega 0,95. Suhteline viga oli 12,5%. Leida, kui suur peaks olema valimi maht, et suhteline viga oleks väiksem kui 5%. VASTUS lk 668.

6.16. Eeluuringus oli valimi suurus 200 isikut. Huvipakkuva tunnuse valimi keskmine tuli 230 ja standardhälve 14,6. Kui palju isikuid tuleks veel küsitleda, et usaldatavuse 0,95 korral oleks usaldusvahemiku poollaius väiksem kui 1? VASTUS lk 668.

Kaheväärtuselise tunnuse osakaalu usalduspiirid

6.17. TNS Emori 2013. aastal läbiviidud uuringus „Heategevusalaste hoiakute uuring“ vastas küsitlusele 1140 Eesti elanikku vanuses 18–60. Küsimusele „Kas olete püsiannetaja (toetate mingit valdkonda või

[4] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel EM026: ettevõtete aastastatistika üldkogum, valim ja vastanud.

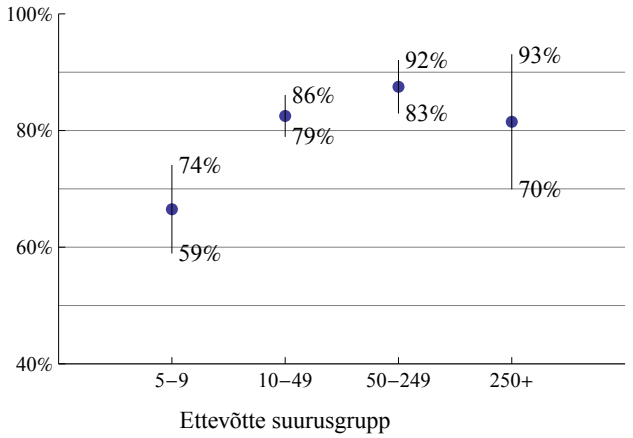
ühendust teatud regulaarsusega, rohkem kui üks kord)?“ vastas „Jah“ 12% vastajatest (Kandla jt, 2013). Leida usaldatavusega 0,90, kui suur osa 18–60-aastastest Eesti elanikest on püsiannetajad? VASTUS lk 668.

6.18. 2013. aasta Eesti sotsiaaluuringu valimis oli 25–64-aastaseid isikuid 7727, nendest 613 oli teinud aasta jooksul sissemakseid erapensioniskeemidesse (pensioni III sammas). Sissemakseid teinutel oli keskmise sissemakse suurus 420,91 eurot standardhälbega 554,74 eurot. (*Eesti sotsiaaluuring* 2013) Leida usaldatavusega 0,95,

a) kui suur osa 25–64-aastastest isikutest oli 2013. aastal teinud sissemakseid erapensioniskeemidesse;
b) milline oli neil keskmine sissemakse suurus aastas.

VASTUS lk 668.

6.19. 2013. aastal viidi Sotsiaalministeeriumi tellimusel Eestis läbi „Töölepingu seaduse uuring“, milles käsitleti individuaalse tööõiguse alast õigusteadlikkust ja õiguskindlust, töölepingu sõlmimise, töötingimuste kokkuleppimise ning töösuhte lõppemise praktikaid (Masso jt, 2013). Valimisse kuulus 4352 asutust ja ettevõtet. Joonisel on toodud ettevõtete ja asutuste osakaal, kellel subjektiivse hinnangu järgi oleks tööõigusest vaja rohkem teada, jaotatud ettevõtete suurusgruppidesse töötajate arvu järgi. Esitatud on osakaalude usaldusvahemike alumised ja ülemised piirid. Leida, kui palju oli valimis igasse suurusgruppi kuuluvaid ettevõtteid, kui on teada, et usaldusvahemikud on antud usaldatavusega 0,95. VASTUS lk 668.



6.20. EUROSTUDENT on üleeuroopaline uuring, mis keskendub kõrghariduse sotsiaalsele dimensioonile üliõpilaste vaatenurgast. Praeguseks on EUROSTUDENT-i uuringut tehtud neljal korral alates 2000. aastast, Eesti on osalenud kahes viimases uuringus — EUROSTUDENT III (2008) ja EUROSTUDENT IV (2010). Viimases uuringus osales 1219 üliõpilast 34 kõrgkoolist. (Kirss jt, 2011)

1. Mitu protsenti on suurim osakaalu usaldusvahemiku poollaius selle valimi korral, kui usaldatavuseks on 0,95?
2. Kõigist tudengitest elas küsitluse ajal ühiselamus 18%. Millistesse piiridesse jääb ühiselamus elavate üliõpilaste osakaal usaldatavusega 0,95?

VASTUS lk 668.

6.21. Arseeni leidub looduses metallirohketes geoloogilistes materjalides ning ka inimtegevuse tagajärjel vase, plii ja kulla tootmise kõrvalsaadusena. Nii satub see vette, õhku ja sealt toiduainetesse. Aastatel 1991–2005 koguti USA-s üldise toitumise uuringu käigus 443 mahlaproovi, millest neljas oli arseenisisaldus lubatust suurem (Bolger, Egan ja Tao, 2008). Leida usaldatavusega 0,90, kui suur osa USA-s müüdavast mahlast sisaldab arseeni üle lubatud piirmäära. VASTUS lk 668.

6.22. Näites 6.8 käsitletud riigikogu valimiste eel toimunud küsitluse tulemusena pidi valima minema 51,0% ± 4,4% kõigist valimisealistest kodanikest. Valimi maht oli 500. Kui suurt valimit oleks tulnud kasutada, et osakaalu usaldusvahemiku poollaius ei oleks suurem kui 1%? Usaldatavuseks võtta 0,95. VASTUS lk 668.

Kvalitatiivse tunnuse osakaalude usaldusvahemikud

6.23. EUROSTUDENT IV küsimustikus oli üks küsimus üliõpilase peamise sissetulekuallika kohta. Peamine sissetulekuallikas on see, mis moodustab kogusissetulekust rohkem kui 50%. Vastajaid oli 1219 ja vastused jagunesid järgmiselt (Kirss jt, 2011):

- enda teenitud palk 48%;
- perekonnalt/partnerilt saadud raha 33%;
- riigi toetused 15%;
- muu 4%.

Leida, kui suure osal Eesti üliõpilastest on peamiseks sissetulekuallikaks enda teenitud palk, kui suurel osal perekonnalt/partnerilt saadud raha ja kui suurel osal riigi toetused. Kasutada usaldatavust 0,95. VASTUS lk 668.

6.24. Invatakso on sotsiaalteenus, mida osutatakse Tallinna rahvastikuregistrisse kantud puuetega inimestele. Teenust finantseeritakse Tallinna eelarvest. Alates 2002. aastast korraldab MTÜ Tallinna Puuetega Inimeste Koda küsitlust, mille eesmärgiks on uurida rahulolu taksoteenusega. Valim moodustatakse juhuvaliku alusel. 1. novembri 2011 seisuga oli invatakso kasutamise võimalus 989 kliendil. 2011. aasta ankeetküsitlus saadeti 125 kliendile, vastas 76 isikut. Rahulolu taksojuhtide tööga oli järgmine (Tiik, 2011):

- „hea“ — 68 vastajat;
- „rahuldav“ — 5 vastajat;
- „mitterahuldav“ — 3 vastajat.

Leida usaldatavusega 0,95, kui suur osa invatakso kasutajatest hindavad taksojuhtide tööd heaks, kui suur osa rahuldavaks ja kui suur osa mitterahuldavaks. VASTUS lk 668.

Mediaani usalduspiirid

6.25. Valimi maht on 500. Leida, mitmenda elemendi väärtus kasvavalt järjestatud valimis on mediaani alumine usalduspiir ja mitmenda elemendi väärtus on ülemine usalduspiir, kui usaldatavuseks võtta

a) 0,95,
b) 0,99.

VASTUS lk 668.

ÜL06Valikvaatlused

Järgmiste ülesannete andmed on failis ÜL06Valikvaatlused

Keskväärtuse usaldusvahemik, suur valim

A.6.1. Määramaks Maroko apelsinide keskmist kaalu, on kaalutud ära 50 juhuslikult valitud apelsini. Leida apelsinide kaalu

a) keskväärtuse, dispersiooni ning standardhälbe punkthinnangud;
b) standardviga;
c) keskväärtuse usaldusvahemik usaldatavusega 0,95.

VASTUS lk 668.

A.6.2. 1994. aastal viis ajakiri Quality Progress oma lugejate hulgas läbi küsitluse. 100 tuhande tellija hulgast valiti juhuvaliku abil välja 9117 isikut, kellele saadeti posti teel vastav küsimustik. Täidetud küsimustikke laekus 4823, s.t vastamisprotsent oli 53%. Tabelis on esitatud uuringul põhinevad andmed ametikohtade ja palkade kohta.

1. Leida iga ametikoha jaoks palga usaldusvahemiku alumine ja ülemine piir usaldatavusega 95%.
2. Milliste ametikohtade palkade usaldusvahemikud osaliselt kattuvad?
3. Leida suhtelised vead. Millise ameti palk on selle uuringu järgi määratud kõige täpsemini? Miks on sellel tulemusel täpsus kõige suurem?

VASTUS lk 668.

A.6.3. Näites 6.4 leiti, tuginedes 2012. aasta Leibkonna eelarve uuringu andmetele, et keskmised kulud toidule ja mittealkohoolsetele jookidele pereliikme kohta aastas on $915{,}3 \pm 17$ eurot. Tabelis on ainult üheliikmeliste leibkondade andmed (*Leibkonna eelarve uuring* 2012). Leida usaldatavusega 0,95 keskmised kulud aastas toidule ja mittealkohoolsetele jookidele üheliikmelise leibkonna korral. Võrrelda tulemusega, mis saadi kõikide leibkondade andmete kasutamisel. VASTUS lk 669.

A.6.4. Valimis on erinevate isikute rahaline netosissetulek palgatööst

aastas, eurodes. Küsitletud on kahest piirkonnast: Põhja-Eestist (kood 1) ja Kirde-Eestist (kood 3). Andmed pärinevad Eesti Statistikaameti sotsiaaluuringust (*Eesti sotsiaaluuring* 2013).

1. Leida keskmise netosissetuleku usaldusvahemik Põhja-Eestis ja Kirde-Eestis. Kasutada usaldatavust 0,95.
2. Võrrelda valimite standardhälbeid ja usaldusvahemike poollaiusi. Milline on järeldus?
3. Kui suur oli kummaski piirkonnas keskmine netosissetulek kuus?

VASTUS lk 669.

A.6.5. Tabelis toodud andmed pärinevad äriregistrist. Ühes valimis on 100 finantsvahenduses tegutsevat ettevõtet ja teises 100 kinnisvara valdkonnas tegutsevat ettevõttet. Toodud on nende ettevõtete kasum ühe töötaja kohta 2010. aastal (eurot). Leida ühe töötaja kohta saadud keskmise kasumi hinnang mõlema tegevusvaldkonna jaoks usaldatavusega 0,95. Kas võib väita, et aastal 2010 oli keskmine kasum ühe töötaja kohta neis tegevusvaldkondades erinev? VASTUS lk 669.

Keskväärtuse usaldusvahemik, väike valim

A.6.6. Paljudel USA ülikoolidel ja kolledžitel on rahaline fond, millesse teevad sissemakseid eraisikud ja ettevõtted. Õppeasutuse tegevuskulud ja töötajate palgad kaetakse selle fondi teenitud intressidega. Tabelis on toodud juhuslikult väljavalitud kaheksa kolledži vastava fondi suurus miljonites dollarites[5]. Leida keskmine kolledži fondi suurus usaldatavustega 75%, 90% ja 95%. VASTUS lk 669.

A.6.7. Artikli „Small Firm Internationalization and Business Strategy“ autorid viisid Suurbritannia tootmisettevõtete seas läbi uuringu, mille eesmärgiks oli selgitada rahvusvahelisustumise arengutrende väikeettevõtete äristrateegias (Bell, Crick ja Young, 2004). Intervjueeriti 30 väikeettevõtte (töötajate arv väiksem kui 250) juhtivtöötajaid. 15 ettevõtet olid n-ö traditsioonilistest tegevusvaldkondadest (kergetööstus, toiduainetetööstus, tekstiilitööstus) ja 15 ettevõtet teadmusmahukatest tegevusvaldkondadest (elektroonika, informatsiooni- ja kommunikatsioonitehnoloogia).

Üheks näitajaks, mida ettevõtete korral analüüsiti, oli ekspordi osakaal kogukäibes. Leida kummagi grupi jaoks keskmine ekspordi osakaal usaldatavusega 0,95. VASTUS lk 669.

A.6.8. Arst soovib paremini oma tööd planeerida. Ühe nädala vältel registreerib ta 25 korral patsiendi peale kulutatud aja. Andmed on toodud tabelis. Leida, millisesse vahemikku jääb keskmine patsiendi kohta kuluv aeg usaldatavusega 95%. Kasutada kaht meetodit: normaaljao-

[5]Allikas: Chronicle of Higher Education Almanac, Sept, 2. 1996, `http://chronicle.com`

tuse põhjal (suur valim) ja t-jaotuse põhjal (väike valim) ning võrrelda tulemusi. VASTUS lk 669.

A.6.9. Tabelis on 599 isiku vanused. See on kogum. Võtta kogumist kolm juhuvalimit mahuga 10 ja leida iga valimi põhjal keskmine vanus ning selle usalduspiirid usaldatavusega 0,95. Seejärel leida kogumi aritmeetiline keskmine ja vaadata, kas valimite põhjal leitud usalduspiirid katavad kogumi keskmise. Juhuvalimite moodustamiseks saab programmis Excel kasutada vahendit *Sampling* komplektist *Data Analysis*.

Kaheväärtuselise tunnuse osakaalu usaldusvahemik

A.6.10. Eesti 2012. aasta leibkonnaeelarve uuringus paluti osalejatel vastata erinevate kodumasinate olemasolu kohta oma leibkonnas (*Leibkonna eelarve uuring* 2012). Tabelis on andmed nõudepesumasinate olemasolu kohta viieliikmelistes leibkondades („1“ — on nõudepesumasin, „2“ — ei ole nõudepesumasinat). Leida usaldatavusega 0,95, kui suurel osal viieliikmelistest leibkondadest on kodus nõudepesumasin. VASTUS lk 669.

A.6.11. 2015. aastal TTÜ-s kaitstud magistritöös „Eesti noorte säästmisharjumused“ viidi läbi küsitlus 18–35-aastaste Eesti noorte seas (Tiitso, 2015). Üheks küsimuseks oli „Kas olete liitunud III pensionisambaga?“. Vastusevariant 1 on „Jah“ ja 2 on „Ei“. Leida usaldatavusega 0,95, kui suur osa 18–35-aastastest Eesti noortest on liitunud III pensionisambaga. VASTUS lk 669.

Kvalitatiivse tunnuse osakaalude usaldusvahemikud

A.6.12. Eesti Konjunktuuriinstituut viib pidevalt läbi tarbijabarometri uuringuid, kus palutakse hinnata majandusliku olukorra muutumist nii vastaja perel kui ka Eestis tervikuna. Küsitlusi viiakse läbi telefoni teel, valimis on 800 inimest. Tabelis on esitatud küsimusele „Kulutamine püsikaupade ostuks järgmisel 12 kuul (võrreldes viimase 12 kuuga)“ vastanute vastused 2004. aasta oktoobris ja detsembris (Josing, 2014).

1. Leida kõigi vastusevariantide jaoks usaldusvahemike alumised ja ülemised piirid mõlema kuu korral.
2. Kas usaldusvahemike põhjal võib väita, et kahe kuuga toimusid muutused?

VASTUS lk 669.

A.6.13. Ülesandes A.6.11 nimetatud uuringus oli üheks küsimuseks „Kas põhi- ja/või keskkooli viimases klassis peaks olema majandusmatemaatika või -õpetus kohustuslik?“ Vastusevariandid olid:

„Jah“ — 1;
„Võiks olla valikaine“ — 2;
„Ei“ — 3.

Leida vastuste osakaalud ja nende usaldusvahemikud usaldatavusega 0,95. VASTUS lk 669.

Mediaani usalduspiirid

A.6.14. Ülesandes A.6.3 toodud andmete põhjal leida üheliikmeliste leibkondade korral toidule tehtavate kulutuste mediaani usaldusvahemiku alumine ja ülemine piir usaldatavusega 0,95. Kui üksinda elav inimene kulutas 2012. aastal toidule ja mittealkohoolsetele jookidele keskmiselt 80 eurot kuus, kas ta oli väiksema kulutustega 50% hulgas või mitte? Aga kui tema kulutused olid 95 eurot kuus? VASTUS lk 670.

A.6.15. Tabelis on müügitulu ühe töötaja kohta 2010. aastal 323 Eesti väikeettevõttes, mille tegevusala on veondus ja laondus. Andmed on eurodes. Leida müügitulu mediaani usalduspiirid usaldatavusega 0,9. VASTUS lk 670.

Peatükk 7

Hüpoteeside statistiline kontrollimine

2013. aasta kevadel rõõmustasid paljud eestlased: kahekordne olümpiavõitja Andrus Veerpalu mõisteti dopinguskandaalis õigeks. Veerpalu süüdistati dopingu tarvitamises, kuna kasvuhormoonide tase tema proovis ületas lubatud piirmäära. Rahvusvaheline Spordiarbitraažikohus leidis, et kasvuhormooni piirmäärad polnud õiged. Teema oli aktuaalne veel kaua aega ning paljud eestlased püüdsid ennast kurssi viia testide ning mõõtmiste statistiliste nüanssidega.

Analoogne piirmäära probleem esineb ka ärinduses ja majandusanalüüsis. Näiteks kavatseb tootja oma toote läbimüügi suurendamiseks tellida reklaamikampaania. Et hinnata reklaamikampaania tulemuslikkust, viiakse läbi küsitlused enne ja pärast kampaaniat. Enne reklaamikampaaniat vastas 100 küsitletavast 32, et tarbivad antud toodet. Pärast kampaaniat läbiviidud küsitluse tulemusel vastasid 45 inimest 100-st, et tarbivad seda toodet. Erinevus on 13 inimest. Kas sellest piisab, et kinnitada: kampaania oli tulemusrikas, tarbimine suurenes? Küsitlusele vastajad olid ju erinevad.

Aastatel 2004–2010 sai Ettevõtluse Arendamise Sihtasutuselt toetust 508 Eesti väikeettevõtet kogusummas 13,8 miljonit eurot (Hartsenko ja Sauga, 2013). Kas toetustest oli kasu? Kas toetust saanud ettevõtetel suurenes müügitulu, kasum ja eksport rohkem kui ettevõtetel, kes toetust ei saanud? Milline on kriteerium?

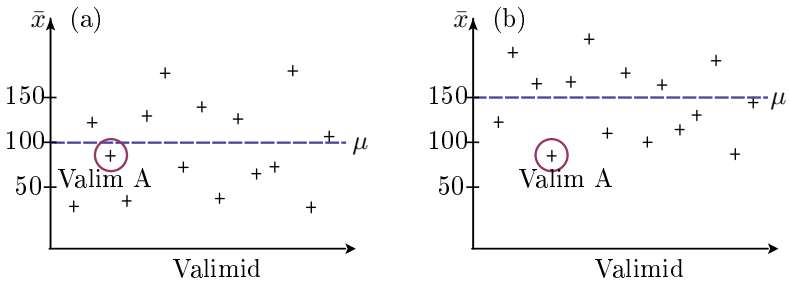
Hüpoteeside statistilise kontrollimise meetod võimaldab kindlaks teha, kas **erinevus on piisav** väitmaks, et selle on põhjustanud üks oluline tegur, või on erinevus seletatav mõõdetava tunnuse väärtuste juhusliku varieerumisega. Otsuse vastuvõtmiseks kasutatakse **statistilist kriteeriumi**.

7.1. Nullhüpotees, sisukas hüpotees ja statistiline kriteerium

Hüpoteeside statistilist kontrollimist kasutatakse väga mitmesuguste probleemide analüüsimisel.

- Kas mingil tegevusel oli mõju? Näiteks muudatused seadusandluses, ettevõtte juhtimises, töökorralduses, uute tehnoloogiate või töövahendite kasutuselevõtt.
- Kas toodang vastab teatud standarditele? Näiteks vastavus Euroopa Liidu, Eesti või mõne muu riigi standarditele, kvaliteedinõuetele, ostu-müügilepingus fikseeritud parameetritele.
- Kas vastus ankeedis esitatud küsimusele sõltub mõnest tausttunnusest nagu sugu, vanus, elukoht?

Miks on vaja statistilist kriteeriumi? Vaatleme järgmist situatsiooni. Joonisel 7.1 on toodud kahest erinevast üldkogumist võetud juhuvalimid. Horisontaaltelg on valimite eristamiseks, vertikaalteljel on valimite keskmised. Kogumi keskväärtus on μ (kriipsjoon), mis pole reaalsetes oludes teada. Hüpoteesiks on, et kogumi keskväärtus $\mu = 100$.



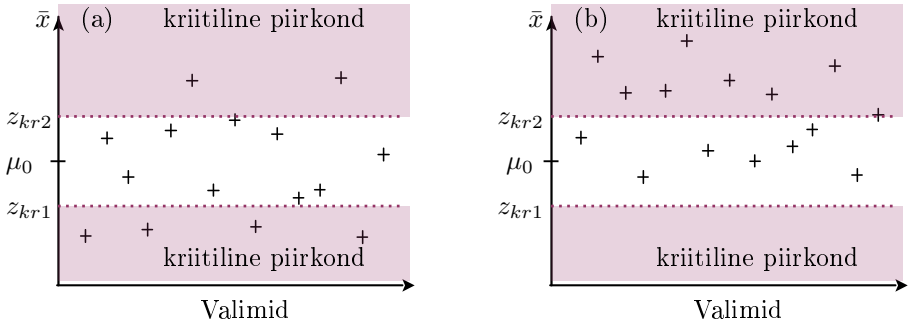
Joonis 7.1. Kahest erineva keskväärtusega üldkogumist võetud juhuvalimid. Vertikaalteljel on valimite keskmised $\bar{x}$, horisontaaltelg loendab valimeid. Kriipsjoon näitab vastava kogumi keskväärtust μ

Otsustamiseks, kas meie hüpotees kehtib, võetakse **üks** juhuvalim. Oletame, et meile satub valim A. Me **ei tea**, kummast kogumist see võeti. Valim võib olla vasakpoolsest kogumist, mille keskväärtus vastab meie püstitatud hüpoteesile $\mu = 100$. Aga valim võib olla ka parempoolsest kogumist (b), mille keskväärtus erineb meie püstitatud hüpoteesist. Kuidas otsustada, kumma situatsiooniga on tegemist?

On vaja teatud **objektiivset kriteeriumi**: kui kaugel peaks valimi keskmine $\bar{x}$ olema meie hüpoteesile vastavast väärtusest 100, et võiksime hüpoteesi tagasi lükata ning väita: pole tõenäoline, et see valim pärineb kogumist, mille keskväärtus on 100, see valim pärineb kogumist, mille keskväärtus $\mu \neq 100$.

Joonisel 7.2 on punktiirjoontega lisatud kriteeriumid z_{kr1} ja z_{kr2}, mis aitavad valimi keskmise põhjal otsustada, kummast kogumist võiks

valim olla. Kui valimi keskmine jääb punktiirjoonte vahele, on tõenäoline, et tegemist on kogumiga, mille keskväärtus vastab meie poolt püstitatud hüpoteesile $\mu = \mu_0$. Kui valimi keskmine on punktiirjoonte vahelt väljas, kriitilises piirkonnas, on vähetõenäoline, et tegemist on kogumiga, mille keskväärtus $\mu = \mu_0$. Sel juhul järeldame, et $\mu \neq \mu_0$.



Joonis 7.2. Määratud on kriteeriumid z_{kr1} ja z_{kr2}. (a) Nullhüpotees tegelikult kehtib, kogumi keskväärtus $\mu = \mu_0$. (b) Nullhüpotees tegelikult ei kehti, kogumi keskväärtus $\mu \neq \mu_0$. Paneme tähele, et mõlemal juhul võib meile sattuda valim, mille keskmine asub kas varjutamata või varjutatud piirkonnas, kuid juhul (b) on varjutatud kriitilisse piirkonda sattumise tõenäosus suurem

Nullhüpotees ja sisukas hüpotees

Kontrollimiseks on otstarbekas esitada hüpoteeside paar. Paarikaupa esitatud hüpoteesid peavad teineteist välistama ja üks neist peab kindlasti kehtima. Üks hüpotees püstitatakse tavaliselt nii, et üldkogumi mingi parameeter on võrdne teatud väärtusega või et kogum vastab teatud standardile või erinevus kahe kogumi vahel puudub vms. Selliselt püstitatud hüpoteesi nimetatakse **nullhüpoteesiks** H_0, seda välistavat aga **alternatiivseks** ehk **sisukaks hüpoteesiks** H_1.

Nullhüpotees vastab mingile konkreetsele väärtusele ja kehtib vaikimisi. Valim võib nullhüpoteesi ümber lükata, aga ei pruugi. Sisukas hüpotees on see, mida uurija tavaliselt soovib tõestada. Järgnevalt on toodud mõningaid näiteid hüpoteesipaaride püstitamise kohta.

H_0: Suur- ja väikeettevõtetes on tööviljakus ühesugune: $\mu_1 = \mu_2$.
H_1: Suur- ja väikeettevõtetes on tööviljakus erinev: $\mu_1 \neq \mu_2$.

H_0: Maisihelbepakkide keskmine kaal on 250 g: $\mu = 250$.
H_1: Maisihelbepakkide keskmine kaal ei ole 250 g: $\mu \neq 250$.

H_0: Reklaamikampaania tulemusena läbimüük ei suurenenud: $\mu_1 \leq \mu_2$.
H_1: Reklaamikampaania tulemusena läbimüük suurenes: $\mu_1 > \mu_2$.

H_0: Väärtpaberi tulumäär allub normaaljaotusele.
H_1: Väärtpaberi tulumäär ei allu normaaljaotusele.

H_0: Töötajate rahulolu ja info kättesaadavuse vahel ei ole olulist seost.
H_1: Töötajate rahulolu ja info kättesaadavuse vahel on oluline seos.

Tuginedes valimi põhjal arvutatud teststatistikule, võime väita üht kahest:

- erinevus nullhüpoteesiga püstitatud väärtusest on nii suur, et on tõenäoline: nullhüpotees ei kehti;
- erinevus nullhüpoteesiga püstitatud väärtusest on nii väike, et on tõenäoline: nullhüpotees kehtib.

Viimasel juhul valimi põhjal leitud teststatistik küll erineb nullhüpoteesiga püstitatud väärtusest, aga sellest ei piisa nullhüpoteesi tagasilükkamiseks. Erinevus on tingitud lihtsalt sellest, et me ei vaatle tervet kogumit, vaid meil on juhuvalim. Nullhüpotees on aga püstitatud kogumi jaoks. Näiteks kui kogumi keskväärtus on μ_0, siis valimite keskmised hajuvad ümber selle vastavalt valimjaotusele (vt alapeatükk 6.4).

Nullhüpotees kas võetakse vastu või lükatakse tagasi

Hüpoteesipaar püstitatakse üldkogumi jaoks. Püstitatud hüpoteesi kontrollimisel juhuvalimi abil võib tulemus olla üks kahest:
1) tuleb jääda nullhüpoteesi H_0 juurde;
2) nullhüpotees lükatakse tagasi ja vastu võetakse sisukas hüpotees H_1.

Erinevate hüpoteeside kontrollimiseks kasutatakse erinevaid teste, millel igaühel on oma statistiline kriteerium (z-test, t-test, χ^2-test, märgitest, Manni-Whitney test, Wilcoxoni test, ühe- ja mitmefaktoriline dispersioonanalüüs). Testi valik sõltub probleemi püstitusest, mõõdetava suuruse iseloomust (kas nimi-, järjestus- või intervallskaala), mõõdetava suuruse nivoode arvust (kaks või enam). Kõikide testide korral võrreldakse vaatlusandmete (empiiriliste andmete) põhjal leitud teststatistiku K empiirilist väärtus vastavast juhusliku suuruse jaotusseadusest leitud kriitilise väärtusega K_{kr}. Kui empiiriline väärtus ületab kriitilise, siis teststatistik langeb **kriitilisse piirkonda**.

Kuidas leitakse kriitiline väärtus? See võetakse selline, et kehtiva nullhüpoteesi korral oleks kriitilisse piirkonda sattumise tõenäosus ehk riskitase väiksem teatud väärtusest, mis tavaliselt on 0,05. Seda nimetatakse olulisuse nivooks.

Olulisuse nivoo

Olulisuse nivoo α näitab, millest väiksem peab olema kriitilisse piirkonda sattumise tõenäosus kehtiva nullhüpoteesi korral.

Vaatame sellist näidet, kus urnis on 100 kuuli, mustad ja valged. Sealt tuleb võtta juhuslikult 10 kuuli ja nende põhjal otsustada, kas urnis on musti ja valgeid kuule võrdselt või ei ole.

Oletame, et musti ja valgeid kuule on urnis võrdselt, mõlemaid 50. See on nullhüpotees. Leiame, millise tõenäosusega satub sel juhul kätte 10 valget või 9 valget ja 1 must või 8 valget ja 2 musta jne. Mustade ja valgete arv juhuslikult valitud 10 kuuli hulgas allub hüpergeomeetrilisele jaotusele (Jõgi, 2000). Kui urnis on nii musti kui valgeid kuule 50, siis tõenäosus, et 10 kuuli hulgas on m valget, on

$$P(X = m) = \frac{C_{10}^{m} C_{90}^{50-m}}{C_{100}^{50}}, \tag{7.1}$$

kus C_n^m on kombinatsioonide arv n elemendist m kaupa. Valemi (7.1) abil leitud tõenäosused on esitatud tabelis 7.1.

Tabel 7.1. Valgete kuulide arvu tõenäosusjaotus

Valgete arv m	Tõenäosus $P(X = m)$		
10	0,000 59	$P(X > 8) = 0{,}00779 < 0{,}025$	Kriitiline pk
9	0,007 2		
8	0,038		
7	0,11		
6	0,21		
5	0,26		
4	0,21		
3	0,11		
2	0,038		
1	0,007 2	$P(X < 2) = 0{,}00779 < 0{,}025$	Kriitiline pk
0	0,000 59		

Väikeseks loetakse tõenäosust, mis on väiksem kui 0,05. Selle järgi määratakse kriitilised väärtused: teststatistiku väärtuse sattumine kriitilisse piirkonda on kehtiva nullhüpoteesi korral väiksem kui 0,05. Kuna kriitiline piirkond jaguneb kaheks võrdseks osaks, siis kummassegi ossa sattumise tõenäosus peab olema väiksem kui 0,025. Tabelist 7.1 näeme, et

$$\begin{aligned} P(X < 2) &= P(X = 1) + P(X = 0) = \\ &= 0{,}0072 + 0{,}00059 = 0{,}00779 < 0{,}025, \\ P(X > 8) &= P(X = 9) + P(X = 10) = \\ &= 0{,}0072 + 0{,}00059 = 0{,}00779 < 0{,}025. \end{aligned}$$

Järelikult kriitiline piirkond on $P(X < 2)$ ja $P(X > 8)$ ning kriitilised väärtused vastavalt 2 ja 8. Kui valgete kuulide arv m on vahemikus

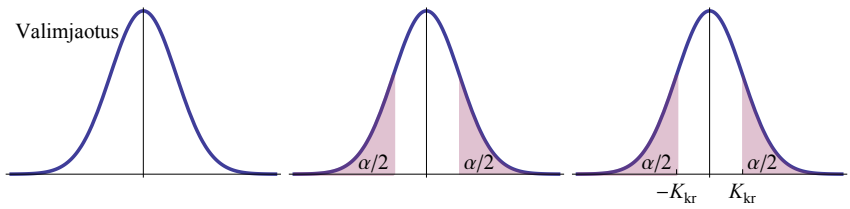
$2 \leq m \leq 8$, võtame vastu nullhüpoteesi: urnis on musti ja valgeid kuule võrdselt. Kui valgete kuulide arv $m < 2$ või $m > 8$, on nullhüpotees ümber lükatud ja võtame vastu sisuka hüpoteesi: valgeid ja musti kuule ei ole urnis võrdne arv.

Nullhüpoteesi võib püstitada ka nii, et urnis on 40 valget kuuli ja 60 musta kuuli. Sellisel juhul tulevad kriitilised väärtused teistsugused. Me ei saa aga püstitada nullhüpoteesi nii, et urnis on valgeid kuule rohkem. Selle väite alusel ei saa leida kriitilisi väärtusi.

Nullhüpotees tuleb püstitada nii, et on võimalik seda ümber lükata.

Et nullhüpoteesi saaks ümber lükata, peab kehtiva nullhüpoteesi korral olema võimalik leida kriitilisi väärtusi. Üldjuhul leitakse kriitilised väärtused järgmiselt (vt ka joonis 7.3):

1) leitakse teststatistiku valimjaotus kehtiva nullhüpoteesi korral;
2) võetakse ette olulisuse nivoo α, millest väiksem peab kriitilisse piirkonda sattumise tõenäosus olema kehtiva nullhüpoteesi korral;
3) teststatistiku valimjaotusest leitakse olulisuse nivoole vastavad kriitilised väärtused.



Joonis 7.3. Teststatistiku kriitilise väärtuse leidmine. Algul leitakse teststatistiku valimjaotus kehtiva nullhüpoteesi korral, siis määratakse olulisuse nivoo α põhjal kriitiline piirkond (varjutatud) ja seejärel kriitilised väärtused $-K_{kr}$ ning K_{kr}

Eeskiri hüpoteesi kontrollimiseks

Üldine eeskiri hüpoteesi kontrollimiseks

1. Otsustada, millist testi tuleb kasutada.
2. Püstitada hüpoteesipaar: nullhüpotees H_0 ja sisukas ehk alternatiivne hüpotees H_1.
3. Kasutades juhuvalimit, leida valitud testile vastava statistilise parameetri empiiriline väärtus K.
4. Võtta ette olulisuse nivoo (tavaliselt 5% või 1%) ja leida teststatistikule vastavast jaotusseadusest statistiku kriitiline väärtus K_{kr}.

5. Võrrelda, kas teststatistiku empiiriline väärtus K langeb väärtusega K_{kr} määratud kriitilisse piirkonda või ei lange.
6. Teha järeldus hüpoteesi kohta:
   - kui K ei lange kriitilisse piirkonda, jäädakse nullhüpoteesi juurde;
   - kui K langeb kriitilisse piirkonda, on nullhüpotees ümber lükatud ja võetakse vastu sisukas hüpotees.

7.2. Keskväärtuse testimine suure valimi korral

Alapeatükis 6.4 esitatud tsentraalse piirteoreemi kohaselt alluvad suurte valimite ($n > 30$) keskmised normaaljaotusele keskväärtusega μ ja standardhälbega $\sigma/\sqrt{n}$, kus μ on kogumi keskväärtus ning σ kogumi standardhälve. Kuna me kogumi standardhälvet ei tea, kasutame selle hinnangut: valimi standardhälvet s.

Igal konkreetsel juhul on normaaljaotuse keskväärtus ja standardhälve erinevad ning järelikult ka statistilise kriteeriumi väärtus tuleb erinev. Otstarbekas on valimi keskmine teisendada standardiseeritud skaalale z (3.9):

$$z = \frac{\bar{x} - \mu_0}{se}, \tag{7.2}$$

kus standardviga (valimi keskmise valimjaotuse standardhälve)

$$se = \frac{s}{\sqrt{n}}. \tag{7.3}$$

Testi, mida kasutatakse suurte valimite korral kogumi keskväärtuse testimiseks, nimetatakse seetõttu **z-testiks**.

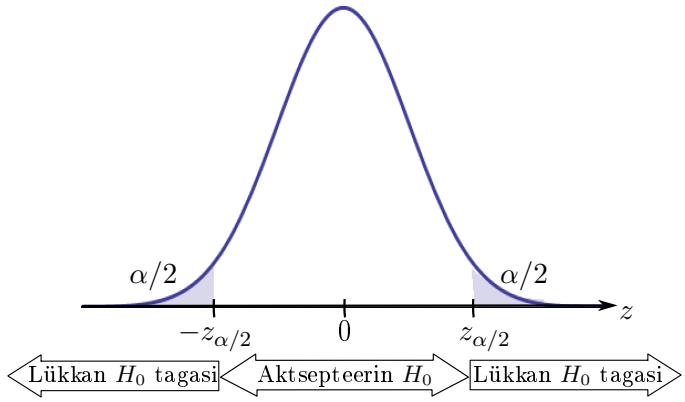
Nullhüpotees on, et kogumi keskväärtus μ võrdub mingi arvuga μ_0:

$$H_0 : \mu = \mu_0. \tag{7.4}$$

Sisukas hüpotees on, et kogumi keskväärtus ei võrdu selle arvuga:

$$H_1 : \mu \neq \mu_0. \tag{7.5}$$

Nullhüpoteesi kehtimise korral on vahe $\bar{x} - \mu_0$ nullilähedane, s.t valimi keskmise $\bar{x}$ ja testitava väärtuse μ_0 vahel statistiliselt oluline erinevus puudub. See erinevus, mis eksisteerib, on tingitud juhuvalimi kasutamisest. On ju selge, et kui 1000 arvu aritmeetiline keskmine on näiteks 500 ja me valime nende hulgast juhuslikult 100 arvu, siis valitud arvude aritmeetiline keskmine ei pruugi olla 500. Kui kehtib nullhüpotees, siis



Joonis 7.4. Nullhüpoteesi vastuvõtmine ja tagasilükkamine keskväärtuse kahepoolsel testimisel, kui olulisuse nivoo on α. Kriitiline piirkond, mille korral H_0 tagasi lükatakse, jaguneb kaheks

kogumist võetud valimite keskmiste standardiseeritud väärtused (7.2) alluvad standardiseeritud normaaljaotusele.

Statistiku z empiirilist väärtust võrreldakse kriitilise väärtusega, milleks on standardiseeritud normaaljaotuse täiendkvantiil $z_{\alpha/2}$. Jooniselt 7.4 on näha, et nullhüpotees lükatakse tagasi, kui $z > z_{\alpha/2}$ või $z < -z_{\alpha/2}$. Need kaks tingimust saab kokku võtta üheks:

$$|z| > z_{\alpha/2}. \tag{7.6}$$

Kahepoolne hüpotees

Viimane tähendab, et kui teststatistiku z väärtus on nullist piisavalt kaugel, kaugemal kui $z_{\alpha/2}$, on nullhüpotees ümber lükatud. Vastupidisel juhul on aga võimalik, et kogumi keskväärtus võrdub arvuga μ_0 ja võtame vastu nullhüpoteesi. Hüpoteesipaari (7.4) ja (7.5) nimetatakse **kahepoolseks hüpoteesiks** (*two-tailed hypothesis*), sest kriitiline piirkond jaguneb kaheks (joonis 7.4).

Enamkasutatavad olulise nivoo α väärtused on 10%, 5% ja 1% ning neile vastavad standardiseeritud normaaljaotuse täiendkvantiilid on esitatud tabelis 7.2.

Tabel 7.2. Statistiku z kriitilised väärtused kahepoolse hüpoteesi korral on standardiseeritud normaaljaotuse täiendkvantiilid $z_{\alpha/2}$

Olulisuse nivoo α	10%	5%	1%
Kriitiline väärtus $z_{\alpha/2}$	1,645	1,96	2,58

Standardiseeritud normaaljaotuse täiendkvantiili leidmiseks tabelarvutuses tuleb kasutada funktsiooni **NORM.S.INV**, kus parameeter *Probability* $= \alpha/2$. See funktsioon väljastab väärtuse $-z_{\alpha/2}$ (vt joonis 7.4) ja positiivse väärtuse saamiseks tuleb ette panna miinusmärk.

Esitame hüpoteesi testimise protseduuri kokkuvõtlikult.

Kogumi keskväärtuse testimine kahepoolse hüpoteesiga
Hüpoteesipaar:

Keskväärtuse testimine, kahepoolne hüpotees

$$\begin{aligned} \mathrm{H}_0: &\quad \mu = \mu_0, \\ \mathrm{H}_1: &\quad \mu \neq \mu_0. \end{aligned}$$

Teststatistik

$$z = \frac{\bar{x} - \mu_0}{se}, \tag{7.7}$$

kus $\bar{x}$ on valimi keskmine ning standardviga

$$se = \frac{s}{\sqrt{n}}. \tag{7.8}$$

Siin on n valimi maht ja s valimi standardhälve.
Olulisuse nivoole α vastav kriitiline väärtus on standardiseeritud normaaljaotuse täiendkvantiil $z_{\alpha/2}$.
Võtta vastu

$$\begin{aligned} &\mathrm{H}_0, \text{ kui } \ |z| \leq z_{\alpha/2}, \\ &\mathrm{H}_1, \text{ kui } \ |z| > z_{\alpha/2}. \end{aligned}$$

Analüüsime, millest sõltub teststatistiku väärtus. Selleks paneme standardvea avaldise (7.8) teststatistiku valemisse (7.7):

$$z = \frac{(\bar{x} - \mu_0)\sqrt{n}}{s}. \tag{7.9}$$

1. Mida rohkem erineb valimi keskmine $\bar{x}$ nullhüpoteesiga püstitatud väärtusest μ_0, seda suurem on z ja seda kergem on nullhüpoteesi ümber lükata.
2. Mida suurem on valimi maht n, seda suurem on z ja seda kergem on nullhüpoteesi ümber lükata.
3. Mida suurem on valimi standardhälve s, seda väiksem on z ja seda raskem on nullhüpoteesi ümber lükata.

Näide 7.1. Duracelli patareide pikkuse testimine

N07Hüpoteesid
N7.1

Duracelli AAA patarei QU2400 tehnilise kirjelduse[a] põhjal on patarei pikkus 44,0 mm. Kontrollimaks, kas tootmisliin on korrektselt seadistatud, võetakse toodangu hulgast juhuvalim mahuga 50 ja mõõdetakse valitud patareide pikkus. Valimi keskmine tuleb 44,25 mm ja standardhälve 1,5 mm. Kontrollida, kas

patareide pikkus vastab etteantud väärtusele. Kasutada olulisuse nivood 0,05.
1. Kasutame z-testi. Hüpoteesipaari püstitamine:

$$H_0 : \mu = 44,$$
$$H_1 : \mu \neq 44.$$

2. Valimi statistilised parameetrid: valimi maht $n = 50$, valimi keskmine $\bar{x} = 44{,}25$ ja valimi standardhälve $s = 1{,}5$. Teststatistiku z empiiriline väärtus valemitest (7.7) ja (7.8):

$$z = \frac{44{,}25 - 44}{\frac{1{,}5}{\sqrt{50}}} = 0{,}17.$$

3. Olulisuse nivoole 0,05 vastav kriitiline väärtus on 1,96.
4. Empiirilise võrdlus kriitilisega: $0{,}17 < 1{,}96$, järelikult teststatistik ei lange kriitilisse piirkonda.
5. Kuna teststatistik ei lange kriitilisse piirkonda, pole alust nullhüpoteesi tagasi lükata. Valimvaatluse tulemus on kooskõlas etteantud väärtusega 44 mm ja tootmisliin on korrektselt seadistatud.

[a] `http://ww2.duracell.com`, *Technical Library*.

Näide 7.2. Üliõpilaste ajakasutus

2005. aasta kevadel Eesti üliõpilaste hulgas läbiviidud uuringust selgus, et keskmiselt kulub üliõpilastel õppetööle 25 tundi nädalas (*Üliõpilaste sotsiaalmajanduslik olukord 2005/2006* 2007). Oletame, et ühe teaduskonna juhtkond soovib teada, kas selles teaduskonnas on üliõpilaste keskmine ajakasutus sama. Selleks viiakse läbi valikuuring, millesse kaasatakse 100 juhuslikult valitud üliõpilast. Küsitletud üliõpilastel kulub nädalas õppetööle keskmiselt 21,2 tundi standardhälbega 18,1 tundi. Kas antud teaduskonnas on üliõpilaste keskmine ajakasutus sama, mis Eesti keskmine?
1. Kasutame z-testi. Hüpoteesipaari püstitamine:

$$\mathrm{H}_0 : \mu = 25,$$
$$\mathrm{H}_1 : \mu \neq 25.$$

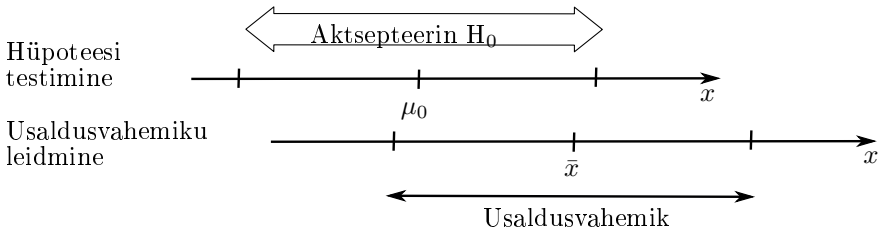
2. Valimi statistilised parameetrid: valimi maht $n = 100$, valimi keskmine $\bar{x} = 21{,}2$ ja valimi standardhälve $s = 18{,}1$. Teststatis-

tiku z empiiriline väärtus valemitest (7.7) ja (7.8)

$$z = \frac{21{,}2 - 25}{\frac{18{,}1}{\sqrt{100}}} = -2{,}10.$$

3. Olulisuse nivoole 0,05 vastav kriitiline väärtus on 1,96.
4. Võrdlemine kriitilisega: $| -2{,}1 | > 1{,}96$, järelikult teststatistik langeb kriitilisse piirkonda.
5. Kuna teststatistik langeb kriitilisse piirkonda, on nullhüpotees ümber lükatud: selles teaduskonnas ei ole üliõpilaste keskmine ajakulu õppetööle 25 tundi nädalas.

Üldkogumi keskväärtuse testimine ja keskväärtuse usalduspiiride leidmine on omavahel seotud. Olgu meil nullhüpoteesiks, et üldkogumi keskväärtus võrdub μ_0. Kui valimi keskmine $\bar{x}$ langeb piirkonda, kus olulisuse nivool α võetakse vastu H_0, siis μ_0 asub valimi põhjal usaldatavusega $\beta = 1 - \alpha$ leitud keskväärtuse usaldusvahemikus (vt joonis 7.5). Hüpoteesi testimisel lähtume nullhüpoteesiga püstitatud väärtusest μ_0, aktsepteerimisvahemik on sellest mõlemale poole ja aktsepteerime nullhüpoteesi, kui valimi keskmine $\bar{x}$ jääb sellesse vahemikku. Keskväärtuse usaldusvahemiku leidmisel lähtume aga valimi keskmisest $\bar{x}$ ning üldkogumi keskväärtus μ langeb antud usaldatavusega sellesse vahemikku.



Joonis 7.5. Hüpoteesi testimine üldkogumi keskväärtuse kohta ja valimi põhjal usaldusvahemiku leidmine on omavahel seotud. Usaldusvahemiku laius on võrdne vahemikuga, kus aktsepteeritakse H_0

7.3. Olulisuse nivoo ja kahte liiki vead

Järelduse tegemiseks vajaliku teststatistiku kriitiline väärtus sõltub uurija poolt ette võetud olulisuse nivoost α, mis näitab, kui suure tõenäosusega me võime väita, et erinevus on oluline. Olulisuse nivoo määramisel tuleb arvestada, et järelduste tegemisel võivad esineda kahte liiki vead.

Kahte liiki vead

Hüpoteesi statistilisel kontrollimisel võib esineda kahte liiki viga:
I liiki viga, kui lükatakse tagasi kehtiv nullhüpotees;
II liiki viga, kui võetakse vastu mittekehtiv nullhüpotees.

Sõltuvalt tegelikkusest ja otsuse vastuvõtmisest realiseerub üks neljast variandist. Kõikvõimalikud variandid on toodud tabelis 7.3.

Tabel 7.3. I ja II liiki viga

	H_0 vastu võetud	H_0 tagasi lükatud
H_0 kehtib	Õige, tõenäosus $1-\alpha$	I liiki viga, tõenäosus α
H_0 ei kehti	II liiki viga, tõenäosus β	Õige, tõenäosus $1-\beta$

Olulisuse nivoo

Olulisuse nivoo α on uurija määratud tõke I liiki vea tõenäosusele.

Kriitilise väärtuse leidmiseks tuleb olulisuse nivoo ette võtta ning see võetakse enamasti kas 10%, 5% või 1%. Kõige sagedamini kasutatakse väärtust 5%.

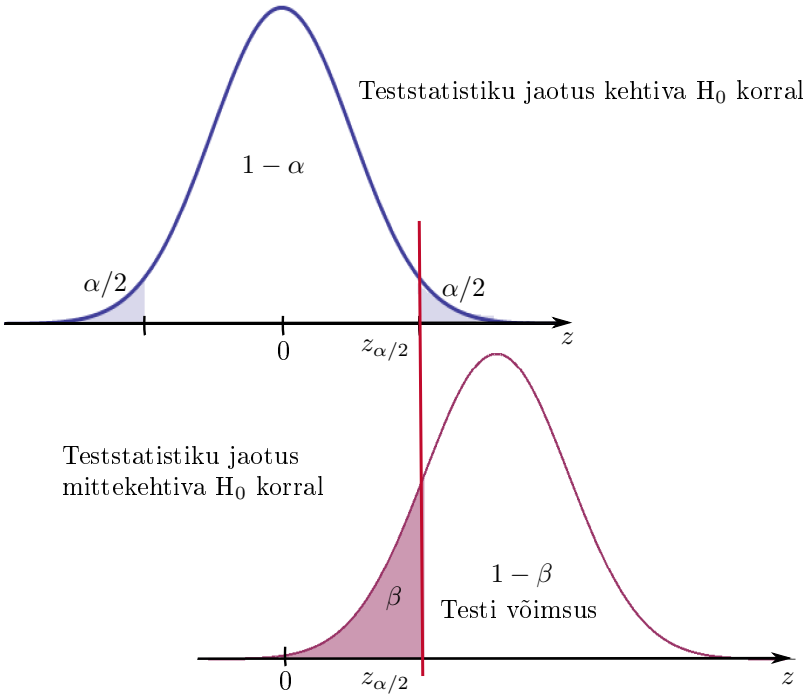
Testi võimsus

Testi võimet lükata tagasi mittekehtiv H_0 nimetatakse testi **võimsuseks** (*power*). Kui II liiki vea tõenäosus on β, siis testi võimsus on $1-\beta$ (vt tabel 7.3 ja joonis 7.6). Järelikult testi võimsus on seda suurem, mida väiksem on II liiki vea tõenäosus.

Erinevat liiki vigade tagajärg on enamasti erineva raskusastmega. Näites 7.1 oli patareide pikkuse mõõtmisel nullhüpoteesiks, et patareide pikkus vastab normile 44,0 mm ja tootmisliin ei nõua seadistamist. Kui tehakse I liiki viga, siis tekib täiendav ajakulu tootmisliini seadistamisele, kuigi liin on korras. Kui aga tehakse II liiki viga, siis läheb müüki toodang, mis ei vasta nõuetele. Ilmselt on II liiki vea tagajärg raskem.

	H_0 võetakse vastu, liini ei seadistata	H_0 lükatakse tagasi, liini seadistatakse
H_0 kehtib, liin ei vaja seadistamist	Otsus õige	I liiki viga
H_0 ei kehti, liin vajab seadistamist	II liiki viga	Otsus õige

Teise näitena vaatame ravimi kliinilist uuringut. Ravimi toime testimisel kasutatakse kahte katseisikute rühma. Ühele rühmale antakse



Joonis 7.6. Teststatistiku jaotus kehtiva ja mittekehtiva nullhüpoteesi korral. Testimisel me ei tea, kumb jaotus esineb

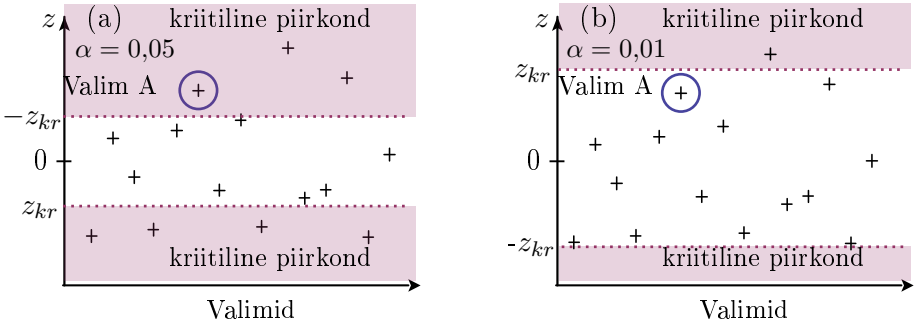
testitavat ravimit, teisele antakse ravimi asemel platseebot ehk tühiravimit, mis maitse- ja lõhnaomadustelt sarnaneb ravimile, kuid ei sisalda uuritavat toimeainet. Viimast rühma nimetatakse kontrollrühmaks. Seejärel võrreldakse katseisikute tervislikke näitajaid kummaski rühmas. Nullhüpoteesiks on, et ravimil toime puudub ning mõlema rühma näitajad on ühesugused.

	H_0 võetakse vastu, järeldus: ravim ei toimi, ravimit müüki ei lasta	H_0 lükatakse tagasi, järeldus:ravim toimib, ravim lastakse müüki
H_0 kehtib, ravim ei toimi	Otsus õige	I liiki viga
H_0 ei kehti, ravim toimib	II liiki viga	Otsus õige

Kui ravim tegelikult toimib, aga eksperimendi tulemustele tuginedes otsustatakse, et ei toimi (erinevused katsealuste ja kontrollgrupi vahel on väikesed), siis seda ravimit müügile ei lasta. Ravimifirma kulutused jäävad kompenseerimata. See on II liiki vea tagajärg.

Kui ravim tegelikult ei toimi, aga testi tulemustele tuginedes otsustatakse, et toimib (I liiki viga), siis lastakse müügile mittetoimiv ravim

ja inimesed ostavad ning tarbivad seda ilmaasjata. Ilmselt on nüüd I liiki viga raskemate tagajärgedega. Et vähendada I liiki vea tõenäosust, peab erinevus ravimit saanute ja kontrollrühma vahel olema piisavalt suur.



Joonis 7.7. Nullhüpotees kehtib. Joonisel (a) lükatakse valimi A korral H_0 tagasi ja tehakse I liiki viga. Joonisel (b) on olulisuse nivood vähendatud, kriitilised väärtused on nullist kaugemal ning sellega I liiki vea tõenäosus väheneb

Sõltuvalt sellest, kumb viga on raskemate tagajärgedega, valitakse olulisuse nivoo suurem või väiksem. Joonisel 7.7 kehtib nullhüpotees. Kui meile satub valim A, siis joonisel 7.7 (a) on see kriitilises piirkonnas, me lükkame kehtiva nullhüpoteesi tagasi ja teeme I liiki vea. Et I liiki vea tõenäosust vähendada, võib kriitilise piirkonna piire nihutada nullist eemale, s.t vähendada olulisuse nivoo väärtust α. Selle tulemusel valim A ei ole enam kriitilises piirkonnas (joonis 7.7 (b)). Nüüd võtame nullhüpoteesi vastu ja otsus on õige.

Aga me **ei tea**, kas nullhüpotees kehtib. Tegelikult võib esineda situatsioon, mis on toodud joonisel 7.8, kus nullhüpotees ei kehti. Kui me vähendasime olulisuse nivood, on nüüd suurem tõenäosus võtta vastu mittekehtiv nullhüpotees, s.t suureneb II liiki vea tõenäosus. Seda on näha ka jooniselt 7.6: väiksema olulisuse nivoo korral nihkub $z_{\alpha/2}$ paremale ning alumisel jaotuskõveral suureneb tõenäosusele β vastav pindala.

Olulisuse nivoo muutmine

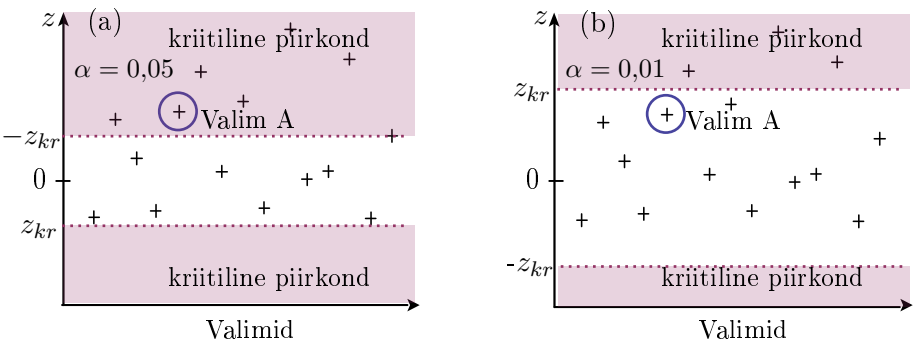
Olulisuse nivoo vähendamine lükkab kriitilise piirkonna nullist kaugemale. Sellega

- väheneb I liiki vea tõenäosus;
- suureneb II liiki vea tõenäosus.

Korraga mõlemat liiki vigade tõenäosust vähendada ei saa. Olulisuse nivoo valik sõltub sellest, kumb on ohtlikum, kas

- kehtiva nullhüpoteesi tagasilükkamine (I liiki viga);
- mittekehtiva nullhüpoteesi vastuvõtmine (II liiki viga).

Praktika on näidanud, et I liiki vead on enamasti raskemate tagajärgedega. Majandusalaste otsuste suhtes loeb tihti see, kumb viga rohkem maksma läheb.



Joonis 7.8. Nullhüpotees ei kehti. Joonisel (a) lükatakse valimi A korral H_0 tagasi ja otsus on õige. Joonisel (b) on olulisuse nivood vähendatud, kriitilised väärtused on nullist kaugemal. Valimi A korral võtame vastu nullhüpoteesi ja teeme II liiki vea

Näide 7.3. Uue kassasüsteemi muretsemine

Kauplusele pakutakse uut kassasüsteemi. Kaubanduse infosüsteemidega tegeleva ettevõtte müügiesindaja väitel vähendab uus süsteem oluliselt ühele ostjale kuluvat aega. Müügiesindaja väite kontrollimiseks otsustab kaupluse juhataja ühte aparaati nädal aega katsetada. Selle aja jooksul mõõdetakse 50 juhuslikult väljavalitud ostja teenindamiseks kulunud aeg. Saadud valimi andmetele tuginedes kontrollitakse müügimehe väite paikapidavust ning kontrollimiseks kasutatakse z-testi.

Nullhüpotees H_0: keskmine ostjate teenindamiseks kulunud aeg on uue kassaaparaadi korral sama, mis vana kassaaparaadi korral (see on varasemast teada), $\mu = \mu_0$.

Sisukas hüpotees H_1: uue kassaaparaadi korral on keskmine aeg erinev $\mu \neq \mu_0$.

Teststatistiku z empiiriliseks väärtuseks saadakse 2,4. Kriitilised väärtused on

olulisuse nivoo 5% korral 1,96;

olulisuse nivoo 1% korral 2,58.

See tähendab, et kasutades olulisuse nivood 5%, tuleb vastu võtta sisukas hüpotees: uus süsteem toimib. Aga olulisuse nivoo 1% juures pole uue süsteemi paremus tõestatud. Mida peab kaupluse juhataja tegema?

Nüüd tuleb lisaks arvestada ka seda, kui palju uus süsteem maksab ja kui palju on rahalisi vahendeid. Kui hind on kõrge ja raha vähe, tuleks kasutada väiksemat olulisuse nivood, kus I liiki vea (ostan olukorda mitteparandava süsteemi) tõenäosus on väiksem. Võib ka suurendada valimit ja testida, kas suurema valimi

korral on nullhüpotees ümber lükatud ja uue süsteemi paremus tõestatud ka nivool 1%.

7.4. Kahepoolne ja ühepoolne hüpotees

Kahepoolsete hüpoteeside korral kontrollitakse seda, kas kogumi keskväärtus **erineb** mingist etteantud väärtusest. Erinevus võib olla mõlemale poole, kriitiline piirkond koosneb kahest osast (joonis 7.4).

Mõnikord on põhjust oodata kindlasuunalist kõrvalekallet standardväärtusest, s.t et kogumi keskväärtus on **suurem** või **väiksem** mingist arvust. Sellisel juhul on tegemist ühepoolse hüpoteesiga. Näites 7.3 kasutati küll kahepoolset hüpoteesi, kuid tegelikult rahuldab uue kassasüsteemi muretsemisel poodi vaid see, kui ostjate teenindamiseks kulunud aeg **väheneb**.

Kui me tahame kontrollida, kas kogumi keskväärtus on **väiksem** mingist arvust μ_0, püstitatakse järgmine hüpoteesipaar:

$$\mathrm{H}_0: \quad \mu \geq \mu_0; \tag{7.10}$$

$$\mathrm{H}_1: \quad \mu < \mu_0. \tag{7.11}$$

Kui me tahame kontrollida, kas kogumi keskväärtus on **suurem** mingist arvust μ_0, püstitatakse hüpoteesid järgmiselt:

$$\mathrm{H}_0: \quad \mu \leq \mu_0; \tag{7.12}$$

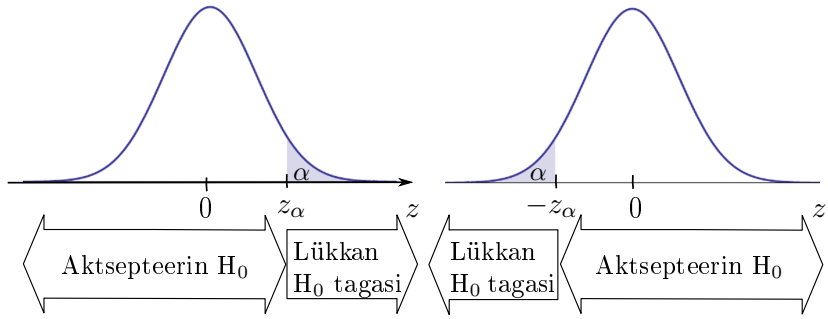
$$\mathrm{H}_1: \quad \mu > \mu_0. \tag{7.13}$$

Ühepoolse hüpoteesi korral on teststatistiku kriitiline piirkond ühel pool nulli, kas paremal või vasakul (joonis 7.9). Seetõttu on kriitiline väärtus nullile lähemal. Teststatistiku kriitiliseks väärtuseks parempoolse hüpoteesi korral on standardiseeritud normaaljaotuse täiendkvantiil z_α ja vasakpoolse hüpoteesi korral $-z_\alpha$.

Tabel 7.4. Statistiku z kriitilised väärtused ühepoolse hüpoteesi korral

Olulisuse nivoo α	10%	5%	1%
Parempoolse hüpoteesi korral z_α	1,28	1,64	2,33
Vasakpoolse hüpoteesi korral $-z_\alpha$	−1,28	−1,64	−2,33

Esitame kogumi keskväärtuse testimise ühepoolse hüpoteesiga kokkuvõtlikult. Vasakpoolseks nimetame situatsiooni, kus tahame kontrollida, kas kogumi keskväärtus on väiksem mingist arvust, sest siis asub kriitiline piirkond vasakul. Parempoolne hüpoteesipüstitus on siis, kui tahame kontrollida, kas kogumi keskväärtus on suurem mingist arvust.



Joonis 7.9. Kriitiline piirkond ühepoolse hüpoteesi korral. Vasakul H_0: $\mu \leq \mu_0$ ja H_1: $\mu > \mu_0$, paremal H_0: $\mu \geq \mu_0$ ja H_1: $\mu < \mu_0$

Keskväärtuse testimine, ühepoolne hüpotees

Kogumi keskväärtuse testimine ühepoolse hüpoteesiga

1. Hüpoteesipaar:

vasakpoolne		parempoolne	
H_0 :	$\mu \geq \mu_0$	H_0 :	$\mu \leq \mu_0$,
H_1 :	$\mu < \mu_0$	H_1 :	$\mu > \mu_0$.

2. Teststatistik

$$z = \frac{\bar{x} - \mu_0}{se}, \tag{7.14}$$

kus $\bar{x}$ on valimi keskmine ning standardviga

$$se = \frac{s}{\sqrt{n}}. \tag{7.15}$$

Siin on n valimi maht ja s valimi standardhälve.

3. Olulisuse nivoole α vastav kriitiline väärtus on standardiseeritud normaaljaotuse täiendkvantiil z_α. Vasakpoolse hüpoteesi korral on kriitiline väärtus täiendkvantiili vastandväärtus $-z_\alpha$.

4. Võtta vastu

H_0, kui	$z \geq -z_\alpha$,	$z \leq z_\alpha$,
H_1, kui	$z < -z_\alpha$,	$z > z_\alpha$.

Nii kahepoolse hüpoteesi kui ka mõlema ühepoolse hüpoteesi korral on nullhüpotees ümber lükatud, kui teststatistiku empiiriline väärtus on **nullist kaugemal** kui kriitiline väärtus.

Järgnevalt vaatame kaht näidet ühepoolse hüpoteesi kasutamise kohta.

Näide 7.4. Muudatus tehnoloogias ja patarei keskmine kasutusiga

Patareide tootja on kindlaks teinud, et patarei keskmine kasutusiga on 299 tundi. Peale tehnoloogiliste muudatuste tegemist patareide täitmisel soovib tootja kontrollida, kas kasutusiga on pikenenud. Selleks valitakse juhuslikult välja 200 patareid ja testitakse neid. Väljavalitud patareide keskmine kasutusiga oli 300,55 tundi standardhälbega 8,28 tundi.
1. Kasutame ühepoolset z-testi.
2. Hüpoteesipaari püstitamine:

$$\mathrm{H}_0 : \mu \leq 299,$$
$$\mathrm{H}_1 : \mu > 299.$$

3. Valikvaatluse põhjal saadud statistilised parameetrid: valimi maht $n = 200$, aritmeetiline keskmine $\bar{x} = 300{,}55$ ja standardhälve $s = 8{,}28$. Teststatistiku empiiriline väärtus

$$z = \frac{300{,}55 - 299}{\frac{8{,}28}{\sqrt{200}}} = 2{,}65.$$

4. Olulisuse nivoole 5% vastav kriitiline väärtus on 1,645.
5. Võrdlus kriitilisega: $2{,}65 > 1{,}645$, teststatistik langeb kriitilisse piirkonda.
6. Kuna teststatistik langeb kriitilisse piirkonda, tuleb nullhüpotees tagasi lükata. Vaatluse tulemus on kooskõlas väitega, et patareide keskmine kasutusiga pikenes.

Näide 7.5. Maisihelveste paki kaal

N07Hüpoteesid
N7.5,6

Poes on müügil maisihelbed 250-grammises pakis. Tarbijakaitseameti esindaja soovib kontrollida, kas pakendil märgitud kaal vastab tegelikkusele. Selleks kaalub ta ära 50 juhuslikult valitud pakki. Nende pakkide keskmine kaal on 248,14 grammi standardhälbega 18,51 grammi. Kontrollida olulisuse nivool 5%, kas pakkides on keskmiselt maisihelbeid vähem kui pakendile märgitud või mitte.
1. Kasutame ühepoolset z-testi.

2. Hüpoteesi püstitamine:

$$\mathrm{H}_0 : \mu \geq 250,$$
$$\mathrm{H}_1 : \mu < 250.$$

3. Valimi statistilised parameetrid: valimi maht $n = 50$, aritmeetiline keskmine $\bar{x} = 248{,}14$ ja standardhälve $s = 18{,}51$. Statistiku z empiiriline väärtus

$$z = \frac{248{,}14 - 250}{\frac{18{,}51}{\sqrt{50}}} = -0{,}71.$$

4. Olulisuse nivoole 5% vastav kriitiline väärtus on 1,64.
5. Võrdlus kriitilisega: $-0{,}71 > -1{,}64$, järelikult teststatistik ei lange kriitilisse piirkonda.
6. Kuna teststatistik ei lange kriitilisse piirkonda, tuleb vastu võtta nullhüpotees. Kontrolli tulemus ei kinnita väidet, et maisihelbeid on pakkides vähem kui 250 grammi.

Tootja risk, tarbija risk

Ühepoolseid hüpoteese kasutatakse tihti juhul, kui tahetakse hinnata tootja riski või tarbija riski. Tegeliku koguse kõrvalekaldumisele normist reageerivad tootja ja tarbija erinevalt. Kui näiteks pakendis on rohkem kui 250 grammi, kannatab kahju tootja, sest tarbija maksab vaid 250 eest. Seega ei tohi tootja seisukohast maisihelveste kogus pakis olla suurem kui 250, võib olla väiksem ning **tootja** kontrollib järgmist hüpoteesipaari:

$$\mathrm{H}_0 : \mu \leq 250,$$
$$\mathrm{H}_1 : \mu > 250.$$

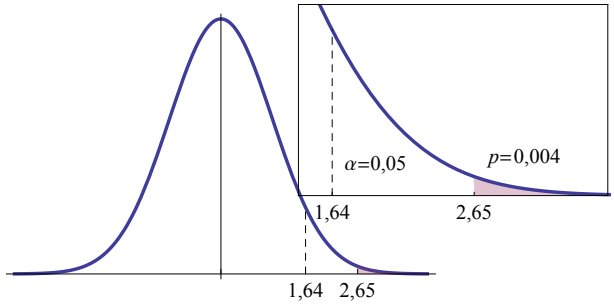
Tarbija seisukohast ei tohi kogus olla väiksem kui 250 grammi, võib olla suurem, siis on tarbija rahul. **Tarbija** peab kontrollima sellist hüpoteesipaari:

$$\mathrm{H}_0 : \mu \geq 250,$$
$$\mathrm{H}_1 : \mu < 250.$$

7.5. Olulisuse tõenäosus

Näites 7.4 patareide kasutusea testimise kohta leidsime, et z-testi empiiriliseks väärtuseks tuli 2,65, aga kriitiline väärtus oli olulisuse nivoo 0,05 korral 1,64. Kui madal võiks olla olulisuse nivoo, et antud andmete põhjal saaks veel vastu võtta sisuka hüpoteesi? See tähendab,

milline tõenäosus vastab väärtusele 2,65? Standardiseeritud normaaljaotusest saame leida, et täiendkvantiilile $z_p = 2{,}65$ vastab $p = 0{,}004$. Joonisel 7.10 on kriitilisele väärtusele 1,64 vastav tõenäosus $\alpha = 0{,}05$ ja empiirilisele väärtusele 2,65 vastav tõenäosus $p = 0{,}004$ (varjutatud piirkond).



Joonis 7.10. Varjutatud ala pindala on teststatistiku empiirilisele väärtusele 2,65 vastav olulisuse tõenäosus $p = 0{,}004$

Olulisuse tõenäosus

Olulisuse tõenäosus on väikseim olulisuse nivoo, mis antud valimi põhjal lubab vastu võtta sisuka hüpoteesi. See näitab antud valimi sobivust nullhüpoteesiga.

Teststatistiku valimjaotusest on võimalik leida siis kaks tõenäosust:

- olulisuse **nivoo** α — teststatistiku **kriitilisele** väärtusele vastav tõenäosus;
- olulisuse **tõenäosus** p — teststatistiku **empiirilisele** väärtusele vastav tõenäosus.

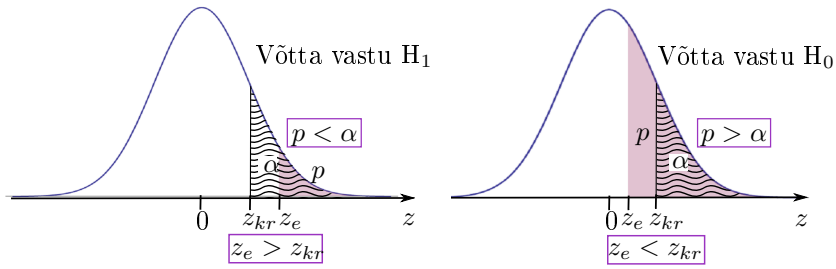
Mida väiksem on olulisuse tõenäosus, seda väiksem on tõenäosus, et kehtib nullhüpotees ja seda kindlamini on sisukas hüpotees tõestatud. Ingliskeelses terminoloogias kasutatakse selle märkimiseks terminit *p-value* või *significance probability*. Analüüsi tulemuste esitamisel tuuakse ära olulisuse tõenäosuse arvuline väärtus, mis võimaldab lugejal otsustada, kui kindlalt on sisukas hüpotees tõestatud või kui palju jäi olulisuse nivoost puudu. Testide läbiviimisel leiab enamik statistikapakette olulisuse tõenäosuse.

Olulisuse tõenäosuse kasutamine

Järelduse tegemiseks võrreldakse olulisuse tõenäosust p olulisuse nivooga α. Kui

$p \geq \alpha$, võetakse vastu nullhüpotees;
$p < \alpha$, võetakse vastu sisukas hüpotees.

Kaks meetodit: olulisuse tõenäosuse võrdlemine olulisuse nivooga ja teststatistiku empiirilise väärtuse võrdlemine kriitilise väärtusega on ekvivalentsed. Jooniselt 7.11 näeme, et kui $p < \alpha$, siis teststatistiku empiiriline väärtus z_e on nullist kaugemal kui kriitiline väärtus z_{kr} ja kehtib nullhüpotees (vasakul). Kui aga $p > \alpha$, siis empiiriline väärtus z_e on nullile lähemal kui kriitiline väärtus ja kehtib sisukas hüpotees (paremal). Seos on lihtne: mida väiksem on mingist väärtusest paremale jääva kõveraaluse piirkonna pindala, seda kaugemal see väärtus nullist on.



Joonis 7.11. Olulisuse tõenäosuse p võrdlemine olulisuse nivooga α on ekvivalentne teststatistiku empiirilise väärtuse z_e ja kriitilise väärtuse z_{kr} võrdlemisega. Lainelise mustriga ala pindala on kriitilisele väärtusele vastav olulisuse nivoo α ja varjutatud ala pindala on empiirilisele väärtusele vastav olulisuse tõenäosus p

See, kas kasutada empiirilise väärtuse võrdlemist kriitilisega või olulisuse tõenäosuse võrdlemist olulisuse nivooga, sõltub sellest, kas meie käsutuses olev tarkvara võimaldab leida olulisuse tõenäosust p või mitte. Kui võimaldab, siis on olulisuse tõenäosuse kasutamine lihtsam.

Tabelarvutuses leiab ühepoolse z-testi olulisuse tõenäosuse p funktsioon **Z.TEST**(*Array*; X; *Sigma*). *Array* on andmete piirkond ja X nullhüpoteesiga püstitatud väärtus μ_0. Funktsiooni kolmas parameeter *Sigma* on mõeldud juhuks, kui kogumi standardhälve σ on teada. Kui see teada pole, tuleb *Sigma* ära jätta: kogumi standardhälbe hinnanguks kasutatakse valimi standardhälvet, mis arvutatakse funktsiooni kasutamisel automaatselt. Funktsioon Z.TEST leiab valimi aritmeetilise keskmise ja standardhälbe, arvutab valemite (7.14) ja (7.15) põhjal statistiku z ning seejärel leiab standardiseeritud normaaljaotusest sellele vastava olulisuse tõenäosuse p.

Kahepoolse hüpoteesi korral tuleb funktsiooniga Z.TEST leitud olulisuse tõenäosust võrrelda suurusega $\alpha/2$, kus α on olulisuse nivoo.

N07Hüpoteesid N7.5,6

Näide 7.6. Olulisuse tõenäosuse kasutamine hüpoteesi kontrollimisel

Näites 7.5 testisime, ega maisihelbepakkides pole helbeid vähem kui 250 grammi. Viime selle testimise uuesti läbi, kasutades olulisuse tõenäosust.
1. Kasutame ühepoolset z-testi.
2. Hüpoteesipaar:

$$\mathrm{H}_0 : \mu \geq 250,$$
$$\mathrm{H}_1 : \mu < 250.$$

3. Vaatlusandmete ja nullhüpoteesiga püstitatud väärtuse 250 põhjal leiame olulisuse tõenäosuse, kasutades tabelarvutuses funktsiooni Z.TEST. Tulemuseks saame, et olulisuse tõenäosus $p = 0{,}76$.
4. Olulisuse nivoo võtame 0,05.
5. Võrdleme olulisuse tõenäosust olulisuse nivooga: $0{,}76 > 0{,}05$, järelikult teststatistik ei lange kriitilisse piirkonda ning tuleb vastu võtta nullhüpotees.
6. Kontrolli tulemus ei kinnita väidet, et maisihelbeid on pakkides keskmiselt vähem kui 250 grammi.

7.6. Väike valim ja keskväärtuse testimine t-testiga

Alapeatükis 6.6 nägime, et väikeste valimite ($n < 30$) korral alluvad valimite keskmiste standardiseeritud väärtused t-jaotusele ehk Studenti jaotusele. Seepärast tuleb väikeste valimite korral kasutada kogumi keskväärtuse testimiseks ***t*-testi** ehk **Studenti testi**.

Hüpoteeside püstitamine ja nende kontrollimise metoodika on sama, mis z-testi korral. Valem teststatistiku t empiirilise väärtuse leidmiseks on ka samasugune, mis z-testi korral. Erinevus z-testist on vaid kriitilistes väärtustes: t-testi kriitiline väärtus on t-jaotuse täiendkvantiil, mis sõltub lisaks olulisuse nivoole ka vabadusastmete arvust ν.

Kogumi keskväärtuse testimine väikese valimi korral

1. Hüpoteesipaar:

kahepoolne	vasakpoolne	parempoolne
$H_0 : \mu = \mu_0$	$H_0 : \mu \geq \mu_0$	$H_0 : \mu \leq \mu_0$
$H_1 : \mu \neq \mu_0$	$H_1 : \mu < \mu_0$	$H_1 : \mu > \mu_0$.

2. Teststatistiku empiiriline väärtus

$$t = \frac{\bar{x} - \mu_0}{se}, \tag{7.16}$$

kus standardviga

$$se = \frac{s}{\sqrt{n}}, \tag{7.17}$$

μ_0 on nullhüpoteesiga püstitatud keskväärtus, n valimi maht, $\bar{x}$ valimi keskmine ja s valimi standardhälve.

3. Olulisuse nivoole α vastav kriitiline väärtus on t-jaotuse täiendkvantiil $t_{\alpha/2}(\nu)$ kahepoolse hüpoteesi korral ja $t_\alpha(\nu)$ ühepoolse hüpoteesi korral, kus vabadusastmete arv

$$\nu = n - 1. \tag{7.18}$$

4. Võtta vastu

H_0, kui	$\lvert t\rvert \leq t_{\alpha/2}(\nu)$,	$t \geq -t_\alpha(\nu)$,	$t \leq t_\alpha(\nu)$,
H_1, kui	$\lvert t\rvert > t_{\alpha/2}(\nu)$,	$t < -t_\alpha(\nu)$,	$t > t_\alpha(\nu)$.

Tabelarvutuses kasutatakse t-testi kriitilise väärtuse leidmiseks ühepoolse hüpoteesi korral funktsiooni **T.INV**. Ette on vaja anda olulisuse nivoo α (*Probability*) ja vabadusastmete arv ν (*Deg_freedom*). Kahepoolse (*two-tail*) hüpoteesi korral kasutatakse funktsiooni **T.INV.2T**, mille korral samuti *Probability*$= \alpha$.

Näide 7.7. Auto bensiinikulu testimine

N07Hüpoteesid
N7.7

Reklaamis väidetakse, et Suzuki Baleno tarbib maanteesõidul bensiini 6,8 liitrit 100 km kohta. Autoomanik soovib kontrollida, ega tema auto ökonoomsus pole lubatust väiksem. Viiel katsel suutis ta 50 liitriga läbida vastavalt 690, 750, 645, 730 ja 740 kilomeetrit. Nende katsete keskmisena tuleb kütusekuluks 7,05 liitrit 100 km kohta standardhälbega 0,45. Kas auto ökonoomsus vastab reklaamis toodule?

1. Hüpoteesi kontrollimiseks kasutame ühepoolset t-testi.

2. Hüpoteesipaar:

$$\mathrm{H}_0 : \mu \leq 6{,}8,$$
$$\mathrm{H}_1 : \mu > 6{,}8.$$

3. Valemitest (7.16) ja (7.17) leiame teststatistiku t väärtuse

$$t = \frac{7{,}05 - 6{,}8}{\frac{0{,}45}{\sqrt{5}}} = 1{,}27.$$

4. Olulisuse nivooks võtame 5%. Vabadusastmete arv on

$$\nu = n - 1 = 5 - 1 = 4.$$

Kriitiliseks väärtuseks on t-jaotuse täiendkvantiil $t_{0{,}05}(4) = 2{,}13$.
5. Võrdleme teststatistiku empiirilist väärtust kriitilisega: $1{,}27 < 2{,}13$, teststatistik ei lange kriitilisse piirkonda.
6. Kuna teststatistiku väärtus ei lange kriitilisse piirkonda, pole alust nullhüpoteesi tagasi lükata. Järelikult võib väita, et auto ökonoomsus vastab reklaamis toodule.

Kuna suurte valimite korral läheb t-jaotus üle normaaljaotuseks, siis võib kõikide valimite korral kasutada t-testi.

z-testi ja t-testi erinevus seisneb vaid kriitilistes väärtustes väikeste valimite korral.

Seepärast räägitakse järgmistes alapeatükkides vaid t-testist. Ka statistikapakettides ei ole z-testi valikut, leiab ainult t-testi.

7.7. Kahe kogumi keskväärtuse t-test ja sõltumatud valimid

Praktikas esineb sagedamini olukordi, kus meil on kaks valimit ning nende abil tuleb võrrelda mingi tunnuse keskväärtusi kahes kogumis. Siis on tegemist kahe keskväärtuse võrdlemisega.

Sõltumatud ja sõltuvad valimid

Põhimõtteliselt võib esineda kaks võimalust:

- tuleb võrrelda erinevaid objekte ning nendest on moodustatud kaks **sõltumatut** valimit;
- mõlemas valimis on ühed ja samad objektid, aga objektide tunnuse väärtus muutub mingi tegevuse tulemusel ning iga objekti

jaoks on kaks väärtust, üks ühes valimis, teine teises valimis: **sõltuvad valimid**.

Sõltumatute valmite (*indenpendent samples*) korral leitakse eraldi mõlema valimi keskmine ja standardhälve. Seejärel testitakse, kas valimite keskmised on oluliselt erinevad või mitte. Nullhüpoteesile vastab erinevuse puudumine.

Näide 7.8. Sotsiaal-majanduslik erinevus Euroopa riikide vahel

2007. aastal ajakirjas International Advances in Economic Research ilmunud artiklis võrreldi 45 sotsiaal-majanduslikku näitajat erinevates Euroopa riikides (Niroomand ja Nissan, 2007). Riigid olid jagatud gruppidesse. Gruppi A kuulusid 15 Euroopa Liidu riiki seisuga 1995. Gruppi B kuulusid kandidaatriigid 1997. aastal: Eesti, Küpros, Poola, Sloveenia, Tšehhi ja Ungari. Tabelis on toodud andmed mõnede näitajate kohta: keskmine kummaski riigigrupis ning keskväärtuste testimisel leitud t-statistik. Statistiku kriitiline väärtus on 1,73 ja tärniga on märgitud statistiliselt oluline erinevus.

Näitaja	Keskmine A	Keskmine B	t-statistik	
Kulutused haridusele (% RKP-st) 1995–1997	5,5	4,6	−0,38	
Kulutused haridusele (% valitsuskuludest) 1995–1997	10,7	16,1	−2,68	*
Oodatav eluiga sünnimomendil	77,4	73,4	4,74	*
Imikute suremus (1000 sünni kohta) 1999	5,0	9,0	−3,14	*
SKP elaniku kohta (PPP 1000 USD)	23,7	12,7	3,94	*
SKP kasv aastas (%) 1990–1999	2,1	2,0	0,22	
Keskmine THI muutus (%) 1990–1999	3,1	19,2	−5,9	*

Olgu meil kaks kogumit keskväärtustega μ_1 ja μ_2. Kui soovime testida, kas keskväärtused on võrdsed või mitte, siis nullhüpoteesiks on, et erinevus puudub (kahepoolne hüpotees):

Kahepoolne hüpotees

$$\mathrm{H}_0: \quad \mu_1 = \mu_2. \tag{7.19}$$

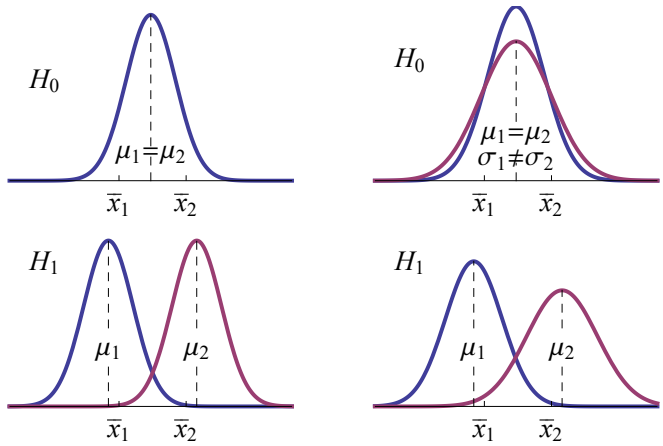
Sisukas hüpotees on, et kogumite keskväärtused ei ole võrdsed:

$$\mathrm{H}_1: \quad \mu_1 \neq \mu_2. \tag{7.20}$$

Hüpoteesipaari (7.19), (7.20) testimiseks on meil üks valim mahuga n_1, keskmisega $\bar{x}_1$ ja standardhälbega s_1 ning teine valim mahuga n_2,

keskmisega $\bar{x}_2$ ja standardhälbega s_2. Nullhüpoteesile ja sisukale hüpoteesile vastavad situatsioonid on toodud joonisel 7.12. Vasakpoolsel ülemisel paneelil on valimkeskmiste valimjaotus ning kahe valimi keskmised kehtiva H_0 korral: valimid tulevad ühest ja samast kogumist. Vasakpoolsel alumisel paneelil on valimite keskmiste valimjaotused ning kahe valimi keskmised kehtiva H_1 korral: valimid tulevad erinevatest kogumitest. Aga kehtiva nullhüpoteesi korral võib esineda ka situatsioon, kus kogumite keskväärtused on võrdsed, aga standardhälbed erinevad. Selline situatsioon on kujutatud ülemisel parempoolsel paneelil.

Otsustamiseks, kumb hüpotees kehtib, tuleb valimite alusel leida vastav teststatistik. Aga enne on vaja kindlaks teha, kas valimid tulevad sama või erineva standardhälbega kogumitest, sest teststatistiku leidmise valem sõltub sellest.



Joonis 7.12. Üleval vasakul kehtib H_0 ning valimid tulevad ühest ja samast kogumist. All vasakul kehtib H_1 ning valimid tulevad erinevate keskväärtustega kogumitest. Üleval paremal kehtib samuti H_0, kuid valimid tulevad erineva standardhälbega kogumitest

t-statistiku valem sõltub tunnuse hajumisest kogumites

See, mismoodi t-testi statistikut leida, sõltub sellest, kas tunnuse hajumine on kummaski kogumis ühesugune või erinev.

Valimite standardhälbeid lihtsalt võrreldes ei pruugi me õigesti otsustada, kas hajumine kogumites on ühesugune või mitte. Nii nagu valimi keskmine erineb kogumi keskväärtusest, nii ka valimi dispersioon erineb kogumi dispersioonist. Ka ühest ja samast üldkogumist võetud valimitel on dispersioon erinev. Seepärast tuleb korrektse tulemuse saamiseks viia eelnevalt läbi dispersioonide testimine, mida vaatame alapeatükis 7.9.

Kui kogumite **dispersioonid on erinevad**, siis kasutatakse teststatistiku arvutamiseks valemit

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(\nu_{WS}). \tag{7.21}$$

Erinev dispersioon, Welchi test

Siin on n_1, $\bar{x}_1$ ja s_1^2 valimi 1 maht, keskmine ja dispersioon ning n_2, $\bar{x}_2$ ja s_2^2 vastavalt valimi 2 maht, keskmine ja dispersioon. Efektiivne vabadusastmete arv ν_{WS} leitakse Welchi-Satterthwaite'i valemist (Satterthwaite, 1946; Welch, 1951):

$$\nu_{WS} = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \tag{7.22}$$

ja ümardatakse täisarvuni. Teststatistiku (7.21) kasutamist nimetatakse ka **Welchi testiks**.

Kui mõlema kogumi **dispersioon on ühesugune**, siis valemis (7.21) asendatakse valimite dispersioonid s_1^2 ja s_2^2 ühendatud dispersiooni (*pooled variance*) hinnanguga s_p^2. Selleks on valimite dispersioonide kaalutud aritmeetiline keskmine

Ühesugune dispersioon, Studenti test

$$s_p^2 = \frac{\nu_1 s_1^2 + \nu_2 s_2^2}{\nu_1 + \nu_2}, \tag{7.23}$$

kus kaaludeks on valimite vabadusastmete arvud $\nu_1 = n_1 - 1$ ja $\nu_2 = n_2 - 1$. Sellisel juhul saadakse teststatistiku jaoks valem

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t(n_1 + n_2 - 2). \tag{7.24}$$

Teststatistik (7.24) allub t-jaotusele vabadusastmete arvuga, mis on valimite vabadusastmete summa;

$$\nu = \nu_1 + \nu_2 = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2. \tag{7.25}$$

Seda testi nimetatakse ka **Studenti testiks**.

Teststatistiku kriitiline väärtus on mõlemal juhul vastava vabadusastmete arvuga t-jaotuse täiendkvantiil $t_{\alpha/2}(\nu)$, kus α on olulisuse tõenäosus ja ν leitakse kas valemist (7.25) või (7.22). Kui aga mõlemad valimid on suured, siis t-jaotus läheneb normaaljaotusele ning kriitiliseks väärtuseks võib võtta normaaljaotuse täiendkvantiili $z_{\alpha/2}$.

Kriitilised väärtused

Näide 7.9. Naissoost juhtide tööstaaži võrdlus väike- ja suurettevõtetes

Ajakirjas American Journal of Small Business ilmus 1988. aastal ülevaade uuringust, mille käigus küsitleti naissoost juhte väike-ettevõtetes (alla 100 töötajaga) ja suurettevõtes (R. L. Anderson ja K. P. Anderson, 1988). Võrreldi väike- ja suurettevõtetes töötavate naisjuhtide vanust, tööstaaži, sissetulekut ning edutamiste arvu viimase kolme aasta jooksul. Väikeettevõtetest saadi tagasi 86 küsitluslehte, suurettevõtetest 91 küsitluslehte.
Väikeetevõtetes oli naisjuhtide keskmine tööstaaž 4,8 aastat standardhälbega 4,6 aastat. Suurettevõtetes oli keskmine tööstaaž 5,3 aastat standardhälbega 5,2 aastat. Testime, kas naisjuhtide keskmine tööstaaž on väike- ja suurettevõtetes erinev. Kuna standardhälbed ei erine väga palju, eeldame, et kogumite dispersioonid on ühesugused.
1. Kasutame kahepoolset t-testi kahe kogumi keskväärtuse võrdlemiseks, kui dispersioonid on võrdsed.
2. Hüpoteesipaar:

$$\begin{aligned} \mathrm{H}_0 &: \mu_1 = \mu_2, \\ \mathrm{H}_1 &: \mu_1 \neq \mu_2. \end{aligned}$$

3. Leiame t-statistiku empiirilise väärtuse. Olgu suurettevõtted valim 1 ja väikeettevõtted valim 2. Ühine dispersioon valemist (7.23)

$$s_p^2 = \frac{(91-1)\cdot 5{,}2^2 + (86-1)\cdot 4{,}6^2}{91+86-2} = 24{,}184$$

ja t-statistik valemist (7.24)

$$t = \frac{5{,}3 - 4{,}8}{\sqrt{24{,}184 \cdot \left(\frac{1}{91} + \frac{1}{86}\right)}} = 0{,}676.$$

4. Kuna tegemist on suurte valimitega, võtame kriitilise väärtuse normaaljaotusest. Olulisuse nivoo 0,05 ja kahepoolse hüpoteesi korral on see 1,96.
5. Võrdleme t-statistiku empiirilist ja kriitilist väärtust: $0{,}676 < 1{,}96$.
6. Kuna statistik ei lange kriitilisse piirkonda, võtame vastu nullhüpoteesi. Naisjuhtide keskmine tööstaaž väike- ja suurettevõtetes ei ole erinev.

Näide 7.10. Naissoost juhtide sissetuleku võrdlus väike- ja suurettevõtetes

Eelmises näites viidatud uurimuses võrreldi erinevaid tunnuseid väike- ja suurettevõtetes töötavate naisjuhtide korral. Sissetuleku kohta saadi järgmised andmed: suurettevõtetes oli naisjuhtide keskmine sissetulek 19 497 dollarit aastas standardhälbega 6118 $. Väikeettevõtetes oli keskmine sissetulek 19 870 $ aastas standardhälbega 8756 $. Kas naisjuhtide keskmine sissetulek on väike- ja suurettevõtetes erinev?
Kuna standardhälbed on väga erinevad, siis ilmselt pole õige kasutada lähenemist, mis eeldab ühesuguseid dispersioone.
1. Kasutame kahepoolset t-testi kahe kogumi keskväärtuse võrdlemiseks, kui dispersioonid on erinevad.
2. Hüpoteesipaar:

$$\mathrm{H}_0 : \mu_1 = \mu_2,$$
$$\mathrm{H}_1 : \mu_1 \neq \mu_2.$$

3. Leiame t-statistiku empiirilise väärtuse. Valim 1 on suurettevõtted ja valim 2 väikeettevõtted. Valimite mahud olid $n_1 = 91$ ja $n_2 = 86$. t-statistik valemist (7.21)

$$t = \frac{19497 - 19870}{\sqrt{\frac{6118^2}{91} + \frac{8756^2}{86}}} = -0{,}327.$$

4. Kuna tegemist on suurte valimitega, siis kriitilise väärtuse võtame normaaljaotusest. Usaldatavuse 0,05 ja kahepoolse hüpoteesi korral on see 1,96.
5. Empiirilise väärtuse võrdlus kriitilisega: $|-0{,}327| < 1{,}96$.
6. Kuna statistik ei lange kriitilisse piirkonda, võtame vastu nullhüpoteesi. Naisjuhtide keskmine sissetulek väike- ja suurettevõtetes ei ole erinev.

Kui mõlema valimi mahud on võrdsed

$$n_1 = n_2 = n, \tag{7.26}$$

Võrdse mahuga valimid

siis valemist (7.23) ühendatud dispersioon

$$s_p^2 = \frac{s_1^2 + s_2^2}{2} \tag{7.27}$$

ning t-statistik ühesuguste dispersioonide korral (7.24) ja t-statistik erinevate dispersioonide korral (7.21) tulevad ühesugused:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2 + s_2^2}{n}}}. \tag{7.28}$$

Vabadusastmete arv on aga erinev. Ühesuguste dispersioonide korral valemist (7.25)

$$\nu = 2n - 2 \tag{7.29}$$

ja erinevate dispersioonide korral valemist (7.22)

$$\nu_{WS} = (n-1)\frac{\left(s_1^2 + s_2^2\right)^2}{s_1^4 + s_2^4}. \tag{7.30}$$

Kuna vabadusastmete arv on erinev, siis ka $t(\nu)$-jaotusest leitavad kriitilised väärtused tulevad erinevad.

Olgu meil kaks valimit mahuga $n = 6$, mis on toodud tabelis 7.5. Testime, kas need on ühesuguse keskväärtusega kogumitest või mitte. Kasutame mõlemat, nii ühesuguse kui ka erineva dispersiooniga kogumite meetodit.

Tabel 7.5. Kaks ühesuguse mahuga valimit

	Valimite elemendid						$\bar{x}$	s
Valim 1	6	8	10	12	18	20	12,33	5,574
Valim 2	12	18	20	28	33	55	27,67	15,32

Kuna valimite mahud on ühesugused, siis t-statistiku leidmiseks võib kasutada valemit (7.28):

$$t = \frac{12{,}33 - 5{,}574}{\sqrt{\frac{5{,}574^2 + 15{,}33^2}{6}}} \approx -2{,}30.$$

Kui võtame olulisuse nivooks $\alpha = 0{,}05$, siis seda statistikut tuleb võrrelda t-jaotuse täiendkvantiiliga $t_{0{,}025}(\nu)$.

1. Kui eeldame, et kogumite dispersioonid on võrdsed, siis valemist (7.29) $\nu = 10$, mille korral täiendkvantiil $t_{0{,}025}(10) = 2{,}23$. Kuna $|-2{,}3| > 2{,}23$, lükkame H_0 tagasi ja võtame vastu H_1.
2. Kui eeldame, et kogumite dispersioonid on erinevad, siis valemist (7.30) $\nu_{WS} = 6{,}30 \approx 6$, mille korral täiendkvantiil $t_{0{,}025}(6) = 2{,}45$. Kuna $|-2{,}3| < 2{,}45$, võtame vastu H_0.

Nägime, et kui kasutame võrdsete dispersioonide meetodit, aga tegelikult kogumite dispersioonid ei ole võrdsed, siis võib juhtuda, et lükkame

tagasi kehtiva nullhüpoteesi. See on I liiki viga. Kui dispersioone pole testitud, siis keskväärtuste testimisel tuleb kasutada erinevate dispersioonide meetodit ehk Welchi testi. Sellega me vähendame I liiki vea tõenäosust.

Kui valimite maht n on suur, läheneb t-jaotus standardiseeritud normaaljaotusele ning kriitilise väärtuse võib leida viimasest. Järelikult pole suure ühesuguse mahuga valimite korral oluline, kas kogumite dispersioonid on võrdsed või mitte. t-statistik leitakse valemist (7.28) ning selle absoluutväärtust võrreldakse standardiseeritud normaaljaotuse täiendkvantiiliga $z_{\alpha/2}$.

Kui soovime tõestada, et kogumis 1 on tunnuse keskväärtus suurem kui kogumis 2, kasutame ühepoolset hüpoteesipaari:

Ühepoolne hüpotees

$$\begin{aligned} &\mathrm{H}_0 : \mu_1 \leq \mu_2, \\ &\mathrm{H}_1 : \mu_1 > \mu_2. \end{aligned} \tag{7.31}$$

Kui tahame kontrollida, kas ühes kogumis on keskväärtus väiksem kui teises kogumis, on see ekvivalentne hüpoteesipaariga (7.31). Suurema keskväärtusega kogum tuleb võtta kogumiks number 1.

Teststatistiku kriitiline väärtus on ühepoolse hüpoteesi korral vastava vabadusastmete arvuga t-jaotuse täiendkvantiil $t_\alpha(\nu)$, kus α on olulisuse nivoo ja ν leitakse kas valemist (7.25) või (7.22). Testimise tulemusena võtta vastu

$$\begin{aligned} &\mathrm{H}_0, \text{ kui } \ t \leq t_\alpha(\nu), \\ &\mathrm{H}_1, \text{ kui } \ t > t_\alpha(\nu). \end{aligned}$$

Kui mõlemad valimid on suured, siis t-jaotus läheneb normaaljaotusele ning kriitiliseks väärtuseks võib võtta normaaljaotuse täiendkvantiili z_α.

Olukord, kus kahe kogumi keskväärtus on ühesugune või erinev, on erijuht üldisemast võrdlusest, kus kahe kogumi keskväärtuse vahe võrdub või ei võrdu mingi arvuga D_0. Hüpoteese (7.19) ja (7.20) võime üldistada:

Kogumite keskväärtus erineb mingi arvu võrra

$$\begin{aligned} &\mathrm{H}_0 : \mu_1 - \mu_2 = D_0, \\ &\mathrm{H}_1 : \mu_1 - \mu_2 \neq D_0. \end{aligned} \tag{7.32}$$

Kui $D_0 = 0$, saame siit hüpoteesid (7.19) ja (7.20).

Kui kogumite dispersioonid on erinevad, kasutatakse hüpoteesipaari (7.32) testimiseks teststatistikut

$$t = \frac{\bar{x}_1 - \bar{x}_2 - D_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \sim t(\nu_{WS}), \tag{7.33}$$

kus kriitilise väärtuse leidmiseks vajalik vabadusastmete arv ν_{WS} leitakse valemist (7.22). Ühesuguste dispersioonide korral kasutatakse teststatistikut

$$t = \frac{\bar{x}_1 - \bar{x}_2 - D_0}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t(n_1 + n_2 - 2), \tag{7.34}$$

kus ühine dispersioon s_p^2 leitakse valemist (7.23).

Analoogse üldistuse võime teha ka ühepoolse hüpoteesi jaoks. Kui soovime kontrollida, kas kahe kogumi keskväärtuste vahe on suurem mingist arvust D_0, siis

$$\begin{aligned} &\mathrm{H}_0 : \mu_1 - \mu_2 \leq D_0, \\ &\mathrm{H}_1 : \mu_1 - \mu_2 > D_0. \end{aligned} \tag{7.35}$$

Tabelarvutuses leitakse t-testi kriitilised väärtused kahepoolse hüpoteesi korral funktsiooni T.INV.2T abil ja ühepoolse hüpoteesi korral funktsiooni T.INV abil.

Programmis Excel on t-testi läbiviimiseks otstarbekas kasutada vastavat andmeanalüüsi vahendit komplektist *Data Analysis*.

1. Kui kogumite dispersioonid on **ühesugused**, siis tuleb kasutada vahendit *t-Test: Two-Sample Assuming Equal Variances*.
2. Kui kogumite dispersioonid on **erinevad**, tuleb kasutada vahendit *t-Test: Two-Sample Assuming Unequal Variances*.

Näited nende vahendite kasutamise kohta on lisades C.4 ja C.5. Seda, kas dispersioonid on ühesugused või mitte, tuleb eelnevalt testida ja seda vaatame alapeatükis 7.9. Kui dispersioonide erinevust pole testitud, tuleks kasutada teist varianti.

Näide 7.11. Meeste ja naiste sissetulek

N07Hüpoteesid
N7.11

Palju on kirjutatud sellest, et Eestis on naiste töö vähemtasustatud kui meestel. Kasutame Eesti sotsiaaluuringu andmeid (*Eesti sotsiaaluuring* 2013) ja uurime, kas meestel on sissetulek palgatööst suurem kui naistel. Kuna sissetulek sõltub ettevõtte tegevusalast ja ka ametikohast, siis on vastajate hulgast välja valitud vaid need, kes märkisid oma ettevõtte tegevusalaks finantsvahendus, kinnisvara, rentimine ja äritegevus ning oma ametialaks keskastme spetsialist, tehnik või ametnik. Mehi oli nende hulgas 28 ja naisi 71. Meestel oli keskmine rahaline netosissetulek palgatööst 9337,23 eurot aastas ja naistel 6757,37 eurot aastas.

Testime, kas meestel on keskmine sissetulek suurem kui naistel. Hüpoteesipaar:

$$\mathrm{H}_0 : \mu_1 \leq \mu_2,$$
$$\mathrm{H}_1 : \mu_1 > \mu_2,$$

kus valim 1 on mehed ja valim 2 naised.
Lähtume erinevate dispersioonide metoodikast ning kasutame programmis Excel vahendit *t-Test: Two-Sample Assuming Unequal Variances* (vt ka lisa C.4). Tulemused on esitatud järgnevas tabelis.

	Mehed	Naised
Valimi keskmine	9337,23	6757,73
Valimi dispersioon	43292030	12846605
Valimi maht	28	71
Nullhüpoteesile vastav erinevus	0	
Vabadusastmete arv	34	
t-statistik	1,963	
Olulisuse tõenäosus ühepoolse hüpoteesi korral	0,029	
Kriitiline väärtus ühepoolse hüpoteesi korral	1,691	
Olulisuse tõenäosus kahepoolse hüpoteesi korral	0,058	
Kriitiline väärtus kahepoolse hüpoteesi korral	2,032	

Väljastatakse kummagi valimi keskmine ning dispersioon. t-statistik leitakse valemist (7.21) ja vabadusastmete arv valemist (7.22). Otsuse vastuvõtmiseks on lisatud olulisuse tõenäosus ja t-statistiku parempoolne kriitiline väärtus olulisuse nivool 0,05 ühepoolse hüpoteesi jaoks. Samad suurused väljastatakse ka kahepoolse hüpoteesi jaoks, kuid see meid praegu ei huvita.
Näeme, et t-statistik on suurem kui kriitiline väärtus, $1{,}963 > 1{,}691$. Järelikult on olulisuse nivool 0,05 nullhüpotees ümber lükatud ning võtame vastu sisuka hüpoteesi: valitud tegevusalal ja ametialal on meeste sissetulek suurem kui naiste oma. Sama järelduse teeme, kui vaatame olulisuse tõenäosust ja võrdleme seda olulisuse nivooga: $0{,}029 < 0{,}05$.
Paneme tähele seda, et kahepoolse hüpoteesi korral tuleks jääda nullhüpoteesi juurde, sest olulisuse tõenäosus $0{,}058 > 0{,}05$. Järelikult kui oleksime soovinud testida, kas meestel ja naistel on sissetulek erinev, oleks tulemuseks olnud, et see pole tõestatud. Seepärast tulekski võimaluse korral püstitada ühepoolne hüpotees.

7.8. Kahe kogumi keskväärtuse t-test ja sõltuvad valimid

Kui mingi tunnuse muutuste määramiseks mõõdetakse samu objekte kahel korral, on tegemist **sõltuvate valimitega** (*paired samples*). Mõõtmistulemustes esineb iga objekti jaoks kaks väärtust. Sellisel juhul ei tohi kasutada eelmises punktis esitatud metoodikat, vaid tuleb leida üksikobjektidele vastavad erinevused ja testida, kas keskmine erinevus on null (nullhüpotees) või mitte (sisukas hüpotees).

Olgu i-nda objekti esimese mõõtmise tulemus x_i ja teise mõõtmise tulemus y_i. Erinevus

$$d_i = y_i - x_i. \tag{7.36}$$

Leitud erinevusi d_i vaadeldakse ühe valimina ning leitakse erinevuste aritmeetiline keskmine $\bar{d}$ (6.3) ja standardhälve s_d (6.5). Kahepoolse hüpoteesi korral kontrollitakse, kas erinevuste keskväärtus on nullist erinev. Ühepoolse hüpoteesi korral kontrollitakse, kas erinevuste keskväärtus on suurem või väiksem nullist.

Keskväärtuse testimine, sõltuvad valimid

1. Hüpoteesipaar:

kahepoolne	vasakpoolne	parempoolne
$H_0: D = 0$	$H_0: D \geq 0$	$H_0: D \leq 0$
$H_1: D \neq 0$	$H_1: D < 0$	$H_1: D > 0,$

kus D on erinevuste keskväärtus kogumis.

2. Statistiku t empiiriline väärtus:

$$t = \frac{\bar{d}}{se}, \quad se = \frac{s_d}{\sqrt{n}}, \tag{7.37}$$

kus $\bar{d}$ on erinevuste $d_i = y_i - x_i$ aritmeetiline keskmine, s_d erinevuste valimi standardhälve ja n valimi maht.

3. Kriitilised väärtused võib suure valimi korral leida normaaljaotusest, väikese valimi korral t-jaotusest vabadusastmete arvuga

$$\nu = n - 1. \tag{7.38}$$

4. Võtta vastu

H_0, kui	$\mid t \mid \leq t_{\alpha/2}(\nu)$,	$t \geq -t_\alpha(\nu)$,	$t \leq t_\alpha(\nu)$,
H_1, kui	$\mid t \mid > t_{\alpha/2}(\nu)$,	$t < -t_\alpha(\nu)$,	$t > t_\alpha(\nu)$.

Näide 7.12. Müügitulu töötaja kohta veonduse ja laonduse tegevusalal

Kuidas mõjutas majanduskriis veondus- ja laondusettevõtete müügitulu? Kas 2009. aastal oli see väiksem kui 2008. aastal? Hüpoteesi kontrollimiseks võrdleme juhuslikult valitud 10 veondus- ja laondusettevõtte müügitulu ühe töötaja kohta aastatel 2008 ja 2009[a].

Kümne ettevõtte andmed on esitatud järgnevas tabelis. Müügitulu töötaja kohta on tuhandetes eurodes. Viimases veerus on erinevus 2009. aasta ja 2008. aasta müügitulu vahel.

Ettevõtte ID	Müügitulu töötaja kohta 2008	2009	Erinevus d_i
1	112,6	115,3	2,7
2	147,2	145,5	−1,7
3	60,9	52,6	−8,3
4	111,7	89,1	−22,6
5	186,5	180,8	−5,7
6	164,2	130,7	−33,5
7	73,3	55,6	−17,7
8	28,7	37,4	8,7
9	60,9	38,7	−22,2
10	405,3	310,9	−94,4

1. Hüpoteesipaar:

$$\mathrm{H}_0 : D \geq 0,$$
$$\mathrm{H}_1 : D < 0,$$

kus D on 2009. aasta ja 2008. aasta väärtuste erinevuste keskmine.

2. Tabeli viimase veeru põhjal leiame erinevuste keskmise

$$\bar{d} = \frac{1}{n}\sum d_i = -19{,}47,$$

kus

$$d_i = mt09_i - mt08_i.$$

Siin $mt08_i$ on i-nda ettevõtte müügitulu töötaja kohta aastal 2008 ja mt_i09 sama näitaja aastal 2009. Erinevuste standardhälve $s_d = 29{,}34$. Valemitest (7.37) leiame t-statistiku:

$$t = \frac{-19{,}47}{\frac{29{,}34}{\sqrt{10}}} \approx -2{,}1.$$

3. Kriitilise väärtuse olulisuse nivool 0,05 leiame t-jaotusest vabadusastmete arvuga $\nu = 10 - 1 = 9$. Tabelarvutuses: T.INV(0,05; 9) = −1,83.
4. Teststatistiku ja kriitilise väärtuse võrdlemine: $-2{,}1 < -1{,}83$.
5. Kuna teststatistik on väiksem kui vasakpoolne kriitiline väärtus (nullist kaugemal), on nullhüpotees ümber lükatud. 2009. aastal oli veondus- ja laondusettevõtetes müügitulu ühe töötaja kohta väiksem kui 2008. aastal.

[a] Allikas: Äriregister

Mis juhtub, kui kasutame sõltuvate valimite korral vale meetodit, sõltumatute valimite t-testi? Analüüsime seda näite 7.12 põhjal.

Sõltumatute valimite korral tuleb eraldi leida müügitulu keskmine ja standardhälve 2008. ja 2009. aastal. 2008. aastal on need $\bar{x}_{08} = 135{,}13$ ja $s_{08} = 107{,}41$ ning 2009. aastal $\bar{x}_{09} = 115{,}66$ ja $s_{09} = 84{,}21$. Sõltumatute valimite korral kasutame erinevate dispersioonidega varianti ehk Welchi testi. Teststatistik valemist (7.21):

$$t = \frac{115{,}66 - 135{,}13}{\sqrt{\frac{84{,}21^2}{10} + \frac{107{,}42^2}{10}}} = -0{,}45.$$

Vabadusastmete arv on 17 ja kriitiline väärtus olulisuse nivoo 0,05 korral −1,74. Näeme, et teststatistik on nullile lähemal kui kriitiline ($-0{,}45 > -1{,}74$) ning nüüd kehtib nullhüpotees.

Probleem on selles, et kui me leiame müügitulu standardhälbe mõlema aasta jaoks eraldi, siis need tulevad väga suured, sest ettevõtete lõikes erineb müügitulu väga palju (vt tabel 7.6). t-statistiku arvutusest näeme, et sellisel juhul on nimetaja suur ning t-statistiku absoluutväärtus tuleb väike, nullile lähemal. Aga kui me leiame iga ettevõtte jaoks müügitulu erinevuse, siis on erinevuste hajumine oluliselt väiksem, selle standardhälve oli 29,34. Sellisel juhul tuleb t-statistiku absoluutväärtus suurem, t-statistik asub nullist kaugemal ning nullhüpoteesi on kergem ümber lükata.

Hoiduda tuleb ka vastupidisest eksimusest, sõltuvate valimite meetodi kasutamine sõltumatute valimite korral. Erinevusel (7.36) on mõte vaid siis, kui väärtused x_i ja y_i kuuluvad ühele ja samale objektile, esinevad paarikaupa. Inglise keeles ongi selle testi nimetuseks tihti *paired samples t-test*. Sõltuvate valimite testi on lubatud kasutada erinevate objektide korral, kui eelnevalt on toimunud objektide paarikaupa sobitamine. Sellisel juhul on valimites erinevad objektid ja igale objektile ühest valimist on leitud sobiv paariline teisest valimist (*matched pairs*). Sobivuse määramine toimub tausttunnuste alusel.

Tabel 7.6. Varieeruvuse võrdlus sõltuvate valimite korral, andmed näitest 7.12

Ettevõtte ID	Müügitulu töötaja kohta 2008		2009		Erinevus d_i	
1	112,6	Suur varieerumine	115,3	Suur varieerumine	2,7	Väike varieerumine
2	147,2		145,5		−1,7	
3	60,9		52,6		−8,3	
4	111,7		89,1		−22,6	
5	186,5		180,8		−5,7	
6	164,2		130,7		−33,5	
7	73,3		55,6		−17,7	
8	28,7		37,4		8,7	
9	60,9		38,7		−22,2	
10	405,3		310,9		−94,4	
Valimi standardhälve	107,41		84,21		29,34	

Sõltuvate valimite testimiseks saab programmis Excel kasutada vahendit *t-Test: Paired Two Sample for Means*, mis asub andmeanalüüsi komplektis *Data Analysis* (vt ka lisa C.6).

Näites 7.12 toodud andmete korral on tulemus Excelis järgmine (*Variable*1 on 2009. ja *Variable*2 2008. aasta andmed):

N07Hüpoteesid
N7.12

	2009	2008
Valimi keskmine	115,66	135,13
Valimi dispersioon	7091	11539
Valimi maht	10	10
Nullhüpoteesile vastav erinevus	0	
Vabadusastmete arv	9	
t-statistik	−2,1	
Olulisuse tõenäosus ühepoolse hüpoteesi korral	0,033	
Kriitiline väärtus ühepoolse hüpoteesi korral	1,83	

t-statistik tuleb sama, mis näites 7.12 arvutatud, sest kasutatakse sama arvutusskeemi. Lisaks on esitatud olulisuse tõenäosus ning kriitiline väärtus olulisuse nivoo 0,05 korral. Paneme tähele seda, et Exceli tabelis esitatakse parempoolne kriitiline väärtus. Kuna tegemist on vasakpoolse hüpoteesiga, siis tuleb t-statistikut võrrelda vasakpoolse kriitilise väärtusega: $-2{,}1 < -1{,}83$. Exceli aruandes on ka kriitiline väärtus kahepoolse hüpoteesi korral (vt lisa C.6), kuid see meid praegu ei huvita ja seda me ei ole välja toonud.

Kui programmi Excel andmeanalüüsi komplekti *Data Analysis* kasutamine pole võimalik, siis t-testi jaoks võib tabelarvutuses kasutada ka funktsiooni **T.TEST**. See leiab t-testi olulisuse tõenäosuse, kui võr-

reldakse kahe kogumi keskväärtust. Funktsiooni parameetrid on järgmised:

Array1 ühe valimi andmed;
Array2 teise valimi andmed;
Tails ühepoolse hüpoteesi korral 1, kahepoolse hüpoteesi korral 2;
Type sõltuvad valimid 1, sõltumatud valimid ja võrdsed dispersioonid 2, sõltumatud valimid ja erinevad dispersioonid 3.

7.9. Dispersioonide võrdlemine *F*-testiga ja *t*-testi valik

Alapeatükis 7.7 nägime, et kahe sõltumatu valimi korral on õige t-testi valikuks vaja teada, kas valimid on ühesuguse või erineva dispersiooniga kogumitest. Selle määramiseks testitakse valimite dispersioone. Dispersioonide testimiseks kasutatakse **F-testi**, mille võttis kasutusele inglise matemaatik Ronald Fisher (1890–1962).

Kahepoolse hüpoteesi korral on nullhüpotees, et kogumite dispersioonid on võrdsed, ja sisukas hüpotees, et ei ole võrdsed:

$$\begin{aligned} &\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2, \\ &\mathrm{H}_1 : \sigma_1^2 \neq \sigma_2^2. \end{aligned} \tag{7.39}$$

Hüpoteesipaari kontrollimiseks leitakse teststatistik, mis on valimite dispersioonide s_1^2 ja s_2^2 suhe ning allub F-jaotusele vabadusastmete arvuga $n_1 - 1$ ja $n_2 - 1$:

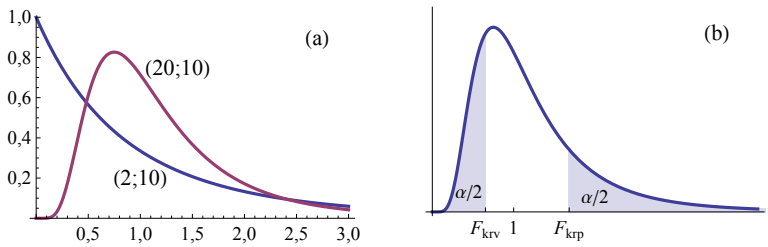
$$F = \frac{s_1^2}{s_2^2} \sim F(n_1 - 1, n_2 - 1). \tag{7.40}$$

F-jaotus on asümmeetriline jaotus. Joonisel 7.13 (a) on esitatud kahe erineva vabadusastmete paariga F-jaotuse tõenäosustiheduse graafikud.

Kui valimite dispersioonid on täpselt võrdsed, siis vastavalt valemile (7.40) $F = 1$. Kui valimite dispersioonid ei ole täpselt võrdsed, võib F statistiku väärtus erineda arvust 1 ühele või teisele poole. Kuna F jaotus on asümmeetriline, siis kahepoolse hüpoteesi korral on kaks kriitilist väärtust F_{krv} ja F_{krp}, mis asuvad arvust 1 vasakul ja paremal pool erinevatel kaugustel. Otsustamine (vaata joonis 7.13 (b)):

- võtta vastu nullhüpotees, kui $F_{krv} \leq F \leq F_{krp}$;
- võtta vastu sisukas hüpotees, kui $F < F_{krv}$ või $F > F_{krp}$.

Praktikas pole mõlemat kriitilist väärtust vaja leida. Kui me oleme leidnud F-statistiku empiirilise väärtuse, siis teame, kas see on ühest väiksem või suurem. Kui see on ühest väiksem, siis võrdleme väärtusega F_{krv}, ja kui ühest suurem, võrdleme väärtusega F_{krp}.



Joonis 7.13. (a) Kaks erinevat F-jaotuse tõenäosustiheduse graafikut vabadusastmete arvudega $\nu_1 = 2$ ja $\nu_2 = 10$ ning $\nu_1 = 20$ ja $\nu_2 = 10$. (b) Kahepoolse F-testi korral võetakse vastu nullhüpotees, kui $F_{krv} \leq F \leq F_{krp}$. Sisukas hüpotees võetakse vastu, kui $F < F_{krv}$ või $F > F_{krp}$

Kui kriitiliste väärtuste leidmiseks kasutada F-jaotuse täiendkvantiile (lisa B.2), siis

$$F_{krv} = F^{-1}_{\alpha/2}(n_2 - 1, n_1 - 1), \tag{7.41}$$

$$F_{krp} = F_{\alpha/2}(n_1 - 1, n_2 - 1). \tag{7.42}$$

Siin juhime tähelepanu sellele, et vasakpoolne kriitiline väärtus F_{krv} on jaotuse $F(n_2 - 1, n_1 - 1)$ täiendkvantiili pöördväärtus.

Tabelarvutuses on kahepoolse F-testi jaoks mugav kasutada funktsiooni **F.TEST**, mis leiab olulisuse tõenäosuse. Funktsiooni argumendid *Array1* ja *Array2* on vastavalt ühte ja teise valimisse kuuluvad arvud.

Kui meil puuduvad andmed üksikute objektide kohta ja on vaid valimite standardhälbed või dispersioonid, siis funktsiooni F.TEST kasutada ei saa. Sellisel juhul tuleb valemist (7.40) leida F-statistik ning võrrelda seda kriitilise väärtusega. F-testi kriitiliste väärtuste leidmiseks on tabelarvutuses kaks funktsiooni. Vasakpoolse kriitilise väärtuse F_{krv} leiab funktsioon **F.INV** (*Probability*; *Deg_freedom1*; *Deg_freedom2*) ja parempoolse kriitilise väärtuse F_{krp} jaoks on funktsioon **F.INV.RT**. Kahepoolse hüpoteesi korral parameeter *Probability* $= \alpha/2$, kus α on olulisuse tõenäosus, ning *Deg_freedom1* ja *Deg_freedom2* on vastavalt lugeja ja nimetaja vabadusastmete arvud:

F_{krv} =F.INV($\alpha/2$; $n_1 - 1$; $n_2 - 1$),
F_{krp} =F.INV.RT($\alpha/2$; $n_1 - 1$; $n_2 - 1$).

Näide 7.13. Sissetulekute dispersioonide võrdlus

Näites 7.11 meeste ja naiste sissetulekute võrdlemisel eeldasime, et kogumite dispersioonid on erinevad. Kontrollime seda nüüd F-testiga.

Meeste arv oli $n_1 = 28$ ning sissetulekute dispersioon $s_1^2 = 43292030$. Naiste arv oli $n_2 = 71$ ning nende sissetulekute dispersioon $s_2^2 = 12846605$.
1. Kasutame kahepoolset F-testi.
2. Hüpoteesipaar:

$$\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2,$$
$$\mathrm{H}_1 : \sigma_1^2 \neq \sigma_2^2.$$

3. Valemist (7.40) F-statistik

$$F = \frac{s_1^2}{s_2^2} = \frac{43292030}{12846605} \approx 3{,}37.$$

4. Kasutame olulisuse nivood 0,05. Lugeja (mehed) vabadusastmete arv

$$n_1 - 1 = 28 - 1 = 27$$

ja nimetaja (naised) vabadusastmete arv

$$n_2 - 1 = 71 - 1 = 70.$$

Parempoolne kriitiline väärtus vastava tabelarvutuse funktsiooni abil $F_{krp} = \mathrm{F.INV.RT}(0{,}025;27;70) = 1{,}81$.
5. F-statistik on suurem kui parempoolne kriitiline väärtus: $3{,}37 > 1{,}81$. Nullhüpotees dispersioonide võrdsuse kohta on ümber lükatud.
6. Kuna F-test näitas, et dispersioonid ei ole võrdsed, oli näites 7.11 valitud t-test õige.

Näide 7.14. Kalendriefekt väärtpaberiturul

N07Hüpoteesid
N7.14

Mitmed autorid on kindlaks teinud, et väärtpaberite keskmised tulumäärad on kuuvahetusel kõrgemad kui tavalistel börsipäevadel. Seda nimetatakse kalendriefektiks. Üldiselt on kalendriefekt aegridades esinev süstemaatiline kõikumine, mis on seotud kalendriga.
2010. aastal TTÜ Majandusteadukonnas kaitstud bakalaureusetöös „Kalendriefektide esinemine NASDAQ OMX Tallinna väärtpaberiturul“ testiti kuuvahetuse efekti olemasolu OMXT indeksi tulumäära aegreas 3.07.2006–30.12.2009 (Noormägi, 2010). Kuuvahetuseks loeti kuu viimast ja järgmise kuu esimest kolme kuupäeva.

Üks valim on tulumäärad kuuvahetustel ($n_1 = 167$) ja teine valim tulumäärad ülejäänud päevadel ($n_2 = 714$). Kontrollimaks hüpoteesi, et kuuvahetustel on tulumäärad kõrgemad, kasutame ühepoolset t-testi. Tegemist on sõltumatute valimitega ja enne t-testi valikut tuleb kontrollida, kas kogumite dispersioonid on võrdsed või mitte.
1. Dispersioonide võrdlemiseks kasutame kahepoolset F-testi.
2. Hüpoteesipaar:

$$\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2,$$
$$\mathrm{H}_1 : \sigma_1^2 \neq \sigma_2^2.$$

3. Leiame teststatistiku olulisuse tõenäosuse, kasutades tabelarvutuses funktsiooni F.TEST. Tulemuseks on $p = 0{,}017$.
4. Olulisuse nivooks võtame 0,05. Kuna olulisuse tõenäosus on väiksem kui olulisuse nivoo ($0{,}017 < 0{,}05$), võtame vastu sisuka hüpoteesi, et dispersioonid ei ole võrdsed.
5. Tuleb kasutada t-testi erinevate dispersioonide korral.
6. Hüpoteesipaar keskväärtuse testimiseks t-testiga:

$$\mathrm{H}_0 : \mu_1 \leq \mu_2,$$
$$\mathrm{H}_1 : \mu_1 > \mu_2.$$

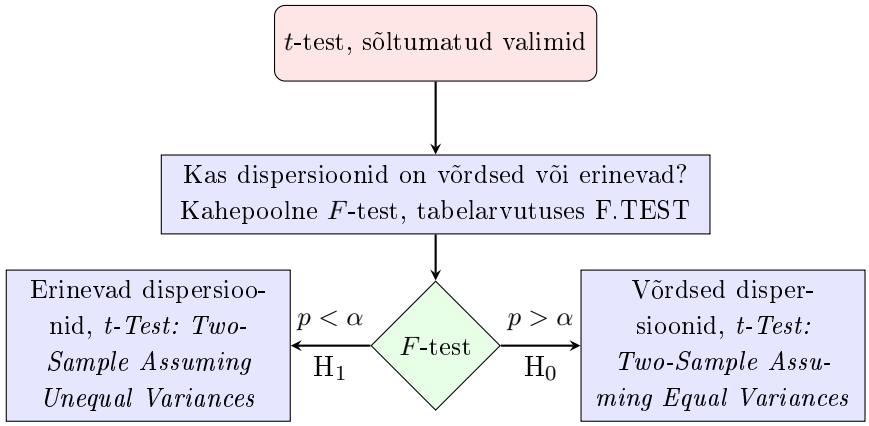
7. Kasutame komplektist *Data Analysis* vahendit *t-Test: Two-Sample Assuming Unequal Variances* (vt ka lisa C.4). Tulemuseks saame järgmise aruande.

	Kuuvahetus	Ülejäänud päevad
Valimi keskmine	0,00129	−0,00075
Valimi dispersioon	0,00016	0,00022
Valimi maht	167	714
Nullhüpoteesiga püstitatud erinevus	0	
Vabadusastmete arv	281	
t-statistik	1,81	
Olulisuse tõenäosus ühepoolse hüpoteesi korral	0,036	
Kriitiline väärtus ühepoolse hüpoteesi korral	1,65	

8. Olulisuse nivooks võtame 0,05. Olulisuse tõenäosus $0{,}036 < 0{,}05$, võtame vastu sisuka hüpoteesi.
9. On tõestatud, et OMXT väärtpaberibörsil esines ajavahemikul 3.07.2006–30.12.2009 kuuvahetuse efekt ning indeksi tulumäär oli kuuvahetustel kõrgem kui ülejäänud päevadel.

Joonisel 7.14 on toodud skeem, kuidas programmis Excel läbi viia kahe kogumi keskväärtuste võrdlemist t-testiga sõltumatute valimite korral.

t-testi valik



Joonis 7.14. t-testi valik sõltumatute valimite korral sõltub dispersioonide võrdlemise F-testi tulemusest. p on F-testi olulisuse tõenäosus, mille väljastab funktsioon F.TEST, ja α on olulisuse nivoo

Mõnikord kontrollitakse eraldi, kas uuritava tunnuse hajumist kirjeldav dispersioon on erinevates kogumites oluliselt erinev. Näiteks, kas sissetulekute varieeruvus erinevates sotsiaalsetes gruppides on oluliselt erinev, kas erinevate väärtpaberite või fondide tulumäära varieeruvus (investeerimisriski näitaja) on oluliselt erinev, kas mõne majandusalase suuruse varieeruvus on erinevatel ajaperioodidel oluliselt erinev jne.

Näide 7.15. Majanduskasvu volatiilsus arengumaades

A. Singh ja S. Fagernäs analüüsisid oma töös "Globalisation, Instability and Economic Insecurity“ majanduskavu volatiilsust (kõikumist) arengumaades. Nad püstitasid hüpoteesi, et vaadeldud riikides oli SKP kasvumäära dispersioon aastatel 1982–2004 oluliselt väiksem kui aastatel 1960–1981. (Singh ja Fagernäs, 2006)
Tabelis on toodud SKP keskmine kasvumäär (%) Lõuna-Aasia ja Ladina-Ameerika riikides neljal perioodil. Viimases veerus on ühepoolse F-testi olulisuse tõenäosus p. On näha, et Lõuna-Aasia riikide korral leidis püstitatud hüpotees kinnitust olulisuse nivool 0,01 ($p < 0{,}01$), kuid Ladina-Ameerika riikide korral mitte.

	1960–1971	1972–1981	1982–1991	1992–2004	p
			Lõuna-Aasia		
Keskmine	4,0	3,6	5,2	5,7	
Standardhälve	2,6	3,6	1,8	1,3	0,001
			Ladina-Ameerika		
Keskmine	5,4	5,1	1,6	2,8	
Standardhälve	1,8	2,2	2,4	2,2	0,79

Ühepoolne F-test

Ühepoolse F-testi korral testitakse, kas ühes kogumis on dispersioon suurem kui teises. Hüpoteesipaar:

$$\begin{aligned} \mathrm{H}_0 &: \sigma_1^2 \leq \sigma_2^2, \\ \mathrm{H}_1 &: \sigma_1^2 > \sigma_2^2. \end{aligned} \tag{7.43}$$

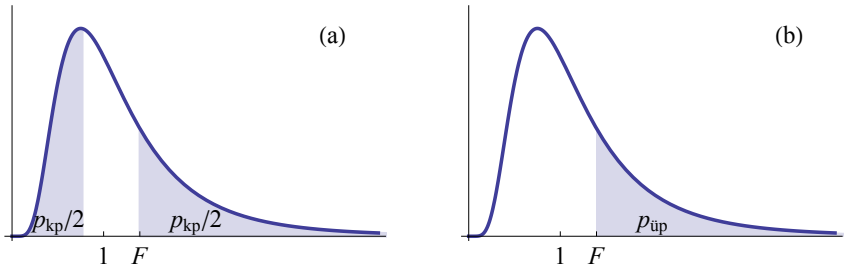
Teststatistik on sama, mis kahepoolse hüpoteesi korral (valem (7.40)), kusjuures lugejas peab olema selle valimi dispersioon, mis on suurem. Järelikult valimiks number 1 tuleb võtta suurema dispersiooniga valim. Kriitiline väärtus on arvust üks paremal ja võrdub F-jaotuse täiendkvantiiliga. Võtta vastu:

H_0, kui $1 \leq F \leq F_{kr}$,
H_1, kui $F > F_{kr}$.

Olulisuse tõenäosuse p kasutamisel:

H_0, kui $p \geq \alpha$,
H_1, kui $p < \alpha$.

Ühepoolse F-testi olulisuse tõenäosuse p leidmiseks tabelarvutuses tuleb funktsiooniga F.TEST leitud väärtus jagada kahega, sest F.TEST leiab olulisuse tõenäosuse kahepoolse hüpoteesi jaoks (vt joonis 7.15). Excelis võib ühepoolse F-testi jaoks kasutada ka vahendit *F-Test Two-Sample for Variances* komplektist *Data Analysis*.



Joonis 7.15. F-statistiku olulisuse tõenäosus kahepoolse (a) ja ühepoolse (b) hüpoteesi korral. On näha, et olulisuse tõenäosus ühepoolse hüpoteesi korral $p_{\text{üp}} = p_{\text{kp}}/2$, kus p_{kp} on olulisuse tõenäosus kahepoolse hüpoteesi korral

7.10. Osakaalu testimine suurte valimite korral

Kas Eestis on naiste osakaal tippjuhtide seas erinev kui teistes riikides? Kas teatud kauba või teenuse tarbijaid on ühes vanusegrupis rohkem kui teises? Sellistele küsimustele saame vastuse, kui testime osakaalusid. Nii nagu kogumi keskväärtuse testimisel võib ka osakaalu testimisel olla tegemist ühe või kahe valimiga. Ühe valimi korral võrreldakse valimi osakaalu mingi etteantud väärtusega. Kahe valimi korral võrreldakse valimite osakaalusid omavahel, et teha kindlaks, kas need tulevad võrdsete osakaaludega kogumitest või mitte.

Kui kogumis on kaheväärtuselise tunnuse ühe väärtuse esinemise tõenäosus p ja me teeme valimi mahuga n, siis valimis allub selle väärtuse jaotus binoomjaotusele $B(n,p)$. Seepärast tuleks testimisel kriitiliste väärtuste leidmiseks kasutada binoomjaotust, mida vaatame järgmises alapeatükis. Suurte valimite korral aga allub valimi osakaal $\hat{p}$ ligikaudu normaaljaotusele:

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right). \tag{7.44}$$

Seepärast võib suurte valimite korral kasutada teststatistiku kriitiliste väärtuste leidmiseks normaaljaotust.

Osakaalu testimine, suur valim

Osakaalu testimine suure valimi korral

1. Hüpoteesipaar:

kahepoolne	vasakpoolne	parempoolne
$H_0 : p = p_0$	$H_0 : p \geq p_0$	$H_0 : p \leq p_0$
$H_1 : p \neq p_0$	$H_1 : p < p_0$	$H_1 : p > p_0$.

2. Teststatistiku empiiriline väärtus:

$$z = \frac{\hat{p} - p_0}{se_p}. \tag{7.45}$$

Kuna teststatistiku jaotus leitakse nullhüpoteesi kehtimise korral, siis standardvea leidmisel võetakse kogumis osakaaluks nullhüpoteesiga püstitatud väärtus p_0. Standardviga

$$se_p = \sqrt{\frac{p_0(1-p_0)}{n}}, \tag{7.46}$$

kus n on valimi maht.

3. Olulisuse nivoole α vastav kriitiline väärtus on standardiseeritud normaaljaotuse täiendkvantiil $z_{\alpha/2}$ kahepoolse hüpoteesi korral ja z_α ühepoolse hüpoteesi korral.
4. Võtta vastu

H_0, kui	$\lvert z\rvert \le z_{\alpha/2}$,	$z \ge -z_\alpha$,	$z \le z_\alpha$,
H_1, kui	$\lvert z\rvert > z_{\alpha/2}$,	$z < -z_\alpha$,	$z > z_\alpha$.

Näide 7.16. Vaesusrisk Euroopa Liidus ja Eestis

Vaesusrisk tähendab seda, kui isiku sissetulek jääb allapoole suhtelist vaesuspiiri, mis on 60% sissetulekute mediaanist. 2013. aastal elas Euroopa Liidus 16,8% elanikkonnast vaesusriskis (*Income distribution statistics* 2015). Kui suur on vaesusriskis elavate inimeste osakaal Eestis?
Kasutame Eesti sotsiaaluuringu andmeid (*Eesti sotsiaaluuring* 2013). Valimi maht oli 15 053 isikut ja nendest 2770 elasid suhtelisest vaesuspiirist allpool. Vaesusriskis elavate inimeste osakaal valimis

$$\hat{p} = \frac{2770}{15053} = 0{,}184.$$

Kas Eestis on selle uuringu järgi vaesusriskis elavate inimeste osakaal suurem kui Euroopa Liidus keskmiselt?
1. Kasutame ühepoolset testi osakaalude testimiseks.
2. Hüpoteesipaar:

$$H_0 : p \le 0{,}168,$$
$$H_1 : p > 0{,}168.$$

3. Standardviga valemist (7.46):

$$se_p = \sqrt{\frac{0{,}168(1-0{,}168)}{15053}} \approx 0{,}003.$$

Statistiku z empiiriline väärtus valemist (7.45):

$$z = \frac{0{,}184 - 0{,}168}{0{,}003} \approx 5{,}3.$$

4. Olulisuse nivoole 5% vastava kriitilise väärtuse leiame normaaljaotusest, kuna tegemist on suure valimiga. Ühepoolse hüpoteesi korral on see 1,64.
5. Teststatistik langeb kriitilisse piirkonda: $5{,}3 > 1{,}64$.

6. Kuna teststatistik langeb kriitilisse piirkonda, tuleb nullhüpotees tagasi lükata ja võtta vastu sisukas hüpotees. Võime väita, et olulisuse nivool 5% on Eestis vaesusriskis elavate inimeste osakaal suurem kui Euroopa Liidus keskmiselt.

Kahe kogumi osakaalude p_1 ja p_2 võrdlemisel kasutatakse kahte valimit. Osakaalud valimites:

$$\hat{p}_1 = \frac{m_1}{n_1},$$
$$\hat{p}_2 = \frac{m_2}{n_2}, \tag{7.47}$$

kus n_1 on valimi 1 maht ja m_1 vastavat väärtust omavate objektide arv selles valimis ning n_2 ja m_2 samad suurused valimi 2 jaoks. Ka nüüd võib suurte valimite korral kasutada teststatistiku kriitiliste väärtuste leidmiseks normaaljaotust, sest osakaalude erinevuse standardviga allub standardiseeritud normaaljaotusele

$$se_{p_1-p_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \sim N(0,1). \tag{7.48}$$

Osakaalude võrdlemine, suured valimid

Kahe kogumi osakaalude võrdlemine suurte valimite korral

1. Hüpoteesipaar:

kahepoolne	vasakpoolne	parempoolne
$H_0 : p_1 = p_2,$	$H_0 : p_1 \geq p_2,$	$H_0 : p_1 \leq p_2,$
$H_1 : p_1 \neq p_2,$	$H_1 : p_1 < p_2,$	$H_1 : p_1 > p_2.$

2. Teststatistik

$$z = \frac{\hat{p}_1 - \hat{p}_2}{se_{p_1-p_2}}, \quad se_{p_1-p_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \tag{7.49}$$

kus $\hat{p}_1$ ja $\hat{p}_2$ on vastava tunnuse osakaal valimites 1 ja 2 ning n_1 ja n_2 valimite mahud.

3. Statistiku (7.49) kriitilised väärtused on samad, mis eelmise testi korral. Olulisuse nivoole α vastav kriitiline väärtus on standardiseeritud normaaljaotuse täiendkvantiil $z_{\alpha/2}$ kahepoolse hüpoteesi korral ja z_α ühepoolse hüpoteesi korral.

4. Võtta vastu

H_0, kui	$\lvert z\rvert \leq z_{\alpha/2},$	$z \geq -z_\alpha,$	$z \leq z_\alpha,$
H_1, kui	$\lvert z\rvert > z_{\alpha/2},$	$z < -z_\alpha,$	$z > z_\alpha.$

Näide 7.17. Naiste osakaal äriühingute juhatustes

2013. aasta aprillis viidi Euroopa Liidus läbi uuring liidripositsioonide hõlvamise kohta naiste ja meeste poolt (*Women and men in leadership positions in the European Union, 2013* 2013). Selle järgi oli 2013. aastal Prantsusmaal naisjuhtide osakaal äriühingute juhatustes 26,8% ja Inglismaal 18,5%. Prantsusmaal kuulus valimisse 35 ettevõtet ja Inglismaal 46 ettevõtet. Kas selle uuringu järgi võib järeldada, et Prantsusmaal kuulub keskmiselt rohkem naisi äriühingute juhatusse kui Inglismaal?
1. Kasutame ühepoolset testi osakaalude testimiseks. Olgu Prantsusmaa ettevõtted valim number 1 ja Inglismaa ettevõtted valim number 2.
2. Hüpoteesipaar:

$$\begin{aligned} \mathrm{H}_0 &: p_1 \leq p_2, \\ \mathrm{H}_1 &: p_1 > p_2. \end{aligned}$$

3. Statistiku z empiiriline väärtus valemitest (7.49):

$$z = \frac{0{,}268 - 0{,}185}{\sqrt{\frac{0{,}268 \cdot (1 - 0{,}268)}{35} + \frac{0{,}185 \cdot (1 - 0{,}185)}{46}}} = 0{,}881.$$

4. Olulisuse nivoole 5% vastava kriitilise väärtuse leiame standardiseeritud normaaljaotusest, kuna tegemist on suure valimiga. Ühepoolse hüpoteesi korral on see 1,64.
5. Teststatistik ei lange kriitilisse piirkonda: $0{,}881 < 1{,}64$.
6. Kuna teststatistik ei lange kriitilisse piirkonda, tuleb vastu võtta nullhüpotees. Olulisuse nivool 5% ei saa väita, et Prantsusmaal on naisi äriühingute juhatustes rohkem kui Inglismaal.

7.11. Märgitest

Eelnevalt vaadeldud z-test, t-test ja F-test olid parameetrilised testid. **Mitteparameetrilised** testid ei kasuta otseselt tunnuste väärtusi, vaid

- väärtuste astakuid (näiteks Wilcoxoni astakmärgitest);
- väärtuste esinemissagedusi (näiteks χ^2-test);
- märkide „+“ ja „−“ esinemissagedusi.

Viimastel põhineb **märgitest** (*sign test*), mida võib kasutada mediaani testimiseks, kahe sõltuva valimi võrdlemiseks või kaheväärtuselise tunnuse osakaalu testimiseks väikese valimi korral. Märgitesti aluseks on

binoomjaotus. Vaatame, kuidas kasutatakse märgitesti, kui soovitakse testida, kas kogumi mediaanil Me on väärtus Me_0 või mitte.

Kahepoolne hüpotees

1. Kahepoolne hüpotees:

$$H_0 : Me = Me_0,$$
$$H_1 : Me \neq Me_0. \tag{7.50}$$

2. Üldkogumist võetakse juhuvalim. Valimi elementide jaoks leitakse, kas need on suuremad kui nullhüpoteesiga püstitatud väärtus Me_0 (tähistatakse „+“) või väiksemad (tähistatakse „−“). Need elemendid, mis võrduvad väärtusega Me_0, jäetakse märkimata ja need jäävad korrigeeritud valimist välja. Seejärel leitakse märkide pluss ja miinus esinemissagedused n^+ ja n^-. Nende summa on **korrigeeritud valimi maht**:

$$n = n^- + n^+. \tag{7.51}$$

3. Kahepoolse hüpoteesi korral jäädakse nullhüpoteesi juurde juhul, kui

$$n_{krv} \leq n^+ \leq n_{krp},$$

kus n_{krv} on vasakpoolne ja n_{krp} parempoolne kriitiline väärtus. Nullhüpotees lükatakse tagasi ja võetakse vastu sisukas hüpotees, kui

$$n^+ < n_{krv} \quad \text{või} \quad n^+ > n_{krp}.$$

4. Kuna nullhüpoteesi kehtimisel on mediaanist Me_0 suurema väärtuse saamise tõenäosus $p = 0{,}5$ (nagu ka väiksema väärtuse saamise tõenäosus), siis n^+ allub binoomjaotusele

$$n^+ \sim B(n, 0{,}5).$$

5. Vasakpoolne kriitiline väärtus n_{krv} leitakse nii, et see on vähim arv, mille korral kehtib võrratus

$$P(n^+ \leq n_{krv}) \geq \frac{\alpha}{2}. \tag{7.52}$$

Tõenäosuse $P(n^+ \leq n_{krv})$ saab leida binoomjaotusest, kasutades Bernoulli valemit (5.49). Parempoolne kriitiline väärtus leitakse seosest

$$n_{krp} = n - n_{krv}. \tag{7.53}$$

Plusside ja miinuste arvu võime vaadata kui kaheväärtuselist tunnust vaid juhul, kui kasutame korrigeeritud valimit. Kui me ei kasutaks korrigeeritud valimit, siis esineks ka kolmas võimalus: erinevus puudub. Sellisel juhul binoomjaotust kasutada ei saa.

Tabel 7.7. Binoomjaotus $B(12;0{,}5)$

m	0	1	2	3
$P(n^+ = m)$	0,00024	0,0029	0,0161	0,0537
$P(n^+ \leq m)$	0,00024	0,0032	0,0193	0,0730

Olgu meil näiteks korrigeeritud valimi maht $n = 12$. Tabelis 7.7 on toodud binoomjaotusest $B(12, 0{,}5)$ leitud tõenäosused mõningate m väärtuste korral.

Meil tuleb leida vähim arv, mille korral kehtib võrratus

$$P(n^+ \leq n_{krv}) \geq 0{,}025.$$

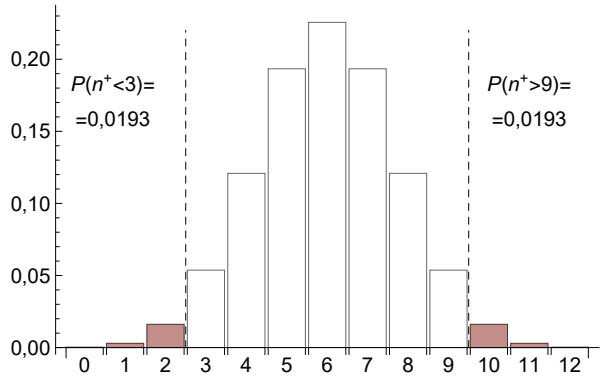
Tabelist 7.7 näeme, et seda tingimust rahuldav väikseim m väärtus on 3. Järelikult, vasakpoolne kriitiline väärtus $n_{krv} = 3$ ja parempoolne valemist (7.53) $n_{krv} = 12 - 3 = 9$. Nullhüpotees võetakse vastu, kui

$$3 \leq n^+ \leq 9$$

ja lükatakse ümber, kui

$$n^+ < 3 \quad \text{või} \quad n^+ > 9.$$

Joonisel 7.16 on kriitilisele piirkonnale vastavad tulbad varjutatud. Kehtiva nullhüpoteesi korral on kriitilisse piirkonda sattumise tõenäosus võrdne nende tulpade kõrguste summaga 0,0386.



Joonis 7.16. Märgitesti kriitiline piirkond kahpoolse hüpoteesi korral, kui korrigeeritud valimi maht on 12. Kehtiva nullhüpoteesi korral on kriitilisse piirkonda sattumise tõenäosus $0{,}0193 + 0{,}0193 = 0{,}0386$

Kui soovime mediaani testimiseks kasutada **ühepoolset hüpoteesi**, siis:

Ühepoolne hüpotees

vasakpoolne	parempoolne
$H_0 : Me \geq Me_0,$	$H_0 : Me \leq Me_0,$
$H_1 : Me < Me_0,$	$H_1 : Me > Me_0.$ (7.54)

Nüüd leitakse vasakpoolne kriitiline väärtus n_{krv} binoomjaotusest $B(n, 0{,}5)$ nii, et see on vähim arv, mille korral kehtib võrratus

$$P(n^+ \leq n_{krv}) \geq \alpha, \tag{7.55}$$

kus α on olulisuse nivoo. Kriitiline väärtus parempoolse hüpoteesi jaoks leitakse valemist (7.53). Märgitesti kriitilised väärtused mõningate korrigeeritud valimi mahu n väärtuste korral on toodud lisas B.3.

Tabelarvutuses on märgitesti vasakpoolse kriitilise väärtuse n_{krv} leidmiseks funktsioon **BINOM.INV**. *Trials* on valimi maht n, *Probability_s* on mediaani testimise korral 0,5 ja *Alpha* on olulisuse nivoo α ühepoolse hüpoteesi korral. Kui meil on

- ühepoolne hüpotees, siis $n_{krv} = \text{BINOM.INV}(n;0{,}5;\alpha)$;
- kahepoolne hüpotees, siis $n_{krv} = \text{BINOM.INV}(n;0{,}5;\alpha/2)$.

Näide 7.18. Elamispinna ruutmeetri mediaanhind Tallinnas

Maa-ameti andmetel oli 2016. aasta mais Tallinna korteri ruutmeetri mediaanhind 1511 eurot. Olgu meil teada ühes Tallinna linnaosas müüdud üheksa korteri ruutmeetri hind. Tabeli teises reas on plussiga märgitud need hinnad, mis on mediaanist suuremad, ja miinusega mediaanist väiksem.

Hind, €	1620	1580	1700	2100	1350	1720	1650	2500	1800
Erinevus mediaanist	+	+	+	+	−	+	+	+	+

1. Kasutame märgitesti.
2. Olgu Me kõigi selles linnaosas müüdud korterite ruutmeetri hinna mediaan. Hüpoteesipaar:
   H_0: $Me \leq 1511$,
   H_1: $Me > 1511$.
3. Valimis on kõik hinnad väärtusest 1511 erinevad, seega korrigeeritud valimi maht $n = 9$.
4. Mediaanist suuremad on 8 hinda: $n^+ = 8$.
5. Olulisuse nivool 5% ja korrigeeritud valimi mahu 9 korral on ühepoolse hüpoteesi puhul vasakpoolne kriitiline väärtus $n_{krv} = 2$ (tabelarvutuse funktsioon BINOM.INV(9;0,5;0,05)) ja parempoolne kriitiline väärtus $n_{krp} = 9 - 2 = 7$. Kuna meil on tegemist parempoolse hüpoteesiga, kasutame parempoolset kriitilist väärtust.
6. Kuna $8 > 7$, on nullhüpotees ümber lükatud. Selles linnaosas on ruutmeetri mediaanhind suurem kui Tallinnas keskmiselt.

Märgitesti kasutatakse ka kogumis toimunud muutuste hindamisel või erinevate kogumite võrdlemisel sõltuvate valimite korral. Positiivsed vahed tähistatakse märgiga „+“ ja negatiivsed märgiga „−“. Hüpoteesipaarid püstitatakse plusside ja miinuste arvu kohta kogumis:

Hüpoteesid märkide arvu kohta

kahepoolne		vasakpoolne		parempoolne	
$H_0:$	$N^+ = N^-,$	$H_0:$	$N^+ \geq N^-,$	$H_0:$	$N^+ \leq N^-,$
$H_1:$	$N^+ \neq N^-,$	$H_1:$	$N^+ < N^-,$	$H_1:$	$N^+ > N^-.$

(7.56)

Kriitilised väärtused leitakse binoomjaotusest $B(n, 0{,}5)$ nii nagu mediaani testimise korral. Vasakpoolse võib leida lisast B.3 või tabelarvutuses funktsiooniga BINOM.INV ja parempoolne leitakse valemist (7.53).

Näide 7.19. Reklaamide hindamine

N07Hüpoteesid N7.19

Juhuslikult valitud 17 inimesele näidati kaht erinevat reklaami ja paluti neil mõlemat hinnata viiepallises skaalas. Tabelis on toodud kummalegi reklaamile pandud hinded. Veerus „Erinevus“ on „+“, kui esimene reklaam oli hinnatud kõrgemalt, ja „−“, kui esimest reklaami oli hinnatud madalamalt. Eesmärgiks on testida, kas hinnang reklaamidele on erinev.

Reklaam 1	Reklaam 2	Erinevus
4	2	+
2	4	−
4	5	−
4	4	
5	4	+
2	5	−
3	3	
4	5	−
3	5	−
5	4	+
3	4	−
3	5	−
4	5	−
5	5	
4	5	−
5	4	+
2	5	−

1. Kuna hinnang pallides ei ole intervallskaalas, siis me ei saa kasutada t-testi keskväärtuste testimiseks ja kasutame märgitesti.
2. Hüpoteesipaar:

H_0: hinnang reklaamidele on ühesugune, $N^+ = N^-$,
H_1: hinnang reklaamidele on erinev, $N^+ \neq N^-$.
3. Reklaame hindas erinevalt 14 inimest, seega korrigeeritud valimi maht $n = 14$.
4. Esimest reklaami hindas kõrgemalt 4 inimest, $n^+ = 4$.
5. Olulisuse nivool 5% ja korrigeeritud valimi mahu 14 korral on kriitilised väärtused $n_{krv} = 3$ (vt lisa B.3) ja $n_{krp} = 14-3 = 11$. Järelikult nullhüpotees kehtib, kui $3 \leq n^+ \leq 11$.
6. Kuna $n^+ = 4$ ei lange kriitilisse piirkonda, tuleb vastu võtta nullhüpotees: hinnang reklaamidele on ühesugune.

Intervallskaalas mõõdetud tunnuse korral on märgitest alternatiivne võimalus t-testile sõltuvate valimite võrdlemiseks. Seda kasutatakse juhul, kui tunnuse väärtus kogumis ei allu normaaljaotusele või kui esinevad ekstreemsed väärtused. Märgitesti korral me ei vaatle vahede arvulisi väärtusi, vaid ainult vahede märke.

Näide 7.20. Müügitulu muutus ja märgitest

Näites 7.12 võrdlesime 10 veondus- ja laondusettevõtte müügitulu ühe töötaja kohta ning kontrollisime hüpoteesi, kas 2009. aastal oli see veondus- ja laondusettevõtetes väiksem kui 2008. aastal.
Kasutame nüüd märgitesti. Selleks märgime need ettevõtted, kus müügitulu oli 2009. aastal väiksem, plussmärgiga.

2008	2009	Erinevus
112,6	115,3	−
147,2	145,5	+
60,9	52,6	+
111,7	89,1	+
186,5	180,8	+
164,2	130,7	+
73,3	55,6	+
28,7	37,4	−
60,9	38,7	+
405,3	310,9	+

1. Kasutame märgitesti.
2. Olgu N^+ kõigi nende ettevõtete arv üldkogumis, kus 2008. aasta müügitulu ühe töötaja kohta oli suurem kui 2009. aastal, ja N^- nende ettevõtete arv, kus see oli 2008. aastal väiksem kui 2009. aastal. Hüpoteesipaar:
H_0: $N^+ \leq N^-$,

H$_1$: $N^+ > N^-$.
3. Valimisse sattunud kõigis ettevõtetes on müügitulu väärtus aastatel 2008 ja 2009 erinev, seega korrigeeritud valimi maht $n = 10$.
4. 2008. aastal oli müügitulu suurem kaheksas ettevõttes, $n^+ = 8$.
5. Olulisuse nivool 5% ja korrigeeritud valimi mahu 10 puhul on ühepoolse hüpoteesi korral vasakpoolne kriitiline väärtus $n_{krv} = 2$ ja parempoolne kriitiline väärtus $n_{krp} = 10 - 2 = 8$. Kuna meil on tegemist parempoolse hüpoteesiga, kasutame parempoolset kriitilist väärtust.
6. Kuna $n^+ = 8$ ei ole suurem kui parempoolne kriitiline väärtus 8, tuleb vastu võtta nullhüpotees: märgitest ei tõesta, et 2008. aastal oli veondus- ja laondusettevõtetes müügitulu ühe töötaja kohta suurem kui 2009. aastal.
t-testi korral oli olulisuse tõenäosus $p = 0{,}033$, seega nivool 0,05 oli nullhüpotees ümber lükatud. t-test võrdleb keskväärtusi, aga keskväärtused võivad olla mõjutatud üksikute ekstreemsete väärtuste poolt. Näiteks ettevõttes 10 vähenes müügitulu väga palju. Märgitesti korral me ei analüüsi, kui palju müügitulu erines ühes või teises ettevõttes, vaid testime ettevõtete arvu, milles müügitulu vähenes. Kui kogumis pooltes ettevõtetes vähenes ja pooltes suurenes, siis valimis mahuga 10 võib neid ettevõtteid, kus vähenes, olla 8. Selle tõenäosus ei ole väiksem kui 0,05.

Osakaalu testimine

Kaheväärtuselise tunnuse osakaalu testimisel märgitakse tunnuse üks väärtus plussiga ja teine väärtus miinusega. Hüpoteesipaari püstitus on samasugune nagu (7.56).

7.12. Jaotuse sobivuse χ^2-test

5. peatükis tutvusime erinevate jaotusseadustega. Teades, et mingi juhuslik suurus allub teatud jaotusseadusele, on võimalik seda jaotusseadust kasutada tõenäosuste ja oodatavate väärtuste leidmiseks. Aga kuidas me saame kindlad olla, et meid huvitav juhuslik suurus allub meie valitud jaotusseadusele? Seda on võimalik testida ja seda nimetatakse **jaotuse sobivuse** testimiseks (*goodness of fit test*). Kõige sagedamini kasutatakse χ^2-testi (hii-ruut test, *chi-square test*), mille teststatistikuks on Pearsoni statistik χ^2:

χ^2-statistik

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i^e - n_i^o)^2}{n_i^o}. \qquad (7.57)$$

Siin on n_i^e juhusliku suuruse erinevate väärtuste esinemissagedused empiirilises jaotuses, n_i^o aga teoreetilise jaotusseaduse alusel leitud oodatavad sagedused. Summeerimine toimub üle kõigi empiirilises jaotuses esinevate juhusliku suuruse väärtuste. Valemist (7.57) on näha, et kui iga i korral empiiriline sagedus langeb kokku oodatava sagedusega, $n_i^e = n_i^o$, siis $\chi^2 = 0$. Kuna χ^2-testis kasutatakse väärtuste esinemissagedusi, siis ka see on mitteparameetriline test.

Näites 5.15 analüüsiti taksofonidest viieminutilistes intervallides tehtud telefonikõnede arvu jaotust ja võrreldi seda Poissoni jaotusega. Kokku vaadeldi 140 intervalli, mille jooksul tehti kõnesid kokku 384. Vastava Poissoni jaotuse keskväärtuseks saadi

$$\lambda = \frac{384}{140} = 2{,}74.$$

Poissoni valemist (5.62) saab leida teoreetilisele jaotusele vastavad tõenäosused $p_i = P(X = m_i)$. Väärtuse m_i esinemise oodatav sagedus on siis

$$n_i^o = 140 p_i.$$

Tabelis 7.8 on toodud empiirilised sagedused n_i^o, tõenäosused p_i ja oodatavad sagedused n_i^o. Paneme tähele, et oodatavate sageduste summa peab olema sama, mis empiirilistel sagedustel. Seepärast viimast väärtust „7“ loetakse „7 ja rohkem“ ning selle tõenäosus leitakse järgmiselt

$$P(X \geq 7) = 1 - P(X \leq 6).$$

Tabel 7.8. Empiirilised ja oodatavad sagedused

N07Hüpoteesid T7.8

i	Kõnede arv intervallis x_i	Empiiriline sagedus n_i^e	Tõenäosus p_i	Oodatav sagedus n_i^o	$\frac{(n_i^e - n_i^o)^2}{n_i^o}$
1	0	5	0,064	9,0	1,788
2	1	33	0,177	24,7	2,770
3	2	24	0,242	33,9	2,895
4	3	38	0,221	31,0	1,580
5	4	23	0,152	21,3	0,143
6	5	9	0,083	11,7	0,607
7	6	4	0,038	5,3	0,332
8	7	4	0,022	3,1	0,260
KOKKU		140	1	140	10,375

Tabeli 7.8 viimases veerus olevate arvude summa on teststatistik

$$\chi^2 = \sum_{i=1}^{8} \frac{(n_i^e - n_i^o)^2}{n_i^o} = 10{,}375. \tag{7.58}$$

Jaotuste võrdlemiseks püsitatakse järgmine hüpoteesipaar:

H_0: empiiriline ja teoreetiline jaotus langevad kokku, erinevus puudub;

H_1: empiiriline ja teoreetiline jaotus erinevad oluliselt.

Hüpoteesipaari matemaatiliseks formuleerimiseks kasutatakse empiirilisi ehk statistilisi tõenäosusi p_i^e ning teoreetilisi ehk oodatavaid tõenäosusi p_i^o. Kuna nullhüpoteesi kehtimisel peavad need tõenäosused võrduma iga i-nda väärtuse korral, kasutatakse tõenäosusvektoreid $\vec{p^e} = (p_1^e, p_2^e, \ldots, p_k^o)$ ja $\vec{p^o} = (p_1^o, p_2^o, \ldots, p_k^o)$:

$$H_0: \quad \vec{p^e} = \vec{p^o}, \tag{7.59}$$

$$H_1: \quad \vec{p^e} \neq \vec{p^o}. \tag{7.60}$$

Valemist (7.57) nägime, et kui empiirilised ja teoreetilised sagedused langevad kokku, siis $\chi^2 = 0$. Kui aga kehtib (7.60), siis $\chi^2 > 0$. Kui palju võib valimi põhjal leitud χ^2-statistik nullist erineda, kui üldkogumis kehtib nullhüpotees? Milline on kriitiline väärtus?

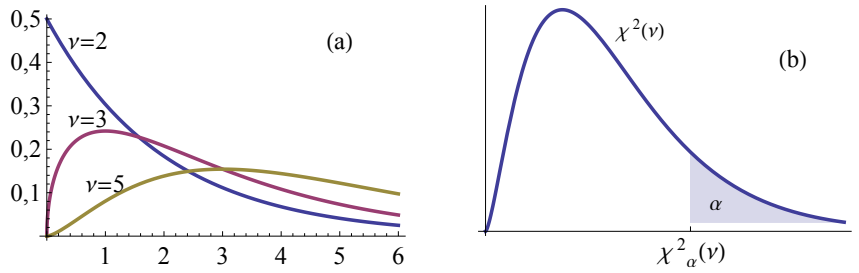
Pearsoni statistik (7.57) allub χ^2-jaotusele vabadusastmete arvuga ν:

$$\chi^2 \sim \chi^2(\nu). \tag{7.61}$$

Seda juhul, kui empiiriliste ja oodatavate sageduste summa on ühesugune:

$$\sum_{i=1}^{k} n_i^e = \sum_{i=1}^{k} n_i^o. \tag{7.62}$$

χ^2-jaotuse jaotustiheduse graafikud mõningate vabadusastmete korral on toodud joonisel 7.17 (a). Kuidas χ^2-jaotus tekib, seda võib lugeda lisast A.5. Võttes ette olulisuse nivoo α, saab leida kriitilise väärtuse, milleks on χ^2-jaotuse täiendkvantiil χ^2_α (joonis 7.17 (b)). Rõhutame, et kuna χ^2 saab olla ainult positiivne, siis kogu kriitiline piirkond on paremal. Kriitilised väärtused mõningate vabadusastmete arvude korral on toodud lisas B.4.



Joonis 7.17. (a) Erineva vabadusastmete arvuga $\chi^2(\nu)$-jaotuste graafikud. (b) $\chi^2(\nu)$-jaotuse kriitiline väärtus olulisuse nivoo α korral on täiendkvantiil $\chi^2_\alpha(\nu)$

Vabadusastmete arv

χ^2-jaotuse **vabadusastmete arv** teoreetilise jaotuse sobivuse testimisel

$$\nu = k - r - 1, \tag{7.63}$$

kus k on sageduste n_i^e arv (liidetavate arv summas (7.57)) ja r vastava teoreetilise jaotuse hinnatavate parameetrite arv, mis erinevate jaotuste korral on järgmine:

- ühtlane jaotus $r = 0$;
- binoomjaotus $r = 1$ (positiivse sündmuse tõenäosus p);
- Poissoni jaotus $r = 1$ (keskväärtus λ);
- eksponentjaotus $r = 1$ (parameeter λ);
- normaaljaotus $r = 2$ (keskväärtus μ ja standardhälve σ).

Tabelis 7.8 on sageduste arv $k = 8$. Poissoni jaotuse hinnatavate parameetrite arv $r = 1$ ja vabadusastmete arv valemist (7.63)

$$\nu = m - r - 1 = 8 - 1 - 1 = 6.$$

Teststatistiku kriitiline väärtus olulisuse nivool 0,05 vabadusastmete arvuga 6 on 12,59 (lisa B.4). Kuna tabeli 7.8 põhjal arvutatud teststatistik 10,375 (valem (7.58)) on väiksem kui kriitiline väärtus, võtame vastu nullhüpoteesi. Telefonikõnede arv viieminutilistes intervallides allub Poissoni jaotusele.

Tabelarvutuses on χ^2-testi kriitilise väärtuse leidmiseks funktsioon **CHISQ.INV.RT**. Argument *Probability* on olulisuse nivoo α ja argument *Deg_freedom* on vabadusastmete arv ν valemist (7.63).

Pideva juhusliku suuruse korral on χ^2-testi kasutamine mõnevõrra keerulisem. Empiiriliste sageduste leidmiseks tuleb variatsioonrida intervallida (vt alapeatükk 1.6 lk 31). Sagedusklassi sobiva laiuse leidmiseks võib kasutada Sturgesi valemit (1.1). Seejärel saab leida sagedusklassidele vastavad empiirilised sagedused n_i^e.

Oodatavate sageduste leidmiseks kasutatakse vastava pideva jaotuse jaotusfunktsiooni F. Kui i-nda sagedusklassi ülemine piir on u_i, siis tõenäosused, et juhusliku suuruse väärtus jääb vastavasse klassi, leitakse järgmiselt:

$$p_1 = P(x < u_1) = F(u_1), \tag{7.64}$$

$$\dots$$

$$p_i = P(u_{i-1} < x < u_i) = F(u_i) - F(u_{i-1}), \tag{7.65}$$

$$\dots$$

$$p_k = P(x > u_{k-1}) = 1 - F(u_{k-1}), \tag{7.66}$$

Esimesse klassi langemise tõenäosus (7.64) ja viimasesse klassi langemise tõenäosus (7.66) garanteerivad selle, et tõenäosuste summa

$$\sum_{i=1}^{k} p_i = 1. \tag{7.67}$$

Klassile i vastava oodatava sageduse määrab ära tõenäosus p_i ja vaatluste koguarv n:

$$n_i^o = np_i. \tag{7.68}$$

Kui empiirilised ja oodatavad sagedused on leitud, saame χ^2-statistiku leidmiseks kasutada valemit (7.57).

Näide 7.21. Xeroxi aktsia tulumäära allumine normaaljaotusele

N07Hüpoteesid
N7.21

Alapeatükis 5.10 tutvusime normaaljaotusega ning erinevate näidetega, kus juhuslik suurus allus normaaljaotusele. Joonisel 5.18 oli esitatud Xeroxi aktsia tulumäära jaotus ajavahemikul 8. veebruar–9. august 2013. Kasutame nüüd χ^2-testi kontrollimaks, kas Xeroxi aktsia tulumäär allus vaadeldaval perioodil normaaljaotusele.

1. Hüpoteesipaar:

   H_0: aktsia tulumäär vaadeldaval ajavahemikul allub normaaljaotusele;

   H_1: tulumäär ei allu normaaljaotusele.

2. Empiiriliste sageduste leidmiseks on tulumäära variatsioonrida intervallitud ning leitud sagedusklassidele vastavad empiirilised sagedused. Teoreetiliste sageduste leidmiseks kasutatakse normaaljaotuse jaotusfunktsiooni. Selleks on eelnevalt leitud valimi aritmeetiline keskmine $\bar{x} = 0{,}2078$ ja standardhälve $s = 1{,}5689$.

Esimesse klassi langemise tõenäosus on määratud selle klassi ülemise piiri jaotusfunktsiooniga (7.64):

$$P_1 = F(u_1) = F(-2{,}4) = 0{,}048.$$

Viimasesse klassi peab langema kõik ülejäänu, mis eelmistesse klassidesse ei langenud. Seepärast viimasesse klassi langemise tõenäosuse leidmiseks kasutame valemit (7.66):

$$P_8 = 1 - F(u_7) = 1 - F(3{,}72) = 0{,}013.$$

Ülejäänud klassidesse langemise sagedused leitakse valemi (7.65) põhjal. Empiiriliste ja teoreetiliste sageduste põhjal leitakse vastavalt valemile (7.57) parameetri empiiriline väärtus $\chi^2 = 6{,}867$.

Klassi nr i	Ülemine piir u_i	n_i^e	$F(u_i)$	p_i	n_i^o	$\frac{(n_i^e - n_i^o)^2}{n_i^o}$
1	−2,4	8	0,048	0,048	5,89	0,760
2	−1,38	8	0,156	0,108	13,12	1,997
3	−0,36	22	0,359	0,203	24,76	0,308
4	0,66	39	0,613	0,255	31,07	2,022
5	1,68	28	0,826	0,213	25,93	0,165
6	2,7	10	0,944	0,118	14,39	1,338
7	3,72	6	0,987	0,043	5,31	0,091
8	4,74	1		0,013	1,54	0,187

3. Klasside arv $m = 8$. Normaaljaotuse hinnatavate parameetrite arv $r = 2$. Vabadusastmete arv

$$\nu = m - r - 1 = 8 - 2 - 1 = 5.$$

Parameetri kriitilise väärtuse olulisuse nivool 0,05 vabadusastmete arvuga 5 leiame tabelarvutuses:

$$\chi^2{}_{kr} = \text{CHISQ.INV.RT}(0{,}05; 5) \approx 11{,}07.$$

4. Teststatistik ei lange kriitilisse piirkonda: $6{,}867 < 11{,}07$.
5. Võtame vastu nullhüpoteesi. Perioodil 8. veebruar–9. august 2013 polnud Xeroxi aktsia tulumäära jaotus normaaljaotusega vastuolus.

Kui me teame väärtpaberi oodatava tulumäära ja standardhälbe suurust ning väärtpaberi tulumäär järgib normaaljaotust, on võimalik leida, kui suure tõenäosusega võib tegelik tulu tulla suurem või väiksem etteantud väärtusest või kui suure tõenäosusega võib see jääda mingisse vahemikku (Sander, 1999). Analüüs näitab, et normaaljaotuse järgimine väärtpaberihindade korral sõltub vaadeldava ajavahemiku pikkusest.

Näide 7.22. Normaaljaotus väärtpaberiturul

2000. aastal Audentese Kõrgemas Ärikoolis kaitstud diplomitöös „Statistiliste meetodite kasutamine Eesti väärtpaberiturul“ uuriti mõningate väärtpaberite päevaste tulumäärade jaotust (Kiivet, 2000). Töö käigus testiti erinevate väärtpaberite korral järgmist hüpoteesipaari:

H_0: tulumäärad alluvad normaaljaotusele;
H_1: tulumäärad ei allu normaaljaotusele.

Testimiseks kasutati erineva pikkusega aegridasid. Kuna tulumäär on pidev suurus, tuli χ^2-testi jaoks kasutada intervallimist. Järgnevalt on toodud tulemuste koondtabel. Tabelist on näha, et vaadeldava perioodi lühenedes vahe χ^2 kriitilise ja empiirilise väärtuse vahel väheneb ning kolme väärtpaberi (Hansapank, Norma ja Eesti Ühispank) jaoks kahe kuu pikkuse perioodi korral on vastu võetud nullhüpotees.

Aktiva	Periood	Sagedus-klasside arv	Kriitiline väärtus	Empiiriline väärtus	Kehtiv hüpotees
HANSA	4 aastat	20	27,59	471	H_1
	1 aasta	20	27,59	31,5	H_1
	2 kuud	13	18,31	11,37	H_0
EÜP	4 aastat	20	27,59	502	H_1
	1 aasta	20	27,59	97	H_1
	2 kuud	20	27,59	18,32	H_0
NORMA	4 aastat	20	27,59	2 141,5	H_1
	1 aasta	20	27,59	56,6	H_1
	2 kuud	20	27,59	23	H_0
EMV	4 aastat	20	27,59	907,3	H_1
	1 aasta	20	27,59	546,3	H_1
	2 kuud	20	27,59	50,2	H_1
SAKU	4 aastat	20	27,59	2 036,5	H_1
	1 aasta	18	25	736,7	H_1
	2 kuud	16	22,36	65,4	H_1
HÜVITUS-FOND	4 aastat	20	27,59	43 443,3	H_1
	1 aasta	15	21,03	18 111,3	H_1
	2 kuud	16	22,36	207,3	H_1

Esitame empiirilise ja teoreetilise jaotuse sobivuse testimise protseduuri kokkuvõtlikult.

Jaotuse sobivuse testimine χ^2-testiga

1. Hüpoteesipaar:
   H_0: empiiriline ja teoreetiline jaotus langevad kokku, erinevus puudub;
   H_1: empiiriline ja teoreetiline jaotus erinevad oluliselt.
2. Vaatlusandmete põhjal leitakse vastava teoreetilise jaotusseaduse parameetrid.

3. Teoreetilisest jaotusseadusest leitakse erinevate väärtuste (või vahemikku langemise) tõenäosused.
4. Tõenäosuste ja valimi mahu põhjal leitakse oodatavad sagedused n_l^o. Peab kehtima võrdus (7.62).
5. Arvutatakse teststatistik

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i^e - n_i^o)^2}{n_i^o},$$

kus n_i^e on erinevate väärtuste esinemissagedused empiirilises jaotuses, n_i^o aga teoreetilise jaotusseaduse alusel leitud oodatavad esinemissagedused.
6. Statistiku kriitiline väärtus on χ^2-jaotuse täiendkvantiil $\chi^2_\alpha(\nu)$, kus α on olulisuse nivoo ja vabadusastmete arv ν leitakse valemist (7.63).
7. Võtta vastu

$$\begin{aligned} &\mathrm{H}_0, \text{ kui } \quad \chi^2 \le \chi^2_\alpha(\nu), \\ &\mathrm{H}_1, \text{ kui } \quad \chi^2 > \chi^2_\alpha(\nu). \end{aligned}$$

7.13. χ^2-test ja kahe tunnuse vaheline seos

Olgu meil läbi viidud ankeetküsitlus. Lisaks vastustele paluti küsitluse täitjatel märkida ka oma sugu. Soovime uurida, kas vastus mingile konkreetsele küsimusele sõltub vastaja soost, s.t kas on **seos** soo ja vastusevariantide jaotuse vahel. Ka siin saab uurimiseks kasutada χ^2-testi. Nullhüpoteesile vastavad oodatavad sagedused arvutatakse aga välja selle sama valimi põhjal.

Näiteks soovime kontrollida hüpoteesi, et konkreetse kauba meeldivus ostjale sõltub ostja soost. Hüpoteesi kontrollimiseks küsitleti juhuvalimit, milles oli 60 meest ja 90 naist. Vastusevariantide sageduste jaotus (empiirilised sagedused) on toodud tabelis 7.9.

N07Hüpoteesid
T7.9-11

Tabel 7.9. Empiirilised sagedused

	mehed	naised	KOKKU
meeldib	17	14	31
neutraalne	29	45	74
ei meeldi	14	31	45
KOKKU	60	90	150

Tabelit 7.9 nimetatakse esinemissageduste **risttabeliks** (*pivot table*) ehk kahemõõtmeliseks sagedustabeliks.

Hüpoteesipaar on järgmine:

H_0: ostja sugu ei mõjuta kauba meeldivust;

H_1: ostja sugu mõjutab kauba meeldivust.

χ^2-testi kasutamiseks tuleb leida nullhüpoteesile vastavad oodatavad sagedused. Kuna nullhüpotees on, et meeldivus ei sõltu ostja soost, peaks tunnuse „meeldivus“ väärtuste jaotus meeste ja naiste vahel olema sama, mis meeste ja naiste jaotus valimis. Meeste osakaal

$$\frac{60}{150} = 0{,}4$$

ja naiste osakaal

$$\frac{90}{150} = 0{,}6.$$

Need 31, kellele kaup meeldis, peaksid nullhüpoteesi kehtimisel jaotuma meesteks ja naisteks samamoodi nagu terves valimis. Oodatav meeste arv, kellele kaup meeldib, on siis

$$0{,}4 \cdot 31 = 12{,}4$$

ja oodatav naiste arv

$$0{,}6 \cdot 31 = 18{,}6.$$

Samamoodi peaksid jaotuma need 74, kes olid neutraalsed, ning need 45, kellele kaup ei meeldinud. Tehes vastavad arvutused, täidetakse ära oodatavate sageduste tabel. Veergudes ja ridades peavad kokku tulema samad arvud, mis empiiriliste sageduste tabelis. Oodatavad sagedused on toodud tabelis 7.10.

Tabel 7.10. Oodatavad sagedused

	mehed	naised	KOKKU
meeldib	12,4	18,6	31
neutraalne	29,6	44,4	74
ei meeldi	18,0	27,0	45
KOKKU	80	90	150

N07Hüpoteesid
T7.9-11

χ^2-statistiku leidmiseks võib nii empiirilised kui ka oodatavad sagedused kirjutada ühte tabelisse, nagu tehti seda eelmises alapeatükis.

Tabeli 7.11 viimases veerus olevate arvude summa on teststatistik

$$\chi^2 = \sum_{i=1}^{6} \frac{(n_i^e - n_i^o)^2}{n_i^o} = 4{,}346. \tag{7.69}$$

Kriitiline väärtus leitakse χ^2-jaotusest vabadusastmete arvuga ν.

Tabel 7.11. χ^2-statistiku arvutamine

N07Hüpoteesid T7.9-11

Klass	Empiiriline sagedus n_i^e	Oodatav sagedus n_i^o	$n_i^e - n_i^o$	$\frac{(n_i^e - n_i^o)^2}{n_i^o}$
mehed, meeldib	17	12,4	4,6	1,706
mehed, neutraalne	29	29,6	−0,6	0,012
mehed, ei meeldi	14	18,0	−4,0	0,889
naised, meeldib	14	18,6	−4,6	1,138
naised, neutraalne	45	44,4	0,6	0,008
naised, ei meeldi	31	27,0	4,0	0,593
			$\chi^2 =$	4,346

Vabadusastmete arv kahe tunnuse võrdlemisel

χ^2-jaotuse **vabadusastmete arv** kahe tunnuse võrdlemisel:

$$\nu = (n_1 - 1) \cdot (n_2 - 1), \tag{7.70}$$

kus n_1 on ühe ja n_2 teise tunnuse väärtuste arv.

Tunnusel „meeldivus“ on kolm väärtust, $n_1 = 3$, ja tunnusel „sugu“ on kaks väärtust, $n_2 = 2$. Vabadusastmete arv

$$\nu = (n_1 - 1) \cdot (n_2 - 1) = (3 - 1) \cdot (2 - 1) = 2.$$

Kriitiline väärtus olulisuse nivool 0,05 vabadusastmete arvuga 2 on täiendkvantiil $\chi^2_{0,05}(2) = 5,99$. Kuna teststatistik ei lange kriitilisse piirkonda, $4,346 < 5,99$, tuleb nullhüpotees vastu võtta. Olulisuse nivool 5% ei õnnestunud tuvastada seost kauba meeldivuse ja ostja soo vahel.

Kahe tunnuse väärtuste sageduste risttabelit saab kasutada juhul, kui kummalgi tunnusel ei ole väärtuste arv eriti suur. Järelikult ei saa intervallskaalas mõõdetud tunnuste korral seda testi kasutada. Sagedustabelis ei tohiks ükski sagedus olla väiksem kui 5.

χ^2-testi kasutamine

χ^2-testi kasutatakse kahe tunnuse vahel oleva seose testimiseks, kui mõlemad tunnused on kas nimiskaalas või järjestusskaalas. Nullhüpotees on, et seos puudub.

Tabelis 7.12 on esitatud tunnuse „meeldivus“ vastusevariantide oodatavate sageduste asemel vastuste osakaalud, eraldi mehed, naised ja kokku. Osakaalude leidmisel on lähtutud tabelist 7.10. Näeme, et variandi „meeldib“ esinemine oli meestel ja naistel ühesugune.

Samuti oli meestel ja naistel ühesugune ülejäänud vastusevariantide „neutraalne“ ja „ei meeldi“ jaotus. See tähendabki, et tunnus „sugu“ ei mõjuta meeldivust. Kui empiirilised osakaalud on samad, mis oodatavad tabelis 7.12, siis ka empiirilised ning oodatavad sagedused on võrdsed ning valemist (7.57) näeme, et $\chi^2 = 0$. Järelikult vastab nullhüpoteesile see, et nende kahe tunnuse vahel seos puudub.

Tabel 7.12. Nullhüpoteesile vastavad oodatavad osakaalud, seos puudub

	mehed	naised	KOKKU
meeldib	21%	21%	21%
neutraalne	49%	49%	49%
ei meeldi	30%	30%	30%
KOKKU	100%	100%	100%

Tabelarvutuses on selle χ^2-testi läbiviimiseks mugav kasutada funktsiooni **CHISQ.TEST** (*actual_range*; *expected_range*). Parameeter *actual_range* on empiiriliste sageduste piirkond ning *expected_range* oodatavate sageduste piirkond. Funktsioon leiab teststatistikule vastava olulisuse tõenäosuse p, mida võrreldakse olulisuse nivooga α. Kui $p > \alpha$, võetakse vastu nullhüpotees ja kui $p < \alpha$, siis sisukas hüpotees. Järelikult ei pea selle funktsiooni kasutamisel tegema arvutusi χ^2-statistiku leidmiseks nagu tabelis 7.11. Meeldivuse ja soo vahelise sõltuvuse testimise korral väljastab CHISQ.TEST tulemuse $p = 0{,}114$, mis on suurem kui $\alpha = 0{,}05$ ja järelikult võtame vastu nullhüpoteesi.

Funktsiooni CHISQ.TEST ei saa aga kasutada empiirilise ja teoreetilise jaotuse võrdlemisel, mida käsitlesime eelmises alapeatükis. Põhjuseks on see, et CHISQ.TEST kasutab vabadusastmete arvu (7.70), kuid empiirilise ja teoreetilise jaotuse võrdlemisel tuleb vabadusastmete arv leida valemist (7.63).

Kahe kvalitatiivse tunnuse vahelise seose testimine χ^2-testiga

1. Hüpoteesipaar:
   H_0: seos puudub;
   H_1: seos esineb.
2. Valimi põhjal koostatakse empiiriliste sageduste risttabel, kus read vastavad ühe tunnuse väärtustele ja veerud teise tunnuse väärtustele.
3. Empiirilise sagedustabeli alusel leitakse nullhüpoteesile vastav oodatavate sageduste risttabel.

4. Arvutatakse teststatistik:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i^e - n_i^o)^2}{n_i^o},$$

kus n_i^e on empiirilised sagedused ja n_i^o oodatavad sagedused.

5. Statistiku kriitiline väärtus on χ^2-jaotuse täiendkvantiil $\chi^2_\alpha(\nu)$, kus α on olulisuse nivoo ja vabadusastmete arv ν leitakse valemist (7.70).
6. Võtta vastu

$$\begin{aligned} &\mathrm{H}_0, \text{ kui } \quad \chi^2 \le \chi^2_\alpha(\nu) \ (p \ge \alpha); \\ &\mathrm{H}_1, \text{ kui } \quad \chi^2 > \chi^2_\alpha(\nu) \ (p < \alpha). \end{aligned}$$

Kui tabelarvutuses on ankeetküsitluse vastused sisestatud nii, et igas veerus on üks tunnus ja igal real ühe vastaja vastused, siis risttabeli koostamiseks saab Excelis kasutada vahendit *PivotTable* (vt lisa C.7).

N07Hüpoteesid
N7.23

Näide 7.23. Tööga rahulolu uuring

Aastal 2005 toimus OÜ-s Ehitus Service tööga rahulolu uuring, mille käigus küsitleti 139 töötajat. Küsimustik sisaldas viit alajaotust, milles kokku oli 40 väidet. Töötajad pidid neid väiteid hindama 5-pallises skaalas. Lisaks oli küsimustikus küsimused töötaja soo, ameti, töökoha (linn), töökogemuse ja haridustaseme kohta. (Pukk, 2005)

Töökogemuse mõõtmiseks kasutati 3-pallist skaalat:

1) kuni 1 aasta;
2) kuni 2 aastat;
3) üle 2 aasta.

Väide A3 oli: „Töötajad usaldavad juhtkonda ja peavad juhtkonna otsuseid organisatsiooni jaoks parimateks.“ Kasutades χ^2-testi, on võimalik analüüsida, kas vastus sellele küsimusele sõltub töökogemusest.

Risttabeli loomine näitas, et vastusevariante 1 ja 5 oli vähe valitud ning χ^2-testi jaoks ühendati need vastavalt variantidega 2 ja 4. Andmetabeli põhjal koostati Exceli vahendiga *PivotTabel* (vt lisa C.7) risttabel empiiriliste sagedustega.

	Väide A3			
Töökogemus	2	3	4	KOKKU
1	7	17	17	41
2	16	8	21	45
3	14	17	20	51
KOKKU	37	42	58	137

Selle tabeli põhjal on arvutatud nullhüpoteesile vastavad oodatavad sagedused, mis on esitatud järgmises tabelis.

	Väide A3			
Töökogemus	2	3	4	KOKKU
1	11,07	12,57	17,36	41
2	12,15	13,80	19,05	45
3	13,77	15,64	21,59	51
KOKKU	37	42	58	137

1. Hüpoteesipaar:
   H_0: hinnang väitele „Töötajad usaldavad juhtkonda...“ ei sõltu töötaja töökogemusest;
   H_1: hinnang väitele „Töötajad usaldavad juhtkonda...“ sõltub töötaja töökogemusest.
2. Tabelarvutuse funktsiooni CHISQ.TEST abil leitud olulisuse tõenäosus $p = 0{,}128 > 0{,}05$ ja võtame vastu nullhüpoteesi.
3. Järeldus: hinnang väitele „Töötajad usaldavad juhtkonda ja peavad juhtkonna otsuseid organisatsiooni jaoks parimateks“ ei sõltu töötaja töökogemusest.

Tavaliselt esitatakse uuringu läbiviimisel palju erinevaid küsimusi, mille seotust tausttunnustega saab uurida χ^2-testi abil. Tulemused on otstarbekas esitada kokkuvõtlikus tabelis, nii nagu järgmises näites.

Näide 7.24. Töötajate motiveerimine erinevates riikides

2004. aastal Audentese Ülikooli kaitstud bakalaureusetöös uuriti töötajate motiveerimist erinevate riikide (Eesti, Soome, Läti) väikeettevõtetes (Kala, 2004). Üheks hüpoteesiks oli, et väikeettevõtetes on töötajate motivatsioon riigiti erinev. Uuringu käigus viidi läbi küsitlus, kus paluti vastata erinevatele küsimustele, vastusevariante oli neli („jah“, „osaliselt“, „üldse mitte“, „ei oska öelda“). Iga küsimuse korral võrreldi vastuste jaotust erinevates riikides ja viidi läbi χ^2-test järgmise hüpoteesipaariga:

H_0: vastuste jaotus antud küsimusele on erinevates riikides ühesugune;
H_1: vastuste jaotus antud küsimusele on erinevates riikides erinev.

Nullhüpoteesi kehtimine tähendab, et tunnus „riik“ ei mõjuta küsimusele vastamist.

Kõikide küsimuste jaoks arvutati teststatistik χ^2. Lisaks leiti olulisuse tõenäosused p. Järgnevas tabelis on toodud tulemused esimese 10 küsimuse kohta. Tärnidega märgitud olulisuse tõenäosuse korral on nullhüpotees vastaval nivool ümber lükatud ja töötajate hinnang erinevates riikides erinev.

Küsimus „Kas Te ...	χ^2 empiiriline väärtus	Olulisuse tõenäosus p
1. ... olete rahul töötasuga, mida momendil teenite?	11,9	0,017 8 **
2. ... olete õiglaselt tasustatud?	27,5	0,000 1 ***
3. ... olete rahul töökeskkonnaga (valitsev õhkkond)?	13,7	0,008 3 *
4. ... olete rahul oma töörežiimiga (tööaeg)?	17,9	0,001 3 ***
5. ... olete rahul informatsiooni liikumisega ettevõttes?	16,6	0,010 8 **
6. ... tunnete ennast turvaliselt oma töökohal (kartus kaotada töö)?	30,2	0,000 04***
7. ... olete rahul oma töö sisuga?	10,0	0,124 3
8. ... tunnete soolist või ealist diskrimineerimist ettevõttes, kus töötate?	12,4	0,054 4 *
9. ... tunnete ennast ettevõttele vajaliku töötajana?	7,0	0,319 2
10. ... lähete meelsasti kaasa ettevõttes toimuvate uuendustega?	16,0	0,013 9 **

* oluline nivool 0,1;
** oluline nivool 0,05;
*** oluline nivool 0,01.

χ^2-testi alusel on otstarbekas kindlaks määrata statistilise seose olemasolu või selle puudumist. Statistik χ^2 ei sobi aga seose tugevuse hindamiseks, kuna selle statistiku väärtus võib olla kuitahes suur. Seose tugevust on sobivam hinnata mõne χ^2 alusel tuletatud **seosekordaja** abil. Mõned neist on järgnevalt toodud. Kõikides valemites on n valimi maht (sageduste koguarv) ja m sagedustabeli lühema külje pikkus (ridade või veergude arv).

Seosekordajad

Pearsoni kontingentsuskordaja (*Contingency coefficient*)

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}. \tag{7.71}$$

Kui tunnused on sõltumatud, siis $C = 0$. Suurim võimalik väärtus $C_{\max} = \sqrt{1 - \frac{1}{m}} < 1$. Puuduseks on, et suurim võimalik väärtus sõltub tabeli suurusest ja ainult lõpmatult suure tabeli korral ($m \to \infty$) on suurim väärtus 1.

Φ-kordaja (*Phi coefficient*)

$$\Phi = \sqrt{\frac{\chi^2}{n}}. \tag{7.72}$$

Kui tunnused on sõltumatud, siis $\Phi = 0$. Suurim võimalik väärtus $\Phi_{\max} = \sqrt{m-1}$. Sobib hästi kasutamiseks 2×2 tabelite korral, sest siis on suurim võimalik väärtus üks.

Craméri V-kordaja (*Cramérs V*)

$$C = \sqrt{\frac{\chi^2}{n(m-1)}}. \tag{7.73}$$

Sobib kasutamiseks suuremate kui 2×2 tabelite korral, $0 \leq C \leq 1$.

Kui tahetakse korraga võrrelda rohkem kui kahe nimiskaalas mõõdetud tunnuse vahelist seost, siis koostatakse ruutmaatriks, mille element (i, j) vastab Craméri V-kordaja väärtusele i-nda ja j-nda tunnuse vahel. See on analoogne korrelatsioonimaatriksile, mida kasutatakse intervallskaalas mõõdetud tunnuste korral (vt ptk 8.3).

7.14. Ühefaktoriline dispersioonanalüüs ANOVA

Keskväärtuste testimisel t-testiga saab korraga võrrelda kahe valimi keskväärtust. Valimite eristamiseks kasutatakse mingit tausttunnust (sugu, kaks vanusegruppi, hõivatud ja töötud, suured ja väikesed ettevõtted) ning öeldakse, et tausttunnusel on kaks erinevat väärtust või nivood. t-test võimaldab kontrollida, kas mõõdetud tunnuse ja tausttunnuse vahel eksisteerib seos: tausttunnuse väärtus mõjutab mõõdetud tunnuse väärtust (valimite keskmised on oluliselt erinevad, kehtib sisukas hüpotees).

Kui tausttunnusel on erinevaid nivoosid rohkem kui kaks, on ka valimeid ja vastavaid keskmisi rohkem kui kaks ning t-testi kasutades tuleks läbi viia paarikaupa testimine. Kolme valimi korral näiteks kolm korda, nelja valimi korral juba kuus korda.

N07Hüpoteesid
N7.25,27

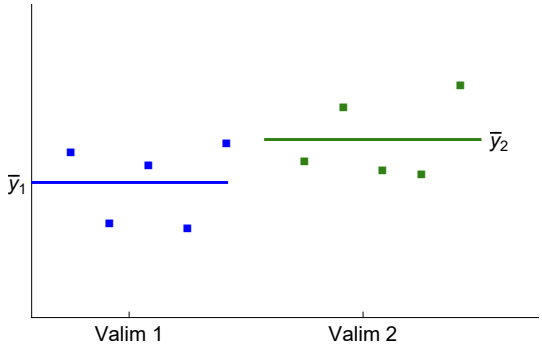
Näide 7.25. Akude võrdlemine

Digitaalfotoaparaatide tootjal tuleb oma uude mudelisse leida sobiv aku. Valida on kolme erineva tarnija vahel ja ettevõte otsustas testida nende kolme erineva tarnija akude mahutavust: millised akud peavad kõige kauem vastu. Testimiseks paluti kõigil tarnijatel saata proovipartii. Testimisel katsetati, mitu fotot saab täislaetud akuga teha, enne kui akupinge langeb teatud väärtuseni. Tabelis on toodud testimise tulemused.

Tarnija	A	B	C
Proovipartii suurus	18	21	19
Valimi keskmine	82,06556	80,66667	87,68421
Standardhälve	7,1247	7,5982	5,2287

Tarnija C proovipartiil on keskmiselt akude mahutavus kõige suurem, järgnevad tarnijad A ja B. Kas nende andmete põhjal võib väita, et akude mahutavus on tarnijatel erinev?
Tegemist on kolme üldkogumiga: tarnija A toodang, tarnija B toodang ja tarnija C toodang. Igast üldkogumist on võetud üks juhuvalim ja nende valimite põhjal tuleb otsustada, kas kogumite keskväärtused on erinevad või mitte. Tausttunnuseks on „tarnija“ ja sellel on kolm erinevat nivood A, B ja C.

Selleks, et võrrelda enam kui kahe üldkogumi keskmisi, kasutatakse **dispersioonanalüüsi** ehk ANOVA-t (*ANalysis Of VAriance*). Dispersioonanalüüsi võttis kasutusele 20. sajandi algul inglise statistik Ronald A. Fisher.

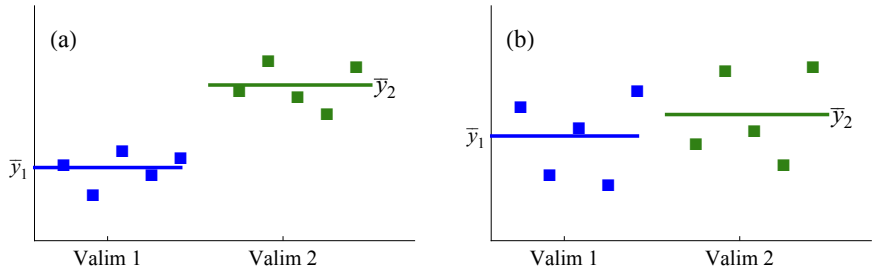


Joonis 7.18. Kaks valimit, mõlemad sisaldavad viit mõõtmistulemust. Jooned vastavad valimite keskmistele

Vaatame, millel põhineb dispersioonanalüüs. Joonisel 7.18 on toodud kaks valimit, mõlemad sisaldavad viit mõõtmistulemust (punktid

joonisel). Lisatud on ka horisontaaljooned, mis vastavad kummagi valimi keskmisele $\bar{y}_1$ ja $\bar{y}_2$. Tuleb otsustada, kas need valimid pärinevad ühest ja samas kogumist või mitte.

Joonisel 7.18 toodud valimite põhjal on sellele küsimusele raske vastata. Kui vaadata aga joonisel 7.19 (a) toodud valimeid, siis võib intuitiivselt väita, et need valimid on võetud erineva keskväärtusega kogumitest. Joonisel 7.19 (b) toodud valimite korral ollakse tavaliselt aga nõus väitega, et need võivad pärineda ühest ja samast kogumist.



Joonis 7.19. Diagrammil (a) esitatud valimid on tõenäoliselt erinevatest kogumitest, diagrammil (b) aga tõenäoliselt ühest ja samast kogumist

Mille alusel võetakse vastu intuitiivsed otsused joonisel 7.19 toodud situatsioonide (a) ja (b) korral? Visuaalselt võrreldakse erinevust valimite keskmiste vahel valimite sisese varieerumisega. Diagrammil (a) on erinevus valimite keskmiste vahel suurem kui varieeruvus valimite sees. Seepärast tehakse otsus, et need valimid pärinevad erineva keskväärtusega kogumitest. Diagrammil (b) on aga erinevus keskmiste vahel väike, võrreldes valimite sisese varieerumisega, ning need valimid võivad pärineda ühesuguse keskväärtusega kogumist.

Visuaalselt tehtud hinnangud on küllaltki subjektiivsed. Järgnevalt leiame **objektiivse kriteeriumi**, mis võtaks arvesse nii valimite keskmiste erinevust kui ka varieerumist valimite sees.

Olgu meil K valimit, mille jaoks on leitud valimeid iseloomustavad statistilised näitajad:

	Valim 1	Valim 2	...	Valim K
	y_{11}	y_{21}	...	y_{K1}
	y_{12}	y_{22}	...	y_{K2}
	...	...	...	...
Valimi keskmine	$\bar{y}_1$	$\bar{y}_2$	...	$\bar{y}_k$
Standardhälve	s_1	s_2	...	s_K
Valimi maht	n_1	n_2	...	n_K

Valimite **üldkeskmine** on kaalutud keskmine kõigi valimite keskmistest:

Üldkeskmine

$$\bar{y} = \frac{\sum_{i=1}^{K} n_i \bar{y}_i}{\sum_{i=1}^{K} n_i}, \tag{7.74}$$

kus i-nda valimi maht n_i on selle valimi keskmise $\bar{y}_i$ kaal ja $\sum_{i=1}^{K} n_i = n$ on elementide arv kõikides valimites kokku.

Näites 7.25 toodud valimite korral üldkeskmine valemist (7.74):

$$\bar{y} = \frac{18 \cdot 82{,}05556 + 21 \cdot 80{,}66667 + 19 \cdot 87{,}68421}{18 + 21 + 19} \approx 83{,}39655.$$

Rühmadevaheline hajumine ja SST

Valimite keskmiste varieerumist üldkeskmise ümber ehk **rühmadevahelist hajumist** (*variation between groups*) iseloomustab hälvete ruutude kaalutud summa SST (*Sum of Squares for Treatments*):

$$\begin{aligned} SST &= n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2 + ... + n_k(\bar{y}_k - \bar{y})^2 = \\ &= \sum_{i=1}^{K} n_i(\bar{y}_i - \bar{y})^2. \end{aligned} \tag{7.75}$$

Näites 7.25 toodud valimite korral annab vastav arvutus

$$\begin{aligned} SST &= 18 \cdot (82{,}05556 - 83{,}39655)^2 + 21 \cdot (80{,}66667 - 83{,}39655)^2 + \\ &+ 19 \cdot (87{,}68421 - 83{,}39655)^2 = 538{,}1629. \end{aligned}$$

Hälvete ruutude kaalutud summa jagatakse läbi vabadusastmete arvuga $K-1$, kus K on valimite arv. Sel teel saadud suurust nimetatakse **keskruuduks** ja tähistatakse MST (*Mean Square of Treatments*):

Keskruut MST

$$MST = \frac{SST}{K-1}. \tag{7.76}$$

Miks on vabadusastmete arv $K-1$? Hälvete ruutude summas (7.75) on K liidetavat. Sõltumatuid liidetavaid on aga $K-1$, sest üldkeskmise $\bar{y}$ arvutusvalemist (7.74) järeldub, et

$$\begin{aligned} &n_1(\bar{y}_1 - \bar{y}) + n_2(\bar{y}_2 - \bar{y}) + ... + n_K(\bar{y}_K - \bar{y}) = \\ &= (n_1\bar{y}_1 + n_2\bar{y}_2 + ...n_K\bar{y}) - (n_1 + n_2 + ... + n_K)\bar{y} = 0. \end{aligned} \tag{7.77}$$

Seega, kui meil on $K-1$ liidetavat olemas, siis K-ndat liidetavat enam sõltumatult valida ei saa, kuna kehtib kitsendav tingimus (7.77). Üks vabadusaste kaob ära üldkeskmise kasutamise tõttu.

Näites 7.25 oli meil valimite arv 3 ning vastav keskruut

$$MST = \frac{538{,}1629}{3-1} = 269{,}0815.$$

Valimite sees esinev hajumine on põhjustatud sellest, et tegemist on juhuvalimitega, seega — juhuslikest teguritest. Valimite sees esinevat varieerumist ehk **rühmasisest hajuvust** (*within groups*) kirjeldab hälvete ruutude summa SSE (*Sum of Squared Errors*). Selle summa leidmine käib järgmiselt:

Rühmasisene hajumine ja SSE

$$\begin{aligned} SSE &= (y_{11}-\bar{y}_1)^2+(y_{12}-\bar{y}_1)^2+...+(y_{1n_1}-\bar{y}_1)^2+ \quad \text{valim 1}\\ &+ (y_{21}-\bar{y}_2)^2+(y_{22}-\bar{y}_2)^2+...+(y_{2n_2}-\bar{y}_2)^2+ \quad \text{valim 2}\\ &\quad .....\\ &+ (y_{K1}-\bar{y}_K)^2+(y_{K2}-\bar{y}_K)^2+...+(y_{Kn_K}-\bar{y}_K)^2 \quad \text{valim } K. \end{aligned}$$

Arvestades, et i-nda valimi dispersiooni s_i^2 jaoks kehtib valem

$$s_i^2=\frac{(y_{i1}-\bar{y}_i)^2+(y_{i2}-\bar{y}_i)^2+...+(y_{in_i}-\bar{y}_i)^2}{n_i-1},$$

võime vastava hälvete ruutude summa esitada valimite dispersioonide kaudu:

$$SSE=(n_1-1)s_1^2+(n_2-1)s_2^2+...+(n_K-1)s_K^2=\sum_{i=1}^{K}(n_i-1)s_i^2. \quad (7.78)$$

Näites 7.25 toodud valimite korral annab vastav arvutus

$$\begin{aligned} SSE &= (18-1)\cdot 7{,}124706^2+(21-1)\cdot 7{,}598245^2+\\ &\quad +(19-1)\cdot 5{,}228688^2=2509{,}716. \end{aligned}$$

Vabadusastmete arv on nüüd $n-K$, sest kõikides valimites kokku on meil n mõõtmistulemust ja nende põhjal on leitud K keskmist, mis vähendabki vabadusastmete arvu K võrra. Vastav keskruut MSE (*Mean Square of Errors*):

Keskruut MSE

$$MSE=\frac{SSE}{n-K}. \quad (7.79)$$

Näites 7.25 oli meil mõõtmistulemusi kokku $n=18+21+19=58$ ning vastav keskruut

$$MSE=\frac{2509{,}716}{58-3}=45{,}6312.$$

Nüüd on meil olemas

- **seletatud hajumist** kirjeldav suurus MST (näites 7.25 on see põhjustatud sellest, et valimid olid võetud erinevatelt tarnijatelt);
- **seletamata hajumist** kirjeldav suurus MSE (näites on see põhjustatud iga üksiku valimi sees toimuvast akude mahutavuse kõikumisest).

Nende suhe ongi teststatistik, mida saab kasutada joonistel 7.18 ja 7.19 toodud situatsioonide objektiivseks hindamiseks:

$$F = \frac{\text{seletatud ehk rühmadevaheline hajumine}}{\text{seletamata ehk rühmasisene hajumine}}.$$

Dispersioonanalüüsi F-statistik

Dispersioonanalüüsi F-statistik on seletatud hajumist ja seletamata hajumist kirjeldavate keskruutude suhe:

$$F = \frac{MST}{MSE}. \tag{7.80}$$

Valemist (7.80) leitud statistik allub samale F-jaotusele ehk Fisheri jaotusele, mida kasutati kahe valimi dispersioonide võrdlemisel. Statistiku empiirilist väärtust võrreldakse F-jaotuse täiendkvantiiliga, mis on kriitiliseks väärtuseks, ja võrdluse põhjal võetakse vastu otsus nullhüpoteesi kehtivuse või tagasilükkamise kohta.

Kui rühmadevaheline hajumine on rühmasisese hajumisega võrreldes piisavalt suur, on F väärtus kriitilisest suurem ning võetakse vastu sisukas hüpotees. Tegemist on ühepoolse hüpoteesiga. Otsuse tegemine võib põhineda kas teststatistiku F empiirilise väärtuse võrdlemisel kriitilise väärtusega F_{kr} või vastava olulisuse tõenäosuse p võrdlemisel olulisuse nivooga α.

Dispersioonanalüüsi tulemused tuuakse tavaliselt ära kokkuvõtvas tabelis 7.13.

Tabel 7.13. Dispersioonanalüüsi tabel

Varieeruvuse allikas	Hälvete ruutude summa SS	Vabadus-astmete arv df	Keskruut MS	F-statistik
Rühmade-vaheline	$SST = \sum_{i=1}^{K} n_i(\bar{y}_i - \bar{y})^2$	$K-1$	$MST = \frac{SST}{K-1}$	$F = \frac{MST}{MSE}$
Rühmasisene	$SSE = \sum_{i=1}^{K}(n_i-1)s_i^2$	$n-K$	$MSE = \frac{SSE}{n-K}$	
Üldine	$SST + SSE$	$n-1$		

Programmis Excel on ühefaktorilise dispersioonanalüüsi läbiviimiseks olemas spetsiaalne vahend *Anova: Single Factor* analüüsivahendite komplektis *Data Analysis* (vt ka lisa C.8). Tabelis *SUMMARY* esitatakse andmed valimite ehk gruppide kohta: valimi maht *Count*,

väärtuste summa *Sum*, valimi aritmeetiline keskmine *Average* ja valimi dispersioon *Variance*. Dispersioonanalüüsi arvutuste tulemused esitatakse tabelis ANOVA: hälvete ruutude summad SS, vabadusastmete arvud df, keskruudud MS, F-statistiku empiiriline väärtus F, sellele vastav olulisuse tõenäosus *P-value* ja F-statistiku kriitiline väärtus valitud olulisuse nivool $Fcrit$.

Näites 7.25 toodud andmete dispersioonanalüüsil saadakse Excelis tabel 7.14. Kuna F-statistiku empiiriline väärtus $F \approx 5{,}9$ on suurem kui kriitiline $Fcrit \approx 3{,}16$, on nullhüpotees ümber lükatud. Samale järeldusele jõuame, kui vaatame olulisuse tõenäosust $p \approx 0{,}0048 < 0{,}05$. Järelikult on erinevate tarnijate akude mahutavus erinev. Võib ka öelda, et aku mahutavus sõltub tunnusest „tarnija“.

Tabel 7.14. Näite 7.25 dispersioonanalüüsi tabel

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
A	18	1477	82,06	50,76
B	21	1694	80,67	57,73
C	19	1666	87,68	27,34

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	538,2	2	269,1	5,90	0,0048	3,16
Within Groups	2509,7	55	45,6			
Total	3047,9	57				

N07Hüpoteesid
N7.25,27

Dispersioonanalüüs kokkuvõtlikult

1. Hüpoteesipaar:
   H$_0$: $\mu_1 = \mu_2 = \ldots = \mu_K$, funktsioontunnuse Y keskväärtused on kõigis rühmades võrdsed ehk funktsioontunnuse keskväärtus on kogumites ühesugune.
   H$_1$: Leidub vähemalt kaks rühma i ja j, mille korral $\mu_i \neq \mu_j$, s.t rühmakeskmiste hulgas leidub oluliselt erinevaid keskmisi. Funktsioontunnuse keskväärtus on kogumites erinev.
2. Hüpoteeside kontrollimiseks leitakse F-statistik (7.80).
3. Võtta vastu

   H$_0$, kui $\quad F \leq F_{kr}$ $(p \geq \alpha)$;
   H$_1$, kui $\quad F > F_{kr}$ $(p < \alpha)$.

Faktor ja funktsioon-tunnus

Ühefaktorilise dispersioonanalüüsi korral on tegemist kahe tunnuse vahelise seose analüüsimisega. Üks tunnus on funktsioontunnus ning teine argument- ehk tausttunnus, mida nimetatakse **faktoriks**. Funktsioontunnus peab olema intervallskaalas, faktor on aga kvalitatiivne tunnus ning võib olla kas nimiskaalas või järjestusskaalas. Faktori võimalikud väärtused on faktori nivood ehk **tasemed**. Näites 7.25 on faktoriks tarnija (tasemed on tarnijad A, B ja C) ja funktsioontunnuseks on akude mahutavus ning uuritakse, kas akude mahutavus sõltub vastavast faktorist, s.t tarnijast.

Faktor X mõjub tunnusele Y, kui tunnuse Y keskmiste erinevus eri rühmades (valimites) on põhjustatud faktori erinevatest tasemetest. Sama testitakse keskväärtuste testimisel sõltumatute valimite t-testiga, kuid siis on võimalik korraga võrrelda vaid kaht valimit.

Kui võetakse vastu nullhüpotees, siis olemasolevad erinevused valimite keskväärtustes ei ole tõenäoliselt põhjustatud faktori erinevatest tasemetest, vaid juhuslikest põhjustest. Faktortunnus **ei mõjuta** funktsioontunnuse väärtusi. Kui nullhüpotees on ümber lükatud ning võetakse vastu sisukas hüpotees, siis erinevus valimite keskväärtustes on tõenäoliselt põhjustatud faktori erinevatest tasemetest, faktortunnus **mõjutab** funktsioontunnuse väärtusi.

Dispersioon-analüüsi kasutamine

> Dispersioonanalüüsi kasutatakse keskväärtuste võrdlemisel siis, kui on vaja võrrelda rohkem kui kaht kogumit (valimite arv on suurem kui 2).

Faktortunnusel võib olla palju erinevaid tasemeid. Kui mõõtmised on tehtud kõigil võimalikel tasemetel, on tegemist **fikseeritud faktoriga**. Kui mõõdetud tasemed on juhuslik valim faktori kõikvõimalike tasemete hulgast, on meil **juhuslik faktor**.

Dispersioonanalüüsi võib iseloomustada ka mõõtmiste arvu alusel. Kui mõõtmiste arv faktori igal tasemel (ehk erinevate valimite maht) on

- ühesugune — tasakaalustatud mudel;
- erinev — tasakaalustamata mudel.

Näites 7.25 olid erinevatelt tarnijatelt saadud proovipartiid erineva suurusega, tegemist oli tasakaalustamata mudeliga. Kui kõigi proovipartiide (valimite) suurus oleks olnud ühesugune, oleks tegemist olnud tasakaalustatud mudeliga. Proovipartiid olid olemas kõigilt võimalikelt tarnijatelt, s.t kõigi võimalike faktori tasemete jaoks, ning tegemist oli fikseeritud faktoriga. Kui aga näiteks tarnijaid oleks olnud kokku kümme ja juhuslikult valiti välja kolm, kelle proovipartiisid testiti, oleks tegemist olnud juhusliku faktoriga.

Näide 7.26. Kreeka hotellide rahandussuhtarvude võrdlemine

Ajakirjas The Journal of Hospitality Financial Management 2011. aastal ilmunud artiklis analüüsiti Kreeka hotellide kapitali struktuuri ja kasumlikkust vahetult enne 2008. aasta majanduskriisi aastatel 2005–2007 (Diakomihalis, 2011). Valimis oli 146 hotelli, nendest 21% olid viietärni hotellid, 20%-l oli neli tärni, 35%-l kolm ja 25%-l kaks tärni. Erineva tärnide arvuga hotellide rahandussuhtarvude võrdlemiseks kasutati dispersioonanalüüsi. Järgnevas tabelis on toodud dispersioonanalüüsi tulemused mõningate rahandussuhtarvude korral. Näeme, et kapitali struktuuri iseloomustavad suhtarvud võlakordaja (kohustused jagatud kogukapitaliga) ning finantsvõimenduse kordaja (keskmine vara jagatud keskmise omakapitaliga) on erineva arvu tärnidega hotellidel erinevad, mõlema korral on olulisuse tõenäosus $p < 0{,}05$ ning nullhüpotees on ümber lükatud. Kasumlikkuse näitaja brutokasumi marginaal aga ei erine, sest $p > 0{,}05$ ja nullhüpotees tuleb vastu võtta.

Rahandus-suhtarv	Varieeruvuse allikas	SS	df	MS	F	p
Võlakordaja	Rühmade-vaheline	1,554	3	0,518	16,3	$2{,}37 \cdot 10^{-9}$
	Rühmasisene	5,37	169	0,032		
	Üldine	12,6	172			
Finants-võimenduse kordaja	Rühmade-vaheline	71,85	3	23,95	6,80	0,000287
	Rühmasisene	419,4	119	3,524		
	Üldine	12,6	172			
Brutokasumi marginaal	Rühmade-vaheline	0,116	3	0,039	0,525	0,665
	Rühmasisene	12,5	169	0,074		
	Üldine	12,6	172			

Keskmiste mitmene võrdlemine

Dispersioonanalüüs ei anna vastust küsimusele, **millis(t)e kogumi(te)** keskväärtus on oluliselt erinev. Selle leidmiseks tuleb läbi viia **keskväärtuste mitmene võrdlemine** (*multiple comparison*), s.t võrrelda kogumite keskväärtusi paarikaupa. Testi, mis järgneb, nimetatakse ka *post-hoc* testiks.

1. Kui F-testi tulemusel erinevus **ei ole oluline**, pole mitmest võrdlemist vaja läbi viia.
2. Kui F-testi tulemusel erinevus **on oluline**, võib minna edasi ja viia läbi mitmene võrdlemine.

Olgu uuritaval faktoril kolm taset. Siis on vaja testida kolme hüpoteesipaari:

$$\begin{array}{lccc} \mathrm{H}_0\colon & \mu_1 = \mu_2, & \mu_1 = \mu_3, & \mu_2 = \mu_3; \\ \mathrm{H}_1\colon & \mu_1 \neq \mu_2. & \mu_1 \neq \mu_3. & \mu_2 \neq \mu_3. \end{array}$$

Mitmese võrdlemise meetodeid on palju. Näiteks Fisheri, Bonferroni, Tukey ja Scheffe testid (lähemalt vt Parring, Vähi ja Käärik, 1997).

Fisheri LSD test

Fisheri LSD-testi (*Least-Significant-Difference Test*) aluseks on varem vaadeldud t-test keskväärtuste võrdlemiseks. Toimub valimite paarikaupa testimine, mis sisaldab järgmisi samme:

1. Hüpoteesipaar
$$\begin{aligned} \mathrm{H}_0 &: \mu_1 = \mu_2, \\ \mathrm{H}_1 &: \mu_1 \neq \mu_2. \end{aligned}$$
2. t-statistiku väärtus leitakse valemist
$$t = \frac{\bar{y}_1 - \bar{y}_2}{se}, \tag{7.81}$$
kus $\bar{y}_1$ ja $\bar{y}_2$ on valimite keskmised ning
$$se = \sqrt{MSE \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}. \tag{7.82}$$
Keskruut MSE on leitud eelnevalt dispersioonanalüüsi käigus (valem (7.79)).
3. Kriitiline väärtus on t-jaotuse täiendkvantiil $t_{\alpha/2}(n-K)$ olulisuse nivoo α ja vabadusastmete arvuga $n-K$, kus K on dispersioonanalüüsis kasutatud valimite arv. Vabadusastmete arv ei sõltu sellest, milliseid valimipaare võrreldakse, sest standardvea arvutamisel kasutatakse kõikide valimite andmeid.
4. Vastu võetakse
$$\begin{array}{ll} \mathrm{H}_0, \text{ kui} & |t| \leq t_{\alpha/2}(n-K); \\ \mathrm{H}_1, \text{ kui} & |t| > t_{\alpha/2}(n-K). \end{array}$$

Näide 7.27. Akude mahutavus ja Fisheri LSD-test

N07Hüpoteesid
N7.25,27

Fisheri LSD-testi läbiviimisel näites 7.25 toodud valimite korral saadi järgmised tulemused:

Tarnijad	A ja B	A ja C	B ja C
Nullhüpotees H_0	$\mu_1 = \mu_2$	$\mu_1 = \mu_3$	$\mu_2 = \mu_3$
Sisukas hüpotees H_1	$\mu_1 \neq \mu_2$	$\mu_1 \neq \mu_3$	$\mu_2 \neq \mu_3$
Parameetri empiiriline väärtus	0,640	−2,533	−3,281
Olulisuse nivoo	5%	5%	5%
Parameetri kriitiline väärtus	2,004	2,004	2,004
Vastu võetud hüpotees	H_0	H_1	H_1

Lõppjäreldused:

1. Tarnijate vahel on oluline erinevus. ANOVA F-test näitas, et erinevatel tarnijatel on akude mahutavus oluliselt erinev.
2. Tarnijalt C saadud akude mahutavus on oluliselt erinev tarnijate A ja B akude mahutavusest.
3. Tarnijatelt A ja B saadud akude mahutavuses olulist erinevust ei ole.

Järelikult tuleb eelistada tarnija C akusid. Kui aga mingil põhjusel tarnijaga C tarnelepingut sõlmida ei saa, siis tarnijate A ja B vahel valiku tegemiseks tuleb kasutada mõnd teist kriteeriumit.

7.15. Sobiva testi valik

Selles peatükis vaatlesime vaid mõningaid sagedamini kasutatavaid teste. Lisaks on tuntud veel Manni-Whitney test sõltumatute valimite korral, Wilcoxoni astakmärgitest sõltuvate valimite korral, Kolmogorovi-Smirnovi test jaotuste võrdlemiseks ja kahefaktoriline dispersioonanalüüs. Nendega tutvumiseks võib kasutada õpikut „Statistilise andmetöötluse algõpetus“ (Parring, Vähi ja Käärik, 1997).

Sõltumatu ja sõltuv tunnus

Nende testide korral, kus on tegemist kahe või rohkema valimiga, eksisteerib alati mingi tunnus, mille alusel toimub valimitesse jagamine. Seda tunnust nimetatakse **sõltumatuks tunnuseks** või faktoriks (dispersioonanalüüsi korral). Teine tunnus, mille väärtuseid mõõdetakse ja testitakse, on **sõltuv tunnus**. Testimisel kontrollitakse, kas sõltuv tunnus sõltub (H_1) või ei sõltu (H_0) sõltumatust tunnusest. Näites 7.8 oli sõltumatuks tunnuseks riigi staatus Euroopa Liidu suhtes: Euroopa Liidu liikmesriik või kandidaatriik. Sõltuvad tunnused olid kulutused haridusele (% RKP-st), oodatav eluiga, imikute suremus jt. Näidetes 7.9 ja 7.10 oli sõltumatuks tunnuseks ettevõtte suurus ja sõltuvateks tunnusteks naisjuhtide tööstaaž ning sissetulek. Näites 7.12 oli sõltumatuks tunnuseks aasta ning sõltuvaks tunnuseks ettevõtte müügitulu ühe töötaja kohta.

Vaadeldud testide hulgast valiku tegemisel tuleb kindlaks teha

1) mitu väärtust ehk taset on sõltumatul tunnusel (faktoril);
2) millist skaalat on kasutatud sõltuva tunnuse mõõtmisel.

Tabelis 7.15 on esitatud üldised juhtnöörid sobiva testi valikuks.

Tabel 7.15. Testi valik

Sõltumatu tunnus ehk faktor	Sõltuv tunnus	Test
Puudub (1 kogum)	intervallskaalas	t-test, üks valim
	järjestus- või intervall-skaalas	märgitest
	kaheväärtuseline, suur valim	osakaalu testimine, üks valim
	kaheväärtuseline, väike valim	märgitest
2 väärtust (taset), sõltumatud valimid	intervallskaalas	t-test, sõltumatud valimid (eelnev dispersioonide F-test)
	kaheväärtuseline, suur valim	2 kogumi osakaalude võrdlemine
2 väärtust (taset), sõltuvad valimid	intervallskaalas	t-test, sõltuvad valimid
	järjestus- või intervall-skaalas	märgitest
2 või rohkem väärtust (taset)	intervallskaalas	ühefaktoriline dispersioonanalüüs
	nimi- või järjestus-skaalas	χ^2-test

Tabelis esitatud valikutele lisaks vaatlesime ka dispersioonide testimist F-testiga ning jaotuse sobivuse testimist χ^2-testiga.

7.16. Ülesanded

Nullhüpotees ja sisukas hüpotees, vead

7.1. Ajakirja Journal of Psychology and Aging 2002. aasta maikuu numbris väideti, et vanematel töötajatel (vanus suurem kui 45) on keskmine tööga rahulolu 5-pallisel skaalal 4,3 palli. Formuleerida nullhüpotees ja sisukas hüpotees, kui soovime testida ajakirjas toodud väidet. VASTUS lk 670.

7.2. Kirde saia pakendil on märgitud: „Netomass 330 g“. Tarbijakaitsele pühendatud ajakiri soovib kontrollida, kas see vastab tõele. Formuleerida nullhüpotees ja sisukas hüpotees. VASTUS lk 670.

7.3. Turundusplaani koostamisel tuleks kindlaks teha, kas restorani külastajate keskmine vanus on oluliselt erinev linna elanike keskmisest

vanusest. Formuleerida nullhüpotees ja sisukas hüpotees. VASTUS lk 670.

7.4. Järgmiste väidete korral otsustada, kas tegemist on nullhüpoteesiga või sisuka hüpoteesiga.

1. Keskmine tootlikkus ühe töötaja kohta on väike- ja suurettevõtetes ühesugune.
2. Finantsvahendusega tegelevates ettevõtetes on keskmine palk suurem kui tööstusettevõtetes.
3. Ameerikas toodetud autode bensiinikulu on keskmiselt suurem kui Euroopas toodetud autodel.
4. Toidukaupade hinnad on linnades A ja B ühesugused.
5. Madalama haridustasemega tarbijad kulutavad riietele vähem kui kõrgema haridustasemega tarbijad.
6. Teenindussaali sisenevate inimeste arv ajaühikus allub Poissoni jaotusele.
7. Erinevas vanuses inimesed reageerivad veebilehtedel olevatele reklaamidele erinevalt.

VASTUS lk 670.

7.5. Uue toote tasuvusanalüüs näitab, et toode tasub end ära, kui vähemalt 25% praegustest tarbijatest hakkab seda tarbima. Hüpoteesi kontrollimiseks tuleb läbi viia tarbijauuring. Formuleerida nullhüpotees ja sisukas hüpotees. Milline on selle uuringu korral I liiki viga ja milline II liiki viga? VASTUS lk 670.

7.6. Olgu uuritavaks tunnuseks ühe detaili valmistamiseks kulutatav aeg, mille normatiivne väärtus on 40 sekundit. Ratsionaliseerimisettepanekus on välja pakutud võtted, mille abil loodetakse vähendada detaili valmistamise keskmist aega. Kontrollimaks, kas eesmärk on saavutatud, tehakse vaatlus, mille käigus registreeritakse 70 juhuslikult välja valitud detaili valmistamiseks kulunud aeg. Püstitada nullhüpotees ja sisukas hüpotees. Milline on selle testimise korral I liiki viga ja milline II liiki viga? VASTUS lk 670.

Keskväärtuse testimine z-testiga

7.7. Ettevõte X on tegutsenud juba üle 50 aasta ja seal töötab üle 1000 töötaja. Ettevõtte juhtkond on alati rõhutanud seda, et nende töötajatel on pikaajaline töökogemus. Uute töötajate värbamiseks koostatud reklaamvoldikus kirjutatakse: „Meie töötajate keskmine tööstaaž on 20 aastat.“ Kuna juhtkond pole kindel, kas see väide on õige, viiakse läbi küsitlus 50 juhuslikult valitud töötaja hulgas. Küsitluse tulemusena saadakse keskmiseks tööstaažiks 19 aastat standardhälbega 2 aastat. Kas vastav väide võib reklaamvoldikusse jääda või tuleb seda muuta? VASTUS lk 670.

7.8. Põllumajandus- ja energiaseadmete edasimüüjad sõltuvad üpris tihti oma peamistest tarnijatest. Tarnijad soovivad tihti hoida edasi-

müüjaid oma kontrolli all. Et uurida, kui tugevasti edasimüüjad sõltuvad tarnijatest, viidi 1986. aastal USA-s läbi vastav uuring. Uuringuks võeti ette organisatsiooni National Farm and Power Equipment Dealers Association kuuluvate liikmete nimekiri ja moodustati juhuvalim 800 edasimüüjast, kellele saadeti vastav küsimustik. Vastas 226 ettevõtet (28,3%). Uuringu tulemused on avaldatud ajakirjas Academy of Management Journal (Provan ja Skinner, 1989). Muuhulgas leiti, et keskmine tarnijate arv ühe edasimüüja kohta oli 3,12 standardhälbega 1,91. Kontrollida hüpoteesi, et keskmine tarnijate arv ühe edasimüüja kohta on suurem kui kaks. Kontrollimiseks kasutada olulisuse nivood 5%. VASTUS lk 670.

7.9. Tööpingil toodetakse naelu pikkusega 1 toll (2,54 cm). Sellele pikkusele mittevastavad naelad, nii pikemad kui ka lühemad, saadab tellija tagasi. Et tagasisaadetava toodangu kogus oleks minimaalne, viib tootja aeg-ajalt läbi kvaliteedikontrolli. Selleks võetakse toodangu hulgast juhuvalim ja mõõdetakse valimisse sattunud naelad täpse mõõtevahendiga üle. Eelmise kontrolli ajal võeti liinilt juhuslikult 50 naela, mille pikkuste aritmeetiline keskmine oli 1,02 tolli ja standardhälve 0,04 tolli. Kas valimi põhjal võib järeldada, et tööpink vajab seadistamist? Kontrollida olulisuse nivool 0,01. VASTUS lk 670.

7.10. Ettevõte testib uut kaupade kättetoimetamissüsteemi. Vana süsteemi korral oli keskmine kättetoimetamisaeg 2,38 päeva. Uue süsteemi testimisel näitas 48 juhuslikult valitud tellimuse jälgimine, et kättetoimetamisaeg oli keskmiselt 2,15 päeva standardhälbega 0,80. Kontrollida, kas uus süsteem on parem, kasutades

a) olulisuse nivood 0,05;
b) olulisuse nivood 0,01.

VASTUS lk 670.

7.11. Paljudes supermarketite kettides on detailselt ära määratud, millistele riiulitele kaup müügisaalis paigutada. Kuna ostja silmade kõrgusel olevatelt riiulitelt võetakse rohkem kaupa kui madalamal või kõrgemal asuvatelt riiulitelt, paigutatakse sinna suuremat kasumit andvad ja kiirema ringlusega kaubad. Traditsiooniliselt on selleks kõrguseks võetud ligikaudu 1,5 m põrandapinnast, arvestades, et keskmise naissoost ostja pikkus on 163 cm. Viimasel ajal on üha rohkem hakanud poodides käima mehed, kelle keskmine pikkus on aga 178 cm. Kas see tendents on mõjutanud ka kaupade paigutust supermarketites? Selleks viidi läbi vastav uuring. 50 supermarketis mõõdeti ära, kui kõrgele põrandast on paigutatud kõige ostetavamad kaubad. Valimi keskmiseks saadi 1,57 m ja valimi standardhälve oli 0,13 m. Kas uuring tõestab

seda, et enammüüdavad kaubad on nüüd paigutatud supermarketites kõrgemale kui 1,5 meetrit põrandapinnast? VASTUS lk 670.

7.12. Kogumi keskväärtuse testimisel on nullhüpotees $\mu = 100$ ja sisukas hüpotees $\mu \neq 100$. Leida olulisuse nivool 0,05, kumb hüpotees tuleb vastu võtta, kui

a) valimi maht on 100, keskmine 110 ning standardhälve 70;
b) valimi maht on 200, keskmine 110 ning standardhälve 70.

Milline on järeldus? VASTUS lk 671.

7.13. Kogumi keskväärtuse testimisel on nullhüpotees $\mu = 100$ ja sisukas hüpotees $\mu \neq 100$. Leida olulisuse nivool 0,05, kumb hüpotees tuleb vastu võtta, kui

a) valimi maht on 150, keskmine 110 ning standardhälve 60;
b) valimi maht on 150, keskmine 110 ning standardhälve 90.

Milline on järeldus? VASTUS lk 671.

Olulisuse tõenäosuse kasutamine

7.14. Kui kasutatakse olulisuse nivood $\alpha = 0{,}05$, siis milliste olulisuse tõenäosuse väärtuste korral on nullhüpotees ümber lükatud?

a) $p = 0{,}07$,
b) $p = 0{,}1$,
c) $p = 0{,}01$,
d) $p = 0{,}034$,
e) $p = 0{,}25$,
f) $p = 1{,}5 \cdot 10^{-5}$.

VASTUS lk 671.

7.15. Milliste olulisuse nivoo α ja olulisuse tõenäosuse p paaride korral on nullhüpotees ümber lükatud?

a) $\alpha = 0{,}05$, $p = 0{,}1$;
b) $\alpha = 0{,}1$, $p = 0{,}05$;
c) $\alpha = 0{,}01$, $p = 0{,}001$;
d) $\alpha = 0{,}1$, $p = 0{,}45$;
e) $\alpha = 0{,}025$, $p = 0{,}05$;
f) $\alpha = 0{,}05$, $p = 0{,}025$.

VASTUS lk 671.

t-test, sõltumatud valimid

7.16. Ajakirjas The American Economic Review 2001. aastal ilmunud artiklis võrreldi riigiettevõtete ja eraettevõtete efektiivsust (Dewenter ja Malatesta, 2001). Kasutati 1369 ettevõtte andmeid aastatest 1975, 1985 ja 1995. Nendest 147 olid riigiettevõtted. Võrreldi ettevõtete tootlus- ja likviidsussuhtarve. Tabelis on toodud vastavate t-testide tulemused.

Suhtarv	Keskmine		t-statistik
	Riigi-ettevõtted	Era-ettevõtted	
Kasumi osatähtsus netokäibes	0,013	0,027	−1,530
Varade tootlus	0,012	0,030	−3,084
Omakapitali tootlus	−0,007	0,089	−2,093
Lühiajaliste kohustuste kattekordaja	0,725	0,709	0,912

Milliste suhtarvude korral on erinevus tõestatud nivool 0,05 ja milliste korral nivool 0,01? VASTUS lk 671.

F-test

7.17. Ajakiri Journal of Occupational and Organizational Psychology viis 1992. aasta detsembris läbi uuringu, kas inimese vaimne tervis ja tööga kindlustatus on omavahel seotud. Üks valim moodustati töötute hulgast, teine hõivatute (töötavate inimeste) hulgast. Uuringuks kasutati GHO (*General Health Questionnaire*) testi. Lisaks kahe valimi keskväärtuse võrdlemisele võrreldi ka testi tulemuste varieerumist erinevates valimites. Näiteks 142 hõivatud mehe korral oli testi tulemuste standardhälve 3,62, aga 49 töötu mehe korral 5,10. Kas vaimse tervise näitaja varieerumine on nendes gruppides oluliselt erinev? VASTUS lk 671.

t-test sõltumatute valimite korral koos eelneva F-testiga

7.18. Indias on üle 150 miljoni internetikasutaja. Ajakirjas Journal of Marketing & Communication 2013. aastal ilmunud artiklis analüüsiti 15–35-aastaste internetikasutajate harjumusi (Dahiya, 2013). Selleks viidi läbi küsitlus, millele vastas 58 noormeest ja 42 neidu. Üheks küsimuseks oli, kui palju aega kulutatakse nädalas *online*-mängude mängimiseks. Noormeeste keskmine aeg oli 4,02 tundi nädalas standardhälbega 5,369 tundi ja neidudel 2,43 tundi nädalas standardhälbega 3,569 tundi. Testida olulisuse nivool 0,05, kas neiud mängivad *online*-mänge vähem. Eelnevalt viia läbi dispersioonide testimine F-testiga, määramaks, kumba sõltumatute valimite t-testi varianti tuleb kasutada. VASTUS lk 671.

Osakaalu testimine, üks valim

7.19. Ettevõte on seadnud klienditeeninduses eesmärgiks, et vähemalt 90% klientidest oleks teenindusega rahul. 13 000 kliendi seast valiti juhuslikult välja 725, keda intervjueeriti. 113 neist väitsid, et nad pole teenindusega rahul. Kas ettevõtte juhtkond peaks klientide rahulolu suurendamiseks midagi ette võtma? VASTUS lk 671.

7.20. Ajavahemikul 15.12.2000–07.01.2001 viis AS Emor läbi küsitluse „Hoiakud web'i keskkonnas tuludeklaratsiooni täitmise suhtes". Uuringu käigus küsitleti kokku 384 inimest. Nendest, kes täitsid 1999. aasta kohta tuludeklaratsiooni, tegi 164 inimest seda paberblanketil ja 82 inimest e-maksuametis. (Emor, 2001) Kontrollida olulisuse nivool

0,01 hüpoteesi, et 1999. aasta tuludeklaratsiooni täideti rohkem paberil kui e-maksuametis. VASTUS lk 671.

Osakaalude võrdlemine, kaks valimit

7.21. 2013. aasta Eesti sotsiaaluuringus oli küsitletute hulgas 6865 tööealist isikut, nendest 3551 meest ja 3314 naist (*Eesti sotsiaaluuring* 2013). Meeste hulgas oli töötuid 434 ja naiste hulgas 288. Kontrollida hüpoteesi, et meeste hulgas on töötute osakaal suurem kui naiste hulgas. Kasutada olulisuse nivood 0,01. VASTUS lk 671.

7.22. Enne 2015. aasta riigikogu valimisi viidi mitmel korral läbi uuringuid erakondade toetuse kohta. TNS Emor viis ühe küsitluse läbi 11.–18. veebruaril ja järgmine oli 23.–26. veebruaril. Mõlemal korral küsitleti ligikaudu 1000 inimest. Esimesel küsitlusel toetas Reformierakonda 23% küsitletutest, teisel korral 26% küsitletutest[1]. Kas olulisuse nivool 0,05 võib väita, et Reformierakonna toetus suurenes? VASTUS lk 671.

7.23. Lähtudes ülesandes 7.22 toodud andmetest, leida, kui suur oleks pidanud Reformierakonna toetusprotsent olema teises küsitluses osalejate hulgas, et olulisuse nivool 0,05 oleks tõestatud sisukas hüpotees: erakonna toetus suurenes. VASTUS lk 672.

Märgitest

7.24. Kas Rootsi tööjõuturul esineb vanuselist diskrimineerimist? Selle uurimiseks korraldasid ajakirjas Applied Economics Letters ilmunud artikli autorid eksperimendi (Ahmed, Andersson ja Hammarstedt, 2012). Eksperimendi käigus saadeti avaldused kandideerimiseks vabadele töökohtadele. 203 ettevõtet vajasid müügiagenti ja 263 ettevõtet restorani abitööjõudu. Need kaks valdkonda valiti välja seepärast, et seal on suhteliselt suur tööjõupuudus. Igasse ettevõttesse saadeti kaks avaldust fiktiivsetelt kandideerijatelt. Mõlemal juhul oli tegemist mehega, kuid vanusevahe oli neil 15 aastat. Noorema vanuseks märgiti 31 aastat ja vanema vanuseks 46 aastat. Muud parameetrid olid fiktiivsetel kandideerijatel samad: tööstaaž vastaval erialal 10 aastat, hea suhtlemisoskus, perfektne rootsi ja inglise keele valdamine, abielus, lapsi pole.

Tabelis on kandideerimisavaldustele vastanud ettevõtete arvud. Tulemused on esitatud kogu valimi kohta ja eraldi kummagi töökoha, täis- või osalise tööaja ning alalise või ajutise töökoha kohta. Testida olulisuse nivool 0,01, kas tööandjad eelistavad nooremat kandideerijat. Testimine viia läbi nii kogu valimi jaoks kui ka valdkondade kaupa eraldi. VASTUS lk 672.

[1] Allikas: TNS Emor, `http://www.emor.ee`

	Ei vastanud kummalegi	Vastasid mõlemale	Ainult nooremale	Ainult vanemale
Kogu valim	419	8	34	5
Müügiagent	188	1	12	2
Abitööline restoranis	231	7	22	3
Täistööaeg	284	8	25	4
Osaline tööaeg	135	0	9	1
Alaline töökoht	265	3	15	3
Ajutine töökoht	154	5	19	2

7.25. Kümme eksperti maitsesid veinimarke Chardonnay ja Cabernet Sauvignon. Tabelis on iga eksperdi hinnang kummalegi veinile. Hinnang anti 20-pallises skaalas, suurem number tähendas paremat maitset. Testida olulisuse nivool 5%, kas hinnang veinidele on erinev. VASTUS lk 672.

Ekspert	Chardonnay	Cabernet Sauvignon
1	18	17
2	19	20
3	20	17
4	18	19
5	20	20
6	20	19
7	17	19
8	17	20
9	18	16
10	20	19

χ^2-test

7.26. 2002. aastal ilmus ajakirjas Technovation uurimus Kanada biotehnoloogia ettevõtete teadus- ja arendustegevuse seosest ettevõtte edukusega (Hall ja Bagchi-Sen, 2002). Üheks edukuse näitajaks oli võetud müügitulu keskmine kasv aastatel 1993–1998. Tabelis on toodud 60 valimis olnud ettevõtte jagunemine kahe tunnuse järgi: kas ettevõttes tegeleti tootearendusega ja kas müügitulu kasvas aastatel 1993–1998?

	Müügitulu kasv	
Tootearendus	Ei	Jah
Ei	6	11
Jah	5	38

1. Püstitada nullhüpotees ja sisukas hüpotees.
2. Leida nullhüpoteesile vastavad oodatavad sagedused.
3. Leida χ^2-statistik.
4. Kontrollida olulisuse nivool 0,05, kas müügitulu kasv on arendustegevusega seotud.

Vastus lk 672.

7.27. Dispersioonanalüüsi korral leida üldine vabadusastmete arv ning vabadusastmete arv, mis vastab rühmadevahelisele keskruudule ja rühmasisesele keskruudule, kui *ANOVA*

a) vaatluste koguarv on 120 ning valimite arv 3;

b) vaatluste koguarv on 150 ning valimite arv 5.

Vastus lk 672.

7.28. Leida ANOVA tabelis küsimärgiga märgitud puuduvad suurused. Vastus lk 672.

Hajuvuse allikas	*SS*	*df*	*MS*	*F*
Rühmadevaheline	385,8	2	?	?
Rühmasisene	6847,3	?	?	
Üldine	?	29		

7.29. Väike- ja suurettevõtete vahel esinevad mitmed suurusest tingitud erinevused. Näiteks kapitali struktuur, ekspordivõimekus, pankrotioht, krediidi saamise võimalus sõltuvad ettevõtte suurusest. Ettevõtte suurust võib aga hinnata erinevate näitajate alusel: töötajate arv, käive, varade suurus. Euroopa Komisjoni soovituse 96/280/CE järgi tuleks kasutada töötajate arvu, kuid mitmed uuringud näitavad, et majandusanalüüsis on parem kasutada ettevõtete grupeerimist käibe järgi.

Ajakirjas Global Finance Journal 2005. aastal ilmunud artiklis analüüsiti, kas ettevõtete iseloomustamiseks kasutatavad rahandussuhtarvud sõltuvad ettevõtte suurusest, kui suuruse näitajaks võtta käive (Cinca, Molinero ja Larraz, 2005). Kokku analüüsiti 15 rahandussuhtarvu 11 riigist ja 19 tegevusvaldkonnast pärinevatel ettevõtetel perioodil 1986–1999. Ettevõtted jagati kolme gruppi: väikeettevõtted käibega kuni 7 mln eurot, keskmise suurusega ettevõtted käibega 7–40 mln eurot ning suurettevõtted käibega üle 40 mln euro aastas. Seda, kas erinevatesse suurusgruppidesse kuuluvatel ettevõtetel on rahandussuhtarvud erinevad, testiti dispersioonanalüüsiga. Tabelis on dispersioonanalüüsi tulemused kahe suhtarvu korral kuues riigis: F-statistik ja sulgudes sellele vastav olulisuse tõenäosus p.

	Belgia	Taani	Soome	Prantsusmaa	Saksamaa	Holland
Netokasumi osatähtsus netokäibes	10,656 (0,000)	25,210 (0,000)	3,367 (0,038)	6,888 (0,001)	19,355 (0,000)	11,007 (0,000)
Omakapitali tootlus puhaskasumi järgi	7,519 (0,001)	1,129 (0,324)	5,419 (0,006)	0,917 (0,400)	11,205 (0,000)	4,011 (0,019)

1. Millistes riikides on netokasumi osatähtsus netokäibes erineva suurusega ettevõtetes erinev?
2. Millistes riikides on omakapitali tootlus puhaskasumi järgi erineva suurusega ettevõtetes erinev?

Otsustamiseks kasutada olulisuse nivood 0,01. VASTUS lk 672.

Erinevad testid

7.30. Perefirma on ettevõttevorm, kus enamik firma kontrollpakist kuulub ühele perekonnale ja firma tegevusse on samaaegselt aktiivselt kaasatud kaks või enam pereliiget. 2004. aastal ajakirjas Family Business Review ilmunud artiklis „Comparison of Family and Nonfamily Business: Financial Logic and Personal Preferences“ (Gallo, Tàpies ja Cappuyns, 2004) võrreldi perefirmasid ja mitte-perefirmasid. Uuringu küsimustik saadeti 4200 Hispaania ettevõttele, mille käive oli üle 21,6 mln euro aastas ja kus töötas rohkem kui 150 töötajat. Tagasi saadi 305 korrektselt täidetud ankeeti. Nende hulgas oli 101 perefirmat ja 204 mitte-perefirmat.

1. Oma tegevusalal oli 10 esimese seas 75 perefirmat 101-st ja 143 mitte-perefirmat 204-st. Kas võib väita, et perefirmad on sagedamini esikümne seas kui mitte-perefirmad?
2. Perefirmadel oli keskmine käive ühe töötaja kohta aastas 0,21 mln eurot standardhälbega 0,19 mln eurot. Mitte-perefirmadel oli keskmine käive töötaja kohta 0,26 mln eurot standardhälbega 0,26 mln eurot. Kas võib väita, et perefirmadel on keskmine käive ühe töötaja kohta väiksem kui mitte-perefirmadel?

VASTUS lk 672.

ÜL07Hüpoteesid

Järgmiste ülesannete andmed on failis ÜL07Hüpoteesid

Kõikide ülesannete juures tuleb püstitada nullhüpotees ja sisukas hüpotees, viia läbi vastavad arvutused ning võtta vastu otsus.

z-test, keskväärtuse testimine

A.7.1. Ülesandes A.5.5 tuli leida, kui suur osa Horvaatias asuva klaasitehase Vetropack Straza D.D. tootmisliinilt L621 tulevatest pudelitest on praak. Klaasitehases on mitu tootmisliini. Liinil L621 on automaatne tootmisprotsessi monitoorimise süsteem WISEPPC, liinil L635 sellist süsteemi pole (Kovačec, Pilipović ja Štefanić, 2010). Pudelite nominaalmass on 190 g. Tabelis on kummaltki liinilt võetud 80 pudeli mass grammides. Mõlema liini korral testida, kas keskmine pudelite mass vastab nominaalsele või mitte. VASTUS lk 673.

A.7.2. Pesuvahendite tootjal on toodete hulgas 1500 grammine pesupulbripakk. Pakkide täitmise liini on vaja perioodiliselt kontrollida, et näha, kas tegelik pakkide täitmine vastab nominaalsele. Selleks võeti

liinilt juhuslikult 50 pakki ja kontrolliti täppiskaaludega üle. Tulemused on toodud tabelis. Kasutades otsuse tegemiseks usaldatavuse nivood 0,05, otsustada, kas pakendamisliin vajab reguleerimist, kui tootjat huvitab,

a) et kõikumine mõlemale poole nominaalsest oleks kontrolli all;
b) ega pakkides liiga palju pesupulbrit pole.

VASTUS lk 673.

A.7.3. Töötajad, kes viibivad iga päev suure müratasemega (üle 85 dB) keskkonnas, peavad kaitsma oma kuulmist ja selleks kasutatakse müra-summutavaid kõrvaklappe. Firma Peltori kõrvaklappide X1A kasutusjuhendis on kirjas, et 250 Hz müra korral on mürasummutuse keskväärtus 15,4 dB[2]. Töökaitseinspektsioon soovib kontrollida, kas see vastab tõele. Selleks testitakse viit paari juhuslikult väljavalitud kõrvaklappe. Testi tulemused on toodud tabelis. Milline on järeldus? VASTUS lk 673.

t-test, keskväärtuse testimine

A.7.4. Analüütik Gerald Appel pani New Yorgi aktsiabörsi (NYSE) uurides tähele, et kui nädala keskmise sulgemishinna ja viimase 10 nädala libiseva keskmise erinevus on väga suur ($-4{,}0$ punkti või rohkem), siis on varsti oodata hindade tõusmist. Analüütik soovitas aktsiaid osta kolm nädalat peale suurimat negatiivset erinevust, hiljem hakkavad hinnad tõusma. Analüütiku väite kontrollimiseks valis ajakiri Barron juhuslikult välja kaheksa kuupäeva, millal oli ostusignaal, ning registreeris indeksi tulumäära 26 nädalat peale ostusignaali. Kui analüütiku sõnu võib uskuda, siis peaks keskmine tulumäär olema positiivne. Kontrollida analüütiku tähelepanekut olulisuse nivool 0,05 ja 0,01. VASTUS lk 673.

A.7.5. Ajakirjas Journal of Applied Econometrics 1997. aastal ilmunud artiklis analüüsiti, kuidas on omavahel seotud ettevõtete arenduskulud ning taotletud patentide arv (Cincera, 1997). Valimis oli 181 ettevõtet USA-st, Euroopast, Jaapanist ja mujalt. Ettevõtted jagunesid 15 erineva tööstusharu vahel. Tabelis on arenduskulud valimisse kuulunud lennuki-, masina- ja keemiatööstusettevõtetes 1991. aastal, mln USD. Kasutades olulisuse nivood 0,1, testida, kas arenduskulud aastas on erinevad

a) lennukitööstuses ja masinatööstuses;
b) lennukitööstuses ja keemiatööstuses.

Eeldada, et kogumite dispersioonid on võrdsed. VASTUS lk 673.

t-test, sõltumatud valimid, võrdsed dispersioonid

A.7.6. Ülesandes 7.30 viidatud artiklis võrreldi ka perefirmade ja mitte-perefirmade rahandussuhtarvusid. Valimvaatluse tulemusel saadud keskväärtused ja standardhälbed on toodud tabelis. Perefirmasid oli valimis 101 ning mitte-perefirmasid 204. Leida, millised suhtarvud

t-test, sõltumatud valimid, erinevad dispersioonid

[2] Allikas: `http://solutions.3m.com/`, 3M Peltor X Series Ear Muffs Brochure.

on perefirmadel ja mitte-perefirmadel erinevad. Eeldada, et kogumite dispersioonid on erinevad ning otsustamiseks kasutada olulisuse nivood 0,05. VASTUS lk 673.

A.7.7. Ajakirjas Journal of Medical Systems ilmunud uuringus võrreldi kasumit taotlevate ja kasumit mittetaotlevate haiglate efektiivsust (Lee, S.-B. Yang ja Choi, 2009). Kasutati USA Florida osariigi haiglate andmeid aastatest 2001–2004. Võrreldi haiglate poolt osutatavate teenuste arvu, töötajate arvu, kulusid, haigusjuhtude ning ambulatoorsete visiitide arvu ja residentuuris õppivate praktiseerivate üliõpilaste ehk residentide keskmist arvu. Testida olulisuse nivool 0,05, kas need näitajad on kasumit taotlevates ja mittetaotlevates haiglates erinevad. Eeldada, et kogumite dispersioonid on erinevad. VASTUS lk 674.

A.7.8. Eesti 2012. aasta tööjõu-uuringus osalejate hulgast on välja valitud need, kellel töötuse kestus oli pikem kui üks aasta ja elukoht Põhja-Eesti või Kirde-Eesti. Tabelis on nende isikute töötuse kestus kuudes (*Eesti tööjõu-uuring* 2012). Kontrollida hüpoteesi, et Kirde-Eesti töötutel on töötuse kestus keskmiselt suurem kui Põhja-Eesti töötutel. Eeldada, et kogumite dispersioonid on erinevad. VASTUS lk 674.

t-test, sõltuvad valimid

A.7.9. Allen F. Jung analüüsis oma artiklis „Price Variations Among Automobile Dealers in Chicago, Illinois“, kas informeeritumad autoostjad saavad müüjalt soodsama hinnapakkumise (Jung, 1959). Ta viis läbi eksperimendi, kus kaks ostjat pöördusid Chicagos 30 erineva autosalongi poole sooviga osta üht ja sama Chevrolet' mudelit. Ostja A käitus isikuna, kes just oli saanud juhiload ja oli veidi murelik oma esimest autot ostes. Ostja B käitus kui isik, kes tunneb autosid, on hästi kursis autoostu detailidega ja teab, mida tahab. Tabelis on toodud müüjate poolt kummalegi ostjale tehtud hinnapakkumised. Kontrollida olulisuse nivool 0,05, kas ostjale B pakuti keskmiselt madalamat hinda. Testimine viia läbi kahel meetodil.

1. Leida hinnapakkumiste erinevused, erinevuste keskmine ja standardhälve ning seejärel t-statistik valemitest (7.37).
2. Programmis Excel kasutada vahendit *t-test: Paired Two Sample for Means* andmeanalüüsi komplektist *Data Analysis*.

VASTUS lk 674.

A.7.10. Ajakiri Hooaeg on üks reklaamikanal, mida Kaubamaja kasutab. Kas ajakirjas avaldatud reklaam suurendab ka toodete läbimüüki? Tabelis on 2005. aasta kevadises Hooaja numbris reklaamitud ekslusiivkategooriasse kuuluvate kosmeetikatoodete müügimahud kuu aega enne ja kuu aega peale reklaami avaldamist. Testida olulisuse nivool

0,05, kas reklaam avaldas müügimahule positiivset mõju. VASTUS lk 674.

A.7.11. Tabelis on erinevate isikute vaba aja veetmiseks tehtud kulutused aastas. Valimis on Eesti leibkonna eelarve uuringus osalenud isikud üheliikmelistest leibkondadest, s.t üksi elavad mehed ja naised (*Leibkonna eelarve uuring* 2012). Tunnuse „Sugu“ väärtus meestel on 1 ja naistel 2. Olulisuse nivool 5% kontrollida hüpoteesi, et meestel varieeruvad kulutused vabale ajale rohkem kui naistel. VASTUS lk 674.

F-test

A.7.12. Näites 7.14 testiti, kas OMXT indeksi tulumääral esines kuuvahetuse efekt ajavahemikul 3.07.2006–30.12.2009. Väärtpaberiturgudel on tähele pandud ka nädalasisese kalendriefekti eksisteerimist, mida nimetatakse nädalapäeva efektiks.

t-test sõltumatute valimite korral koos eelneva F-testiga

Tabelis on OMXT indeksi tulumäärad neljapäeviti ja ülejäänud nädalapäevadel (E, T, K, R) ajavahemikul 3.07.2006–30.12.2009 (Noormägi, 2010). Testida, kas neljapäeviti on OMXT indeksi tulumäär madalam kui ülejäänud nädalapäevadel. Eelnevalt teha F-testi abil kindlaks, kumba t-testi varianti kasutada: ühesuguste või erinevate dispersioonidega. VASTUS lk 674.

A.7.13. Lennufirmad teavad oma kogemusest, et mõned koha reserveerinud reisijad muudavad või tühistavad oma broneeringu enne väljalendu või lihtsalt ei ilmu oma lennule. Seepärast praktiseerivad lennufirmad sageli kohtade ülebroneerimist, mis tähendab, et reservatsioonisüsteem lubab broneerida ja müüa rohkem pileteid, kui on konkreetsel lennul kohti. Kui ülebroneerimist ei kasutataks, lahkuks lennuk tühjade kohtadega, mille on endale kinni pannud lennule mitteilmunud või viimasel hetkel broneeringu tühistanud reisijad. Lende broneeritakse üle, kuna lennufirmad soovivad saavutada paremat kohtade täituvust igal lennul.

Tabelis on Estonian Airi lendudele mitteilmumiste arv 100 reisija kohta aastal 2001, nädalate kaupa. Nende järgi määratakse ning prognoositakse ülebroneeringute hulk (Haidla, 2004). Testida, kas lennule mitteilmumiste arv 100 reisija kohta on äriklassis ja turismiklassis erinev. Eelnevalt teha F-testi abil kindlaks, kumba t-testi varianti kasutada: ühesuguste või erinevate dispersioonidega. VASTUS lk 674.

A.7.14. Suhteline vaesuspiir on 60% leibkonnaliikmete aasta ekvivalentnetosissetuleku mediaanist. Ekvivalentnetosissetulek on leibkonna sissetulek, mis on jagatud leibkonnaliikmete tarbimiskaalude summaga. Tabelis on alamvalim ($n = 500$) 2010. aastal Eesti sotsiaaluuringus osalenud isikutest. Iga isiku jaoks on kirjas, kas ta elab suhtelises vaesuses või mitte ja haridustase (*Eesti sotsiaaluuring* 2013).

Osakaalude testimine, kaks valimit

Suhteline vaesus	0	suhtelisest vaesuspiirist ülalpool;
	1	suhtelisest vaesuspiirist allpool.
Haridustase	1	I taseme haridus: alghariduseta, alghariduse-ga, põhiharidusega, baashariduseta kutsehari-dus;
	2	II taseme haridus: keskharidus, kutseõpe põ-hihariduse baasil;
	3	III taseme haridus: kutseõpe keskhariduse baasil, kõrgharidus, magister, doktor.

Kontrollida olulisuse nivool 0,01 hüpoteesi, et I taseme haridusega inimeste hulgas on rohkem suhtelisest vaesuspiirist allpool elavaid isikuid kui II ja III haridustasemega isikute hulgas. VASTUS lk 675.

Märgitest

A.7.15. 1990-ndate alguses alustasid USA-s paljud ettevõtted kulude kokkuhoiu pärast töötajate koondamist. Vanemad töötajad süüdistasid seejuures ettevõtteid vanuselises diskrimineerimises. Föderaalne seadusandlus kaitseb vanuselise diskrimineerimise eest töövõtjaid, kellel vanust üle 40 aasta.

Olgu ettevõtte A töötajate mediaanvanus 37 aastat. Ettevõte plaanib koondada 15 töötajat, kelle vanused on toodud tabelis. Et süüdistada ettevõtte juhtkonda vanuselises diskrimineerimises ja vastava seaduse rikkumises, on kohtule vaja esitada statistiline tõestusmaterjal, mille olulisus on 10%. Kas ettevõte võib väljavalitud töötajad koondada ilma kohtuprotsessi kartmata? VASTUS lk 675.

A.7.16. 2011. aastal viidi külaelanike hulgas läbi küsitlus, kuidas nad hindavad külavanema tegevust. Küsitlusele vastas 15 külaelanikku. Hinnang paluti anda viie palli skaalas, suurem pallide arv tähendas paremat hinnangut. 2013. aastal paluti samadel külaelanikel anda uus hinnang. Kas hinnang paranes? Hüpoteesi kontrollimiseks kasutada olulisuse nivood 5%. VASTUS lk 675.

A.7.17. Tabelis on ühe tonni bensiini hulgihind ajavahemikul 2.01.–28.12.2007. Testida, kas hinnatõususid esineb sagedamini kui hinnalangusi. Testimiseks kasutada kaht meetodit:

a) märgitest;
b) osakaalu testimine suure valimi korral.

Mõlemal juhul kasutada olulisuse nivood 0,05. VASTUS lk 675.

Jaotuse sobivuse χ^2-test

A.7.18. XX sajandi algul, kui ei olnud veel automaattelefonijaamu ja telefoniühenduse lõid operaatorid, esines tihti valeühendusi. Tabelis on ühe New Yorgi telefonijaama päevas loodud valeühenduste arvu jaotus (Thorndike, 1926). Kontrollida olulisuse nivool 0,05, kas see jaotus allub Poissoni jaotusele. VASTUS lk 675.

A.7.19. Kas kodulehe külastatavus on erinevatel nädalapäevadel eri-

nev? Kui ei ole, siis peaks külastuste arv erinevatel nädalapäevadel alluma ühtlasele jaotusele. Tabelis on õpiku autori kodulehe külastatavus 28 päeval 2.–29.11.2015. Nende 28 päeva hulgas esineb kõiki nädalapäevi neli korda ning külastuste arv on nädalapäevade kaupa summeeritud. Kontrollida olulisuse nivool 0,05, kas külastuste arv

a) kõigi nädalapäevade lõikes allub ühtlasele jaotusele;
b) tööpäevade lõikes allub ühtlasele jaotusele.

Vastus lk 675.

A.7.20. Sobiva ülebroneerimiste arvu leidmiseks võib lennufirma kasutada binoomjaotust nagu ülesandes 5.29. Binoomjaotuse kasutamisel eeldatakse, et iga broneeringu kinnitamine on sõltumatu ülejäänud broneeringute kinnitamisest, s.t ignoreeritakse mitmekesi lendamist. Ülesandes A.7.13 olid toodud andmed Estonian Airi lendudele mitteilmumise kohta aastal 2001, nädalate kaupa. Mitteilmumiste arv on 100 reisija kohta eraldi äriklassis ja turismiklassis. Mõlema reisijaklassi korral kontrollida, kas mitteilmumiste arv allub binoomjaotusele $B(p, n)$, kus $n = 100$ ja üksiku sündmuse tõenäosus p on 52 nädala keskmine (kummaski klassis eraldi). Kasutada olulisuse nivood 0,05. Näpunäide: empiiriliste sageduste leidmiseks ümardada mitteilmumiste arv täisarvuni ja seejärel loendada saadud täisarvude esinemissagedused. Vastus lk 676.

A.7.21. Testida, kas Ford Motor Company aktsia tulumäär päevas[3] allub normaaljaotusele. Kasutada on kolme kuu andmed (märts kuni mai 2016). Vastus lk 676.

χ^2-test ja kahe tunnuse vaheline seos

A.7.22. 2001. aastal tegid Audentese Kõrgema Ärikooli üliõpilased sotsioloogia õppeaines uurimistöö „Inimeste arvamus progresseeruva tulumaksu kohta“ (Michelson, Niils ja Kala, 2001). Uurimistöö jaoks koostati küsitlus, millele vastas 90 inimest. Kokku esitati 9 küsimust, mis kõik olid seotud progresseeruva tulumaksuga ja millele tuli vastata kas „jah“ või „ei“. Lisaks paluti vastajatel märkida sugu, vanus ning hinnang oma majanduslikule olukorrale: „väga halb“, „ pigem halb“, „pigem hea“, „väga hea“.

Tabelis on toodud vastuste jaotus küsimusele „Kas pooldate progresseeruva tulumaksu kehtestamist?“ Testida, kas vastus sellele küsimusele sõltub vastaja majanduslikust olukorrast. Vastus lk 676.

A.7.23. Neli protsentiili järkudega 0,2, 0,4, 0,6 ja 0,8 jagavad variatsioonrea viieks võrdseks osaks, kus igas osas on 20% variatsioonrea liikmetest. Sissetulekute jaotuse uurimisel nimetatakse neid vahemikke sissetuleku kvintiilideks. Näiteks esimesse kvintiili kuuluvad isikud,

[3] Allikas: Yahoo Finance, `http://finance.yahoo.com`

kelle sissetulek on madalam kui protsentiil 0,2. Teise kvintiili kuuluvad isikud, kelle sissetulek on protsentiilide 0,2 ja 0,4 vahel.

Risttabelis on andmed 2013. aasta Eesti sotsiaaluuringus osalenud 15 040 isiku kohta (*Eesti sotsiaaluuring* 2013). Isikud on rühmitatud kahe tunnuse järgi: alla 25-aastaste laste arv leibkonnas ning ekvivalentsissetuleku kvintiil. Ekvivalentsissetulek on leibkonna sissetulek, mis on jagatud leibkonnaliikmete tarbimiskaalude summaga. Tabelis on toodud isikute arv, kellel need tunnused omasid vastavaid väärtusi. Laste arv „4“ tähendab „4 või rohkem“ last. Kontrollida hüpoteesi, kas kuulumine sissetuleku kvintiili sõltub laste arvust leibkonnas. VASTUS lk 676.

ANOVA

A.7.24. Reklaamiagentuur soovis analüüsida, millises kanalis (ajaleht, raadio või televisoon) avaldatud reklaam suurendab kõige rohkem firma käivet. Eksperimendi jaoks valiti välja 12 enam-vähem ühesuurust linna. Neljas linnas avaldati ajalehereklaam, neljas raadioreklaam ja neljas TV-reklaam. Kuu aega hiljem koguti andmed vastava toote müüginumbrite kohta kõigis linnades ja leiti, mitu eurot suurenes vastava toote käive ühe reklaamile kulutatud euro kohta. Kas reklaami mõju käibe suurenemisele sõltub reklaamikanalist? VASTUS lk 676.

A.7.25. Tänapäeval on väga levinud kiirtoidud hamburger, *hot dog* jms. Samal ajal tunnevad inimesed muret oma tervise pärast ja jälgivad toiduainete kalorisisaldust. Ajakiri Consumer Reports viis 1986. aastal läbi uurimuse, kus analüüsiti 54 erinevat sorti *hot dog*'i kalorisisaldust. *Hot dog*'id jagati kolme gruppi: sealihaga, loomalihaga ja kanalihaga. Tulemused on toodud tabelis. (Moore ja McCabe, 1989)

1. Hinnata, kas liha valik mõjutab kalorisisaldust.
2. Kui mõjutab, siis kas kalorisisaldus on oluliselt erinev kõigis *hot dog*'ides või mitte?

VASTUS lk 676.

Erinevad testid

A.7.26. USA Postiteenuse Columbia lennuposti keskuses viidi läbi uuring kirjade tähtaegse edastamise kohta. Kolme kuu jooksul toimetati edasi 3780 kirja, nendest 665 jõudsid kohale ettenähtud tähtajast hiljem (Franchetti, 2015). Tabelis on edasitoimetatud kirjad rühmitatud sihtkoha geograafilise piirkonna järgi, eraldi tähtaegselt kohale jõudnud ja hilinenud. Kontrollida olulisuse nivool 0,05, kas kirjade hilinemine sõltub sihtkohast. VASTUS lk 676.

A.7.27. The Wall Street Journal küsitleb kaks korda aastas majandusanalüütikuid ja muuhulgas palub neil prognoosida keskmist intressimäära, inflatsioonimäära, SKP kasvu ja muid majandussuurusi. Tabelis on toodud erinevate analüütikute prognoosid USA SKP aastase kasvumäära kohta. 2000. aasta prognoos oli tehtud 2000. aasta juulis ja 2001. aasta prognoos 2000. aasta algul (The Wall Street Journal,

2.01.2001). Kontrollida olulisuse nivool 0,01, kas analüütikute prognoos 2000. aastaks oli optimistlikum kui 2001. aastaks. VASTUS lk 676.

A.7.28. Kas sporditarvete poe külastatavus sõltub ilmast? Selle hüpoteesi kontrollimiseks registreeriti 1997. aasta maikuus ühes sporditarvete poes tehtud ostude arv päevas ning see, kas ilm oli vihmane (0) või ilus (1). Kontrollida olulisuse nivool 0,05, kas ostude arv sõltub ilmast. VASTUS lk 677.

A.7.29. Kolmkümmend ajakirja jaotati kolme lugejarühma haridustaseme järgi. 1. rühma kuuluvate ajakirjade lugejate haridustase oli kõige kõrgem, 3. rühma ajakirjade lugejaskonnal oli haridustase kõige madalam. Seejärel võeti igast rühmast juhuslikult kolm ajakirja (tabel 1). Igast väljavalitud ajakirjast valiti juhuslikult kuus reklaami, kokku 54 reklaami. Kõikide reklaamide korral pandi kirja kaks suurust: sõnade arv reklaamis ja lausete arv reklaamis. Tulemused on toodud tabelites 2 ja 3. Teha kindlaks, kas

a) sõnade arv reklaamis on ajakirjades erinev;
b) lausete arv reklaamis on ajakirjades erinev.

Millise järelduse võib teha? VASTUS lk 677.

A.7.30. Kas veebilehtedel esinevate reklaamibännerite kujundamisel arvestatakse sihtrühma kultuuriliste iseärasustega? Ajakirjas Journal of Targeting, Measurement and Analysis for Marketing 2010. aastal ilmunud artiklis võrreldi Hiina, Jaapani, Korea ja USA veebilehtedel avaldatud bännereid (Jin, 2010). Iga riigi esikümne seas olevatelt veebilehtedelt valiti juhuslikult välja 305 bännerit, kokku 1220. Bännerite iseloomustamiseks kasutati mitmeid tunnuseid, nende seas sõnade arvu bänneris, mis oli kodeeritud järgmiselt:

1 — kuni 8;
2 — 9 kuni 13;
3 — 14 ja rohkem.

Kas reklaamibänneris esinevate sõnade arv on erinevate riikide veebilehtedel erinev? VASTUS lk 677.

A.7.31. Alates 2011. aastast vaatleb TNS Emor Euroopa pealinnades toidukaupade hindu. Hinnavaatlust viiakse läbi neli korda aastas ja igas tooterühmas fikseeritakse kõige suurema müügipinnaga toote hind, võrdsete pindade korral valitakse odavam. Tabelis on toodud 2012. aasta detsembri hinnavaatluse tulemused Tallinnas ja Riias[4]. Kontrollida, kas Tallinnas on hinnad keskmiselt odavamad kui Riias. VASTUS lk 677.

A.7.32. Aastatel 2004–2008 oli võimalus paigutada oma pensionifondi

[4] Allikas: TNS Emor, `http://www.emor.ee/`

osakuid erinevatesse LHV fondidesse. Tabelis on fondide LHV Dünaamilised Võlakirjad (LHV DV) ja LHV Kvaliteetsed Võlakirjad (LHV KV) keskmine tootlus ja tootluse dispersioon aastatel 2004–2008. Vastavate fondide tutvustuses oli kirjas, et fondi LHV Kvaliteetsed Võlakirjad oodatav tootlus on kõigist LHV II samba pensionifondidest kõige madalam, kuid samuti ka risk, et fondi osakute väärtus keskmises või pikas perspektiivis väheneb.

Võrrelda nende kahe pensionifondi riski aastatel 2004 kuni 2008. Riski iseloomustamiseks kasutada tootluse dispersiooni. Olulisuse nivool 1% kontrollida kõigi aastate jaoks hüpoteesi, kas LHV Dünaamilised Võlakirjad tootluse dispersioon on suurem kui fondil LHV Kvaliteetsed Võlakirjad. VASTUS lk 677.

A.7.33. Kui hästi oskavad omavalitsused kulude planeerimisel arvestada majanduskeskkonna muutusi? Tabelis on toodud 15 juhuslikult valitud omavalitsuse planeeritud kulud vabatahtlike omavalitsuslike ülesannete täitmiseks (eurot elaniku kohta). Kulude planeerimine toimub eelneval aastal, s.t 2008. aasta kulud planeeriti aastal 2007 ja 2009. aasta kulud aastal 2008.

1. Aastal 2007 jätkus kaua aega kestnud majanduskasv. Kontrollida hüpoteesi, et aastaks 2008 planeeritud kulud olid keskmiselt suuremad kui aastaks 2007 planeeritud kulud.
2. Aastal 2008 algas majanduskriis. Kas omavalitsused arvestasid majanduskriisi võimalike mõjudega? Kontrollida hüpoteesi, et aastaks 2009 olid planeeritud kulud keskmiselt väiksemad kui aastaks 2008 planeeritud kulud.

VASTUS lk 678.

A.7.34. Eesti Statistikaameti poolt läbiviidavas leibkonna eelarve uuringus on üheks küsimuseks „Kas Teie leibkonnal või leibkonnaliikmetel on hoiuseid, kogumiskindlustusi, väärtpabereid või muid sääste?“. Tabelis on 2010. ja 2012. aasta uuringus osalejate hulgast võetud alamvalimid, mõlemad mahuga 500 (*Leibkonna eelarve uuring* 2010; *Leibkonna eelarve uuring* 2012). Tunnused on järgmised:

Kood	uuringus osaleja kood uuringu andmebaasis;
Säästmine	vastus küsimusele säästude kohta:
	1 jah,
	2 ei.

Kontrollida, kas hoiuseid, kogumiskindlustusi, väärtpabereid või muid sääste omavate leibkondade osakaal oli 2012. aastal muutunud, võrreldes 2010. aastaga. VASTUS lk 678.

A.7.35. Vältimaks kontrolltöö või kirjaliku eksami vastuste mahakirjutamist, kasutatakse erinevaid ülesannete variante. Variandid peavad

olema koostatud ühesuguse raskusastmega. Tabelis on toodud 2009. aasta kevadsemestril toimunud statistika kontrolltöö tulemused (punktide arv) variantide kaupa. Kas erinevad variandid olid ühesuguse raskusastmega või mitte? VASTUS lk 678.

A.7.36. Poest piima või leiba ostes saame kohe võrrelda selles poes müügil olevate erinevate toodete hindu. Kui me otsime aga võimalikult soodsat hinda, tuleb külastada erinevaid poode. Ka internetist kaupa ostes tuleb käia erinevate pakkujate veebilehtedel. Müüja otsimisest ning erinevate pakkumiste võrdlemisest tingitud ajakulu nimetatakse tarbija otsimiskuluks ning tarbija poolt vaadatuna lisandub see toote hinnale. Müüjad, kelleni jõudmiseks peab tarbija vähem aega kulutama (otsimiskulud on väiksemad), võivad selle võrra panna oma tootele kõrgema hinna.

Ajakirjas Journal of Applied Econometrics ilmus 2013. aastal artikkel, mille autorid analüüsisid lõpptarbija otsimiskulusid kauba ostmisel internetist (Moraga-González, Sándor ja Wildenbeest, 2013). Analüüsi teostamiseks vaatlesid nad sülearvutite mälukiipide hindasid erinevates internetipoodides. Tabelis on mälu KTHZD8000A hind 39 poes. Iga poe kohta on märgitud, kas selle logo on esindatud portaalis shopper.com või pricegrabber.com (kood 1). Kas nendes poodides, mille logo esineb nimetatud portaalides, on vastava mälukiibi hind kõrgem? VASTUS lk 678.

A.7.37. Suure poe külastajate arv päevas peaks alluma normaaljaotusele. Kõrvalekaldumised normaaljaotusest võivad esineda, kui mõnedel päevadel mõjutab külastajate arvu mingi oluline tegur. Tabelis on toodud ühe poe külastajate arv päevas ajavahemikul 1.09–30.11.2004. Poes toimusid iga kahe nädala tagant müügikampaaniad. Kontrollida, kas külastajate arv päevas allub normaaljaotusele. VASTUS lk 678.

A.7.38. 9.–14.01.2015 küsitles uudisteagentuur Bloomberg News 33 majandusanalüütikut saamaks prognoose Jaapani Keskpanga tegevuse kohta lähitulevikus. Üheks küsimuseks oli, millal võtab Jaapani Keskpank kasutusele rahapoliitilised vahendid Jaapani majanduse elavdamiseks. Analüütikute prognoosid on toodud tabelis[5]. Kas võib väita, et enamiku majandusanalüütikute arvamus oli: see toimub 2015. aasta teise pooles või hiljem? VASTUS lk 679.

A.7.39. Ülesandes A.7.1 tuli testida, kas kahelt liinilt tulnud klaaspudelite mass vastab nominaalsele. Liinil L621 oli automaatne tootmisprotsessi monitoorimise süsteem, mis liinil L635 aga puudus. Kasutades samu andmeid, testida, kas liinilt L635 tulevate pudelite mass varieerub rohkem kui liinilt L621 tuleval toodangul. VASTUS lk 679.

[5] Allikas: BloombergBusiness, January 18, 2015. `http://www.bloomberg.com/`

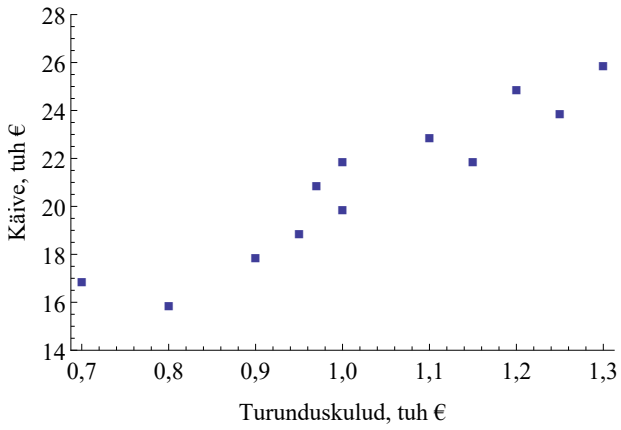
Peatükk 8

Korrelatsioonanalüüs

Sageli pakub huvi kahe tunnuse ühise käitumise uurimine. Tuleb välja selgitada, kas need tunnused on sõltuvad või sõltumatud. Kui tunnused on sõltuvad, siis tuleks mõõta nendevahelise seose tugevust. Sõltumatute tunnuste korral pole mõtet neid koos vaadata, see ei anna meile lisainformatsiooni.

8.1. Korrelatsiooni mõiste

Joonisel 8.1 on ühe ettevõtte turunduskulud ja käive erinevatel kuudel. Iga punkt vastab ühele kuule. Diagrammilt on näha, et keskmiselt turunduskulude suurenemisel käive suureneb. Öeldakse, et nende kahe suuruse vahel on olemas **korrelatsioon** (*correlation*). Sõna „korrelatsioon“ tuleb ladinakeelsest sõnast *correlãtio*, mis tähendab vastastikust seotust.



Joonis 8.1. Turunduskulude ja käibe hajumisdiagramm. Turunduskulude kasvades kasvab ka käive. Diagramm on loodud tabeli 8.2 põhjal, iga punkt vastab ühele kuule

Korrelatsioon

Korrelatsioon on juhuslike suuruste X ja Y vahel esinev statistiline seos.

Hajumis-diagramm

Korrelatsioon väljendab paarikaupa esinevat seost. Joonisel 8.1 esitatud diagrammi nimetatakse **hajumisdiagrammiks** (*scatter diagram*) ja see sobib korrelatsiooni tugevuse visuaalseks hindamiseks. Hajumisdiagrammi saamiseks peab meil olema valim, milles on mõõdetud mõlema tunnuse väärtused, nii et saame väärtuste paarid (tabel 8.1). Hajumisdiagrammil vastab igale paarile üks punkt, mille üheks koordinaadiks on tunnuse X väärtus ja teiseks koordinaadiks tunnuse Y väärtus. Mõnikord nimetatakse hajumisdiagrammi ka korrelatsiooniväljaks.

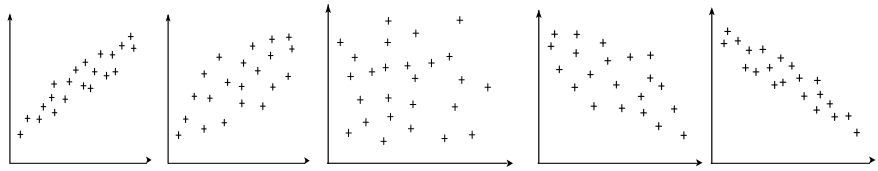
Tabel 8.1. Kahe tunnuse väärtuste paarid hajumisdiagrammi jaoks

Objekt	1	2	...	n
Tunnus X	x_1	x_2	...	x_n
Tunnus Y	y_1	y_2	...	y_n

Korrelatsiooni korral iseloomustatakse eraldi

- suunda;
- tugevust.

Neid mõlemaid võib visuaalselt hinnata hajumisdiagrammilt (vt joonis 8.2). Järgnevates alapeatükkides tutvume ka korrelatsiooni iseloomustamiseks kasutatavate näitarvudega.



Tugev positiivne korrelatsioon | Nõrk positiivne korrelatsioon | Korrelatsioon puudub | Nõrk negatiivne korrelatsioon | Tugev negatiivne korrelatsioon

Joonis 8.2. Korrelatsiooni suuna ja tugevuse visuaalne hindamine hajumisdiagrammilt

Suund võib korrelatsioonil olla kas positiivne või negatiivne:

- **positiivne korrelatsioon** — ühe suuruse kasvades teine suurus keskmiselt samuti kasvab;
- **negatiivne korrelatsioon** — ühe suuruse kasvades teine suurus keskmiselt kahaneb.

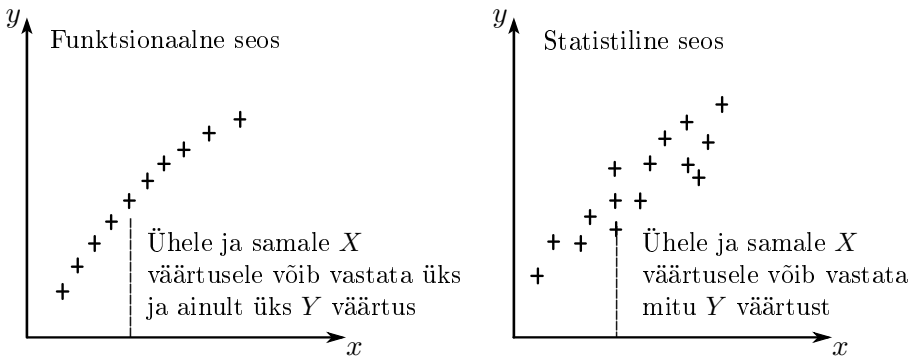
Keskmine kasvamine tähendab seda, et enamikul juhtudel suuruse X kasvades suurus Y kasvab, kuid võib olla ka selliseid arvupaare, kus X väärtuse suurenedes Y väärtus kahaneb. Sama kehtib keskmise kahanemise korral: mõningatel juhtudel võib suuruse Y väärtus kasvada.

Tabelis 8.2 on andmed, mille põhjal on konstrueeritud joonisel 8.1 esitatud hajumisdiagramm. Näeme, et jaanuaris olid turunduskulud tuhat eurot ja käive 20 tuhat eurot, veebruaris ja märtsis turunduskulud kasvasid ning käive ka kasvas. Aprillis ja mais turunduskulud vähenesid ning käive samuti vähenes. Juulis olid turunduskulud samad, mis jaanuaris, kuid käive oli suurem. Kui me võrdleme aga märtsi ja detsembrit, siis detsembris olid turunduskulud suuremad kui märtsis, aga käive väiksem. On selge, et käivet ei mõjuta ainult turunduskulud, vaid ka hulk muid tegureid, näiteks hooaeg, konkurentide müügitegevus, tarbija sissetulekute muutus jpt.

Tabel 8.2. Turunduskulud ja käive

N08Korrelatsioon T8.2

Kuu	1	2	3	4	5	6	7	8	9	10	11	12
Turunduskulud, tuh €	1,0	1,1	1,2	0,9	0,8	1,3	1,0	0,7	0,95	1,15	0,97	1,25
Käive, tuh €	20	23	25	18	16	26	22	17	19	22	21	24



Joonis 8.3. Funktsionaalne ja statistiline seos

Funktsionaalne ja korrelatiivne seos

Üldiselt võib seos kahe juhusliku suuruse vahel olla kahte tüüpi:

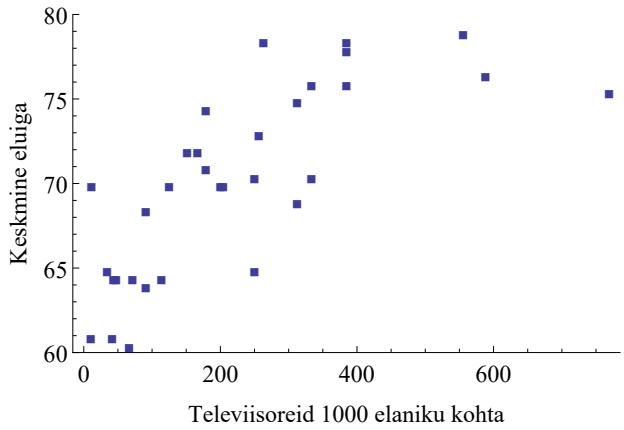
- funktsionaalse seose korral vastab argumendi X mingile väärtusele üks ja ainult üks funktsiooni Y väärtus;
- korrelatiivse ehk statistilise seose puhul võib ühe suuruse X mingile väärtusele vastata mitu erinevat teise suuruse Y väärtust (joonis 8.3).

Korrelatiivse seose korral ei saa me tunnuse X väärtust teades täpselt määrata tunnuse Y väärtust. Statistiline seos väljendab ühe juhusliku suuruse Y **keskväärtuse** sõltuvust teise juhusliku suuruse X väärtustest.

Seose olemasolu ei tähenda, et suurused on omavahel põhjuslikult seotud. **Põhjuslik seos** on seos, mille korral üks nähtus on põhjus ja teine tagajärg. Põhjus avaldab mõju tagajärjele, põhjuslik seos on alati kindla suunaga.

Tuginedes 38 riigi andmetele, on konstrueeritud hajumisdiagramm, kus on televiisorite arv 1000 elaniku kohta ja elanike keskmine eluiga. Joonisel 8.4 toodud hajumisdiagrammilt näeme, et riikides, kus televiisorite arv 1000 elaniku kohta on suurem, on keskmine eluiga kõrgem. Kas need suurused on omavahel põhjuslikult seotud? Kas keskmist eluiga mõnes riigis on võimalik tõsta, kui suurendada seal televiisorite arvu?

Korrelatiivne seos ei tähenda veel põhjusliku seose olemasolu. Põhjuslikkust saab tõestada näiteks eksperimendi abil. Siis muudetakse ühte suurust ja vaadatakse, kas teine suurus selle tagajärjel muutub.



Joonis 8.4. Televiisorite arv 1000 elaniku kohta ja keskmine eluiga 38 riigis (Rossman, 1994). Korrelatiivne seos on, põhjuslik seos puudub

Kui korrelatiivne seos on tugev, vihjab see küll põhjusliku seose võimalusele, ent ei tõesta veel selle olemasolu.

Võimalikud seosed

Võimalikud variandid, mil kahe suuruse X ja Y vahel esineb korrelatsioon:

1) suurus X mõjutab suurust Y, ühepoolne seos $X \to Y$;
2) suurus Y mõjutab suurust X, ühepoolne seos $Y \to X$;
3) suurused X ja Y mõjutavad teineteist vastastikku, kahepoolne seos $X \leftrightarrow Y$;
4) eksisteerib kolmas suurus Z, mis mõjutab nii suurust X kui ka suurust Y: $Y \leftarrow Z \to X$;

5) põhjuslik seos puudub, tegemist on näiva korrelatsiooniga (*spurious correlation*).

Korrelatsiooni uurimine üksinda ei võimalda neid võimalusi eristada. Kuna majanduses eksperimenti teha ei saa, tuleb seoste analüüsimisel lähtuda majandusteoreetilistest kaalutlustest, püüda seletada, kas ja kuidas võivad kaks suurust olla põhjuslikult seotud. Korrelatsiooni olemaolu annab meile vaid idee, et siin on mõtet otsida põhjuslikku seost.

Näiteks palkade (X) ja hindade (Y) vahel on tugev korrelatiivne seos. Mõned majandusteoreetikud usuvad, et palkade tõus põhjustab hindade tõusu (variant 1), kuna suuremad tööjõukulud nõuavad ettevõtjatelt hindade tõstmist. Seda nimetatakse nõudlusinflatsiooniks. Teised arvavad, et hindade tõusust tingitud elukalliduse tõus põhjustab palkade tõusu (variant 2). See on kuluinflatsioon. Inflatsioonispiraal vastab variandile 3. Selle korral kõigepealt nõudluse kasv hakkab kergitama hindasid, mille tõttu tekib surve palgakasvuks, mis omakorda suurendab nõudlust ja põhjustab uue hindade tõusu. Inflatsiooni monetaarse seletuse järgi põhjustab nii hindade kui ka palkade tõusu raha pakkumise (Z) suurenemine (4). Inflatsiooni kontrolli all hoidmiseks on vaja neid alternatiivseid variante eristada.

Joonisel 8.4 toodud korrelatiivsele seosele vastab neljas variant. Mõlemat tunnust, nii keskmist eluiga (X) kui ka televiisorite arvu 1000 elaniku kohta (Y) mõjutab riigi elanike elatustase (Z). Arvuliselt mõõdab seda näiteks SKP elaniku kohta.

8.2. Kovariatsioon

Seose tugevuse hindamine hajumisdiagrammi põhjal on subjektiivne. Objektiivseks hinnanguks on vaja arvkarakteristikut. Kui vaatame korrelatiivses seoses olevat kahte suurust eraldi, siis kirjeldab kummagi hajumist dispersioon (3.3):

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

Kahe tunnuse kooshajumise kirjeldamiseks sobib analoogne suurus, mida nimetatakse kovariatsiooniks (*covariance*).

Kovariatsioon

Juhuslike suuruste X ja Y vaheline **kovariatsioon**:

$$cov_{XY} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y}\ )}{n}, \qquad (8.1)$$

kus n on väärtuste x_i ja y_i paaride arv ning $\bar{x}$ ja $\bar{y}$ aritmeetilised keskmised.

Termin kovariatsioon tuleb ingliskeelsest sõnast *covariation* ehk „koos varieerumine“. Võrreldes kovariatsiooni ja dispersiooni valemeid, võime öelda, et suuruse X dispersioon on selle suuruse kovariatsioon iseendaga:

$$\sigma_X^2 = cov_{XX}. \tag{8.2}$$

Dispersioon on kovariatsiooni erijuht või kovariatsioon on dispersiooni üldistus. Sellest tulenevalt kasutatakse mõnikord kovariatsiooni tähistamiseks tähistust σ_{XY}^2.

Vaatame, kuidas on seotud punktide asend hajumisdiagrammil ja kovariatsiooni väärtus. Tabelis 8.3 on toodud kovariatsiooni arvutus kaheksa punktipaari korral. Keskmised on $\bar{x} = 4{,}5$ ja $\bar{y} = 4{,}625$. Kovariatsiooni leidmiseks liidetakse kokku tabeli viimases veerus olevad korrutised $\sum(x_i - \bar{x}) \cdot (y_i - \bar{y}) = 28{,}5$. Kovariatsioon $cov_{XY} = 28{,}5/8 \approx 3{,}56$.

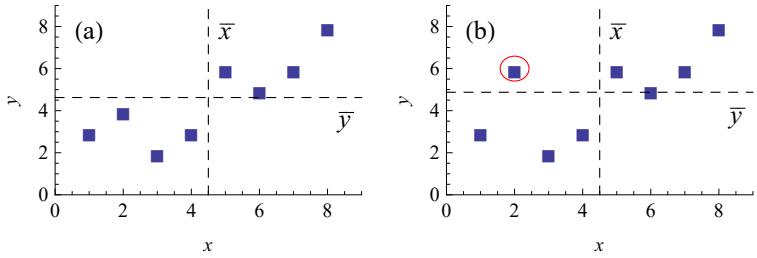
N08Korrelatsioon
T8.3,4

Tabel 8.3. Kovariatsioon $cov_{XY} = 3{,}56$, vt ka joonis 8.5 (a)

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
1	3	−3,5	−1,625	5,688
2	4	−2,5	−0,625	1,563
3	2	−1,5	−2,625	3,938
4	3	−0,5	−1,625	0,813
5	6	0,5	1,375	0,688
6	5	1,5	0,375	0,563
7	6	2,5	1,375	3,438
8	8	3,5	3,375	11,813

Tabel 8.4. Kovariatsioon $cov_{XY} = 2{,}94$, vt ka joonis 8.5 (b)

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
1	3	−3,5	−1,875	6,563
2	**6**	−2,5	**1,125**	**−2,813**
3	2	−1,5	−2,875	4,313
4	3	−0,5	−1,875	0,938
5	6	0,5	1,125	0,563
6	5	1,5	0,125	0,188
7	6	2,5	1,125	2,813
8	8	3,5	3,125	10,938



Joonis 8.5. Hajumisdiagrammid: (a) tabeli 8.3 põhjal $cov_{XY} = 3{,}56$, (b) tabeli 8.4 põhjal $cov_{XY} = 2{,}94$

Tabelis 8.4 on muutunud teise punkti y-koordinaat, mis nüüd on 6. Uus aritmeetiline keskmine $\bar{y} = 4{,}875$. Kui enne oli 2. punkti y-koordinaat väiksem kui keskmine ($4 < 4{,}625$), siis nüüd on see suurem ($6 > 4{,}875$) ja punkt asub keskmisele vastavast joonest kõrgemal (joonis 8.5 (b)). Selle tagajärjel muutus vastav korrutis $(x_2 - \bar{x})(y_2 - \bar{y})$

negatiivseks ning vähenes korrutiste summa $\sum(x_i - \bar{x})(y_i - \bar{y}) = 23{,}5$. Nüüd on kovariatsioon väiksem: $cov_{XY} = 23{,}5/8 \approx 2{,}94$.

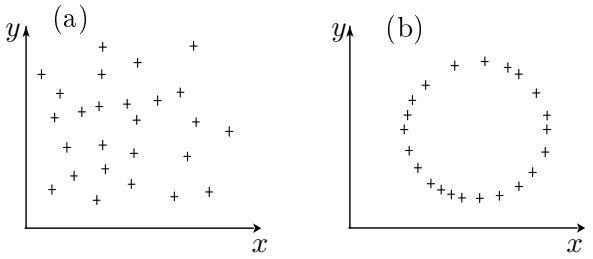
Kovariatsiooni omadusi

Mõningaid kovariatsiooni **omadusi**, mis järelduvad valemist (8.1).

1. Kovariatsioon on suuruste X ja Y suhtes sümmeetriline, $cov_{XY} = cov_{YX}$.
2. Kovariatsiooni ühikuks on suuruste X ja Y ühikute korrutis.
3. Kovariatsiooni väärtus võib olla vahemikus

$$-\sigma_X\sigma_Y < cov_{XY} < \sigma_X\sigma_Y. \tag{8.3}$$

4. Kui $cov_{XY} > 0$, siis on tegemist **positiivse** korrelatsiooniga: suuruse X kasvades kasvab keskmiselt ka suurus Y;
5. kui $cov_{XY} < 0$, siis on tegemist **negatiivse** korrelatsiooniga: suuruse X kasvades suurus Y keskmiselt kahaneb.
6. Kui juhuslikud suurused X ja Y on **sõltumatud**, siis $cov_{XY} = 0$ (joonis 8.6 (a)). Vastupidine ei kehti, kovariatsioon võib olla null ka siis, kui X ja Y sõltuvad teineteisest (joonis 8.6 (b)).
7. Kui $cov_{XY} \neq 0$, siis nimetatakse suurusi X ja Y **korreleeruvateks**.



Joonis 8.6. Mõlema punktiparve korral $cov_{XY} = 0$. Paneel (a): X ja Y on sõltumatud juhuslikud suurused. Paneel (b): X ja Y vahel on sõltuvus $(x_i - a)^2 + (y_i - b)^2 = r^2$, kus a ja b on ringjoone keskpunti koordinaadid ning r ringjoone raadius

Sõltumatute juhuslike suuruste kovariatsioon on null.

Kovariatsioon on üks statistilistest momentidest ja nimelt **2. järku segakeskmoment**:

- 2. järku, sest summeerimine toimub üle kahe teguri **korrutise**;
- segamoment, sest leitakse kahe **erineva** suuruse X ja Y põhjal;
- keskmoment, sest summeeritavates korrutistes on teguriteks erinevused aritmeetilisest **keskmisest** $(x_i - \bar{x})$ $(y_i - \bar{y})$.

Juhuslike suuruste summa dispersioon avaldub nende kovariatsioonide kaudu. Üldiselt, kui meil on tegemist n juhusliku suurusega X_k,

$k = 1, \ldots, n$, siis nende summa $\sum X_k$ dispersioon on

Juhuslike suuruste summa dispersioon

$$\sigma^2_{\sum X_k} = \sum_{i=1}^{n} \sum_{j=1}^{n} cov_{X_i X_j}. \tag{8.4}$$

Rakendame valemit 8.4 kahele juhuslikule suurusele. Arvestame seda, et kui summas (8.4) $i = j$, siis $cov_{X_i X_i} = \sigma^2_{X_i}$.

Kahe juhusliku suuruse **summaarne dispersioon** on võrdne nende juhuslike suuruste dispersioonide summaga, millele on liidetud kahekordne nendevaheline kovariatsioon:

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y + 2cov_{XY}. \tag{8.5}$$

Näiteks tuleb summaarset dispersiooni (8.4) kasutada paljudest väärtpaberitest koosneva väärtpaberiportfelli tulumäära standardhälbe arvutamiseks. Kahest väärtpaberist A ja B koosneva portfelli tulumäära dispersioon on

$$\sigma^2_P = w^2_A \sigma^2_A + w^2_B \sigma^2_B + 2 w_A w_B cov_{AB}, \tag{8.6}$$

kus σ^2_A ja σ^2_B on vastavate väärtpaberite tulumäärade dispersioonid, w_A väärtpaberi A osakaal portfellis, $w_B = 1 - w_A$ väärtpaberi B osakaal ning cov_{AB} tulumäärade vaheline kovariatsioon.

Näide 8.1. Aktsiaportfelli tulumäära dispersioon ja portfelli diversifitseerimine

Ajavahemikul 1.05.–1.06.2012 olid kahe New Yorgi börsil kaubeldava aktsia tulumäär ja standardhälve ning nendevaheline kovariatsioon järgmised[a]:

	Keskmine tulumäär	Tulumäära standardhälve
JPMorgan Chase & Co (JPM)	−1,39%	2,64%
AT&T Inc. (T)	0,116%	0,66%
Kovariatsioon	$-2{,}42 \cdot 10^{-5}$	

Aktsiaportfellis olgu aktsia JPM osakaal 9% ja aktsia T osakaal 91%. Valemist (8.6) on portfelli tulumäära dispersioon

$$\sigma^2_P = 0{,}09^2 \cdot 0{,}0264^2 + 0{,}91^2 \cdot 0{,}0066^2 + \\ + 2 \cdot 0{,}09 \cdot 0{,}91 \cdot (-2{,}42 \cdot 10^{-5}) = 3{,}78 \cdot 10^{-5}.$$

Portfelli tulumäära standardhälve on siis $\sigma_P = \sqrt{3{,}78 \cdot 10^{-5}} = 0{,}61\,\%$, mis on väiksem kui kummagi aktsia tulumäära standardhälve.
Negatiivses seoses olevate aktsiate korral ühe aktsia tulumäära langemisel teise oma tõuseb ja diversifitseerimine on investeerimisriski hajutamine.

[a] Allikas: Wolfram Mathematica *Financial data*

Kui meil on tegemist valimiga ning soovime hinnata üldkogumi kovariatsiooni, siis tuleb kasutada valemit

Valimi kovariatsioon

$$cov_{XY} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y}\,)}{n-1}, \tag{8.7}$$

mida nimetatakse **valimi kovariatsiooniks**.

Tabelarvutuses kasutatakse kogumi kovariatsiooni leidmiseks funktsiooni **COVARIANCE.P**. See leiab kogumi kovariatsiooni, kasutades valemit (8.1). Valimi kovariatsiooni jaoks on funktsioon **COVARIANCE.S**.

Kas kovariatsiooni saab kasutada erinevate suurusepaaride vahelise seose võrdlemiseks? Soovime näiteks võrrelda, kumb on imiku vanusega rohkem seotud, kas kehakaal või pikkus. Kuue kuu jooksul paneme perioodiliselt kirja kõigi kolme suuruse väärtused. Vanust mõõdame kuudes, kaalu kilogrammides ja pikkust sentimeetrites. Kui kovariatsioon vanuse ja kaalu vahel tuleb $1{,}44\,\text{kuu} \cdot \text{kg}$ ning vanuse ja pikkuse vahel $7{,}33\,\text{kuu} \cdot \text{cm}$, siis kumb seos on tugevam? Neil kahel arvul on erinevad mõõtühikud ja me ei saa neid võrrelda. Võrdlemiseks oleks vaja ühikuta suurust.

8.3. Lineaarne korrelatsioonikordaja

Kovariatsiooni puuduseks on see, et kovariatsiooni väärtus sõltub kasutatud ühikutest. Seepärast ei saa seda kasutada erinevates ühikutes mõõdetud tunnusepaaride vaheliste seoste võrdlemiseks. Seose tugevuse hindamiseks sobib paremini selline arvkarakteristik, mis ei sõltu ühikutest ja mille absoluutväärtus on vahemikus 0 kuni 1. Selleks tuleb kovariatsiooni normeerida.

Kui me vaatame kovariatsiooni omadust (8.3), siis sobiv oleks kovariatsioon läbi jagada standardhälvete korrutisega $\sigma_X \sigma_Y$. Selline jagatis muutub vahemikus $[-1, 1]$ ning on ühikuta suurus. Normeerimiseks jagataksegi kovariatsioon läbi mõlema suuruse standardhälbega ja saadud suurust nimetatakse korrelatsioonikordajaks:

$$r = \frac{cov_{XY}}{\sigma_X \sigma_Y}. \tag{8.8}$$

Asendades seosesse (8.8) kovariatsiooni avaldise (8.1), saame valemi korrelatsioonikordaja arvutamiseks.

Lineaarne korrelatsioonikordaja

Lineaarne ehk **Pearsoni korrelatsioonikordaja**:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n\sigma_X\sigma_Y}, \tag{8.9}$$

kus n on juhuslike suuruste X ja Y väärtuste x_i ja y_i paaride arv (valimi maht), $\bar{x}$ ja $\bar{y}$ aritmeetilised keskmised ning σ_X ja σ_Y vastavad standardhälbed.

Korrelatsioonikordaja omadusi

Mõningaid lineaarse korrelatsioonikordaja **omadusi**:

1) korrelatsioonikordaja on tunnuste X ja Y suhtes sümmeetriline, $r_{XY} = r_{YX}$;
2) korrelatsioonikordaja on ühikuta suurus;
3) korrelatsioonikordaja väärtus on vahemikus $-1 \leq r \leq 1$;
4) korrelatsioonikordaja **absoluutväärtus** iseloomustab seose tugevust, **märk** seose suunda:
   - kui $r_{XY} > 0$, siis on tegemist **positiivse** korrelatsiooniga: suuruse X kasvades kasvab keskmiselt ka suurus Y;
   - kui $r_{XY} < 0$, siis on tegemist **negatiivse** korrelatsiooniga: suuruse X kasvades suurus Y keskmiselt kahaneb;
   - kui $r = 0$, siis korrelatsioon puudub;
   - kui $|r| = 1$, on tegemist täielikult korreleeruvate suurustega, hajumisdiagrammil asuvad kõik punktid ühel sirgel.

Millal nimetada seost tugevaks, millal nõrgaks? See on kokkuleppe küsimus, sageli kasutatakse järgmisi piire:

- nõrk seos $|r| \leq 0{,}3$;
- keskmise tugevusega seos $0{,}3 < |r| < 0{,}7$;
- tugev seos $|r| \geq 0{,}7$.

Kuid palju sõltub valimi mahust. Näiteks kui $r = 0{,}6$ valimi mahu $n = 100$ korral, võib rääkida keskmise tugevusega seosest. Aga sama korrelatsioonikordaja väärtus valimi mahu $n = 10$ korral ei tõesta, et seos üldse olemas oleks. Kriitiline väärtus, millest korrelatsioonikordaja absoluutväärtus peab olema suurem, et seos eksisteeriks, sõltub valimi mahust. Korrelatsioonikordaja statistilise olulisuse testimist vaatame alapeatükis 8.4.

Tabelarvutuses kasutatakse korrelatsioonikordaja leidmiseks funktsiooni **CORREL** või **PEARSON**, mõlemad arvutavad valemi (8.9) järgi ja väljastavad sama tulemuse.

Tabelis 8.2 ja joonisel 8.1 toodud turunduskulude ning käibe vahelise seose korrelatsioonikordaja on 0,945, mis näitab tugevat positiivset seost.

Näide 8.2. Korrelatsioonikordajad mõningate näitajate vahel Eesti maakondades


N08Korrelatsioon
N8.2

Kasutame Eesti maakondade andmeid aastast 2009 ja analüüsime seoseid järgmiste näitajate vahel[a]:

- töötuse määr, protsent kogu tööhõivelisest elanikkonnast;
- keskmine brutokuupalk, eurot;
- sündinud ettevõtete arv aastas;
- hõivatute osatähtsus sekundaarsektoris (töötlev tööstus), protsentides;
- hõivatute osatähtsus tertsiaalsektoris (teenindav sektor), protsentides.

Maakondadest on välja jäetud Harjumaa, sest koos Tallinnaga erineb see teistest väga palju. Paarikaupa arvutatud korrelatsioonikordajad on esitatud tabelis.

	Töötuse määr	Keskmine bruto-kuupalk	Sündinud ette-võtted	Sekundaar-sektor	Tertsiaal-sektor
Töötuse määr	1				
Keskmine brutokuu-palk	−0,521	1			
Sündinud et-tevõtted	−0,219	0,852	1		
Sekundaar-sektor	0,175	−0,173	−0,260	1	
Tertsiaal-sektor	−0,043	0,310	0,467	−0,891	1

Kõige **tugevamini** on seotud hõivatute osakaal sekundaar- ja tertsiaalsektoris, korrelatsioonikordaja absoluutväärtus on kõige suurem. Tegemist on negatiivse korrelatsiooniga: nendes maakondades, kus sekundaarsektoris on hõivatuid rohkem, on tertsiaalsektoris vähem ja vastupidi. See on ka loomulik, sest need on osakaalud ühest ja samast hõivatute kogumist.

Kõige **nõrgem** on seos töötuse määra ja tertsiaalsektoris hõivatute osakaalu vahel, korrelatsioonikordaja absoluutväärtus on kõige väiksem.

Analüüsime töötuse määra seost teiste näitajatega. Kõige tugevamini on töötuse määr seotud keskmise brutokuupalgaga (korrelatsioonikordaja absoluutväärtus on kõige suurem). Kuna vastav korrelatsioonikordaja on negatiivne, siis nendes maakondades, kus keskmine kuupalk on suurem, on töötuse määr väiksem.

Nõrk negatiivne seos on töötuse määra ja sündinud ettevõtete arvu vahel: neis maakondades, kus luuakse rohkem ettevõtteid, on töötuse määr üldiselt väiksem. Töötuse määral puudub seos hõivatute osakaaluga tertsiaalsektoris (vastav korrelatsioonikordaja praktiliselt null). Nõrk positiivne seos on hõivatute osakaaluga sekundaarsektoris.
Kui vaadata ülejäänud tunnuste omavahelisi seoseid, siis keskmise kuupalga ja sündinud ettevõtete arvu vahel on tugev positiivne seos. Sündinud ettevõtete arv on positiivselt seotud ka hõivatute osakaaluga tertsiaalsektoris ja negatiivselt seotud hõivatute osakaaluga sekundaarsektoris. Siit võib järeldada, et sündinud ettevõtted tekivad just tertsiaalsektoris (teenindus).

[a] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabelid PA5321, TT442 ja TT241.

Näites 8.2 kasutasime korrelatsioonikordajate esitamiseks korrelatsioonimaatriksit. See on sobiv korrelatsioonikordajate esitamise viis, kui on vaja analüüsida paarikaupa esinevaid seoseid paljude erinevate tunnuste korral.

Korrelatsioonimaatriks

Korrelatsioonimaatriks on juhuslike suuruste $X_1, X_2, \ldots X_K$ vahelist statistilist sõltuvust iseloomustav ruutmaatriks, mille elementideks on lineaarsed korrelatsioonikordajad:

$$r_{ij} = r_{X_i X_j},$$

kus i loendab ridu ja j veerge.

Korrelatsioonikordaja iseendaga on 1 (vt valem (8.9) ning korrelatsioonimaatriksi diagonaalil asuvad seetõttu ühed. Korrelatsioonikordaja sümmeetriaomadusest lähtuvalt (omadus 1) on korrelatsioonimaatriks peadiagonaali suhtes sümmeetriline ning seetõttu jäetakse ülevalpool peadiagonaali asuvad kohad tavaliselt tühjaks, sest pole mõtet korelatsioonikordajate väärtusi topelt esitada.

Korrelatsioonimaatriksi leidmiseks programmis Excel kasutatakse vahendit *Correlation* komplektist *Data Analysis*.

Autokorrelatsioon

Aegridade analüüsimisel kasutatakse mingi suuruse muutumise juhuslikkuse või mittejuhuslikkuse hindamisel autokorrelatsiooni (*serial correlation*). **Autokorrelatsioon** on ajas muutuva tunnuse väärtuste korrelatsioon ühe perioodi võrra nihkes oleva sama suuruse väärtustega. Kui suuruse X väärtuste rida on $x_1, x_2, \ldots, x_t, \ldots, x_T$, siis autokorrelatsiooni kordaja leidmiseks moodustatakse $T-1$ järjestikuste

väärtuste paari (x_t, x_{t-1}) ja leitakse vastav korrelatsioonikordaja. Kui see on nullilähedane, on tegemist juhuslikult muutuva aegreaga. Kui aga korrelatsioonikordaja absoluutväärtus on suur, siis on ajahetkele t vastav väärtus x_t tugevas korrelatsioonis eelmise väärtusega x_{t-1} ja esineb autokorrelatsioon.

Näiteks, kui aktsia hinna muutused ei ole juhuslikud, siis hinnatõusule ühel päeval peaks järgnema hinnatõus ka järgmisel päeval. Efektiivse turu hüpotees väidab, et aktsiate hindade muutus on juhuslik.

Näide 8.3. Aktsia tulumäärade autokorrelatsioon

N08Korrelatsioon
N8.3

Leiame Harju Elektri aktsia tulumäära autokorrelatsiooni kordaja ajavahemikul 1.03.–30.11.2010. Aegrea pikkus on 193. Tabelis on toodud näide, kuidas võib arvutusi organiseerida, x_t on tulumäär protsentides.

Kuupäev	x_t	x_{t-1}
1.03.2010	1,96	
2.03.2010	0,77	1,96
3.03.2010	−0,76	0,77
4.03.2010	0,77	−0,76
5.03.2010	−1,15	0,77
...	...	...

Autokorrelatsioonikordaja leidmiseks on 192 arvupaari, sest 1.03.2010 esinenud väärtuse jaoks paarilist pole. Harju Elektri aktsia tulumäära autokorrelatsiooni kordaja tuleb −0,149. Võrdluseks on leitud sama perioodi alusel Baltika aktsia tulumäära autokorrelatsiooni kordaja, mis on 0,009. Järelikult oli Baltika aktsia tulumäära muutumises juhuslikkust rohkem.

8.4. Korrelatsiooni statistiline olulisus

Empiiriliste andmete põhjal leitud lineaarse korrelatsioonikordaja väärtus võib erineda nullist ka täiesti sõltumatute tunnuste puhul, seepärast on vaja hinnata korrelatsioonikordaja statistilist olulisust. Selleks kasutatakse standardset hüpoteeside statistilise kontrollimise skeemi, kus tuleb püstitada nullhüpotees ja sisukas hüpotees, leida parameetri empiiriline ja kriitiline väärtus ning otsuse vastuvõtmiseks neid võrrelda. Korrelatsioonikordaja testimisel kasutatakse t-testi.

Korrelatsiooni olulisuse testimine

Korrelatsiooni statistilise olulisuse testimine

1. Hüpoteesipaar

H_0: $r = 0$ korrelatsioon puudub,
H_1: $r \neq 0$ korrelatsioon esineb.

2. Parameetri empiiriline väärtus leitakse valemist

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \tag{8.10}$$

kus r on korrelatsioonikordaja ja n arvupaaride arv.

3. Kriitiline väärtus olulisuse nivool α on t-jaotuse täiendkvantiil $t_{\alpha/2}(\nu)$, kus vabadusastmete arv

$$\nu = n - 2. \tag{8.11}$$

4. Võtta vastu

H_0, kui $\quad |t| \leq t_{\alpha/2}(\nu)$;
H_1, kui $\quad |t| > t_{\alpha/2}(\nu)$.

Mõnikord, kui soovitakse testida positiivse või negatiivse korrelatsiooni olemasolu, kasutatakse ühepoolset testi.

Ühepoolne hüpotees

	Soovime tõestada positiivse seose olemasolu	Soovime tõestada negatiivse seose olemasolu
H_0	$r \leq 0$	$r \geq 0$
H_1	$r > 0$	$r < 0$

Parameetri empiiriline väärtus leitakse valemist (8.10) ja kriitiliseks väärtuseks olulisuse nivool α on t-jaotuse täiendkvantiil $t_\alpha(\nu)$.

Võtta vastu

H_0, kui $\quad t \leq t_\alpha(\nu), \qquad t \geq -t_\alpha(\nu)$,
H_1, kui $\quad t > t_\alpha(\nu), \qquad t < -t_\alpha(\nu)$.

Näide 8.4. Kas keskmine eluiga sõltub elanikkonna rikkusest?

N08Korrelatsioon
N8.4

Kasutame andmeid 91 riigi kohta: meeste eluiga ja rahvuslik kogutoodang (RKT). Leiame, kas nende suuruste vahel on seos.

1. Kasutame kahepoolset testi.
2. Hüpoteesipaar:
   H_0: $r = 0$ korrelatsioon puudub,
   H_1: $r \neq 0$ korrelatsioon esineb.
3. Tabelarvutuses leiame korrelatsioonikordaja $r = 0{,}643$. Parameetri t empiiriline väärtus

$$t = \frac{0{,}643\sqrt{91-2}}{\sqrt{1-0{,}643^2}} \approx 7{,}9.$$

4. Olulisuse nivoo võtame $\alpha = 0{,}05$ ja vabadusastmete arv on $\nu = 91 - 2 = 89$.
Statistiku kriitilise väärtuse leiame tabelarvutuses: T.INV.2T(0,05;89)= 1,99.
5. $|7{,}9| > 1{,}99$, empiiriline väärtus on kriitilisest väärtusest suurem.
6. Võtame vastu sisuka hüpoteesi: meeste keskmise eluea ja RKT vahel on statistiliselt oluline seos olulisuse nivool 5%.

Korrelatsioonimaatriksi kasutamisel ei ole mõtet iga üksiku korrelatsioonikordaja jaoks läbi teha statistilise olulisuse testi. Korrelatsioonimaatriks leitakse valimi põhjal, mille maht on n. See tähendab, et kõigi maatriksis olevate korrelatsioonikordajate arvutamisel on arvupaaride arv n ühesugune. Kõikide korrelatsioonikordajate testimisel kasutame ka üht ja sama olulisuse nivood α. Järelikult on t-testi kriitiline väärtus ühesugune kõigi korrelatsioonimaatriksis olevate korrelatsioonikordajate jaoks ning selle põhjal võime arvutada kriitilise korrelatsioonikordaja väärtuse antud korrelatsioonimaatriksi jaoks. Kriitilise korrelatsioonikordaja valem saadakse valemist (8.10), avaldades sealt korrelatsioonikordaja r.

Kriitiline korrelatsioonikordaja

Kui valimi maht on n ja kasutame olulisuse nivood α, siis lineaarse korrelatsioonikordaja **kriitiline väärtus** on

$$r_{kr} = \frac{1}{\sqrt{1 + \dfrac{\nu}{t^2_{\alpha/2}(\nu)}}}, \tag{8.12}$$

kus $t_{\alpha/2}(\nu)$ on t-jaotuse täiendkvantiil ja vabadusastmete arv $\nu = n - 2$.
Korrelatsioonikordaja absoluutväärtust tuleb võrrelda kriitilise väärtusega. Kui

$|r| \leq r_{kr}$, siis võtta vastu H_0, korrelatsioon puudub;
$|r| > r_{kr}$, siis võtta vastu H_1, korrelatsioon esineb.

Lineaarse korrelatsioonikordaja kriitilised väärtused mõningate erineva suurusega valimite korral on toodud lisas B.5.

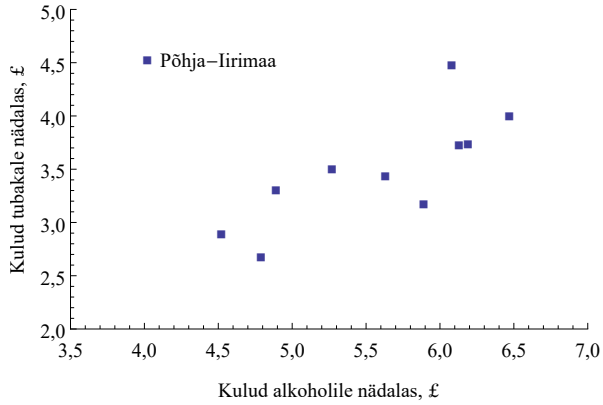
8.5. Lineaarse korrelatsioonikordaja puudused

Lineaarse korrelatsioonikordaja kasutamisel tuleb arvestada mõningate nüanssidega, alati ei pruugi lineaarse korrelatsioonikordaja suurus anda meile objektiivset informatsiooni.

Näide 8.5. Kulud alkoholile ja tubakale


N08Korrelatsioon
N8.5

Suurbritannias viidi 1981. aastal läbi küsitlus perede kulutuste kohta. Muu hulgas uuriti kulutusi tubakale ja alkoholile 11 erinevas Suurbritannia piirkonnas (Moore ja McCabe, 1989). Kui analüüsiti, kas kulud alkoholile ja tubakale on omavahel seotud, siis korrelatsioonikordaja tuli väga väike: 0,22. See on väiksem kui kriitiline väärtus, mis valimi mahu $n = 11$ korral olulisuse nivool 0,05 on 0,602 (vt lisa B.5).



Hajumisdiagrammi uurides aga selgub, et üks punkt asub teistest eraldi. See vastab Põhja-Iirimaale (*Nothern Ireland*). Kui see piirkond välja jätta ja leida korrelatsioonikordaja ülejäänud 10 piirkonna põhjal, saame, et seos on statistiliselt oluline, korrelatsioonikordaja on 0,78.

Näites 8.5 mõjutas seose tugevust üks, teistest tugevasti erinev vaatlus. See on erind. Erind võib tugevasti vähendada lineaarset kor-

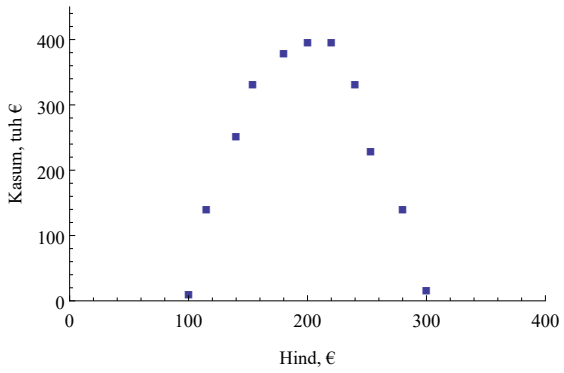
relatsioonikordajat juhul, kui tugev seos on olemas. Samuti võib erind oluliselt suurendada korrelatsioonikordajat juhul, kui tegelikult tugevat seost ei ole. Seetõttu tuleb lisaks korrelatsioonikordaja arvutamisele uurida alati ka hajumisdiagrammi. Iga erindit tuleb analüüsida ja otsustada, kas see kuulub samasse kogumisse koos ülejäänud vaatlustega või tuleb erind välja jätta.

Erindid mõjutavad kergesti lineaarset korrelatsioonikordajat.

Lineaarse korrelatsioonikordaja väärtus võib olla eksitav ka erindite puudumisel. Tunnuste vahel võib olla väga tugev seos, kuid lineaarne korrelatsioonikordaja ei viita selle olemasolule.

Näide 8.6. Mittelineaarne seos

Ettevõttes analüüsiti toote hinna ja kasumi vahelist seost. Hajumisdiagrammilt (joonis 8.7) on näha, et vaatluspunktid asuvad piki paraboolset kõverat. Ka teoreetilistest arvutustest on teada, et kui kulufunktsioon ja nõudlusfunktsioon on lineaarsed, avaldub kasumi sõltuvus hinnast ruutfunktsioonina. Seose tugevuse leidmiseks leiti ka korrelatsioonikordaja mille väärtus tuli 0,0058.

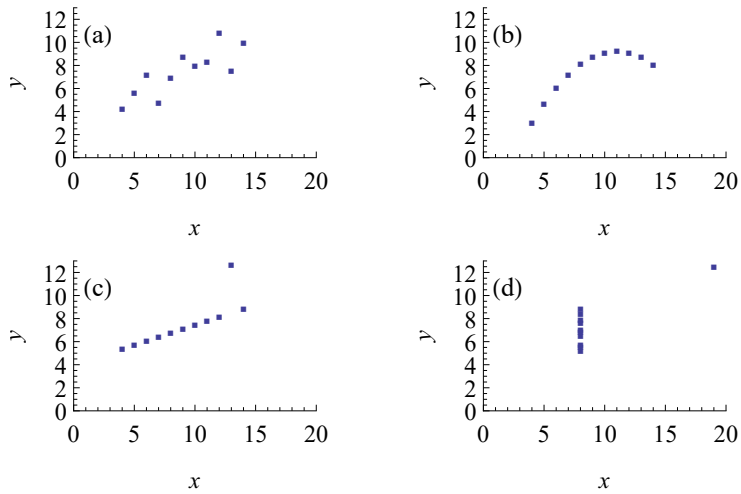


Joonis 8.7. Mittelineaarne seos

Lineaarse korrelatsioonikordaja arvutusvalemi tuletamisel on lähtutud sellest, et kahe tunnuse vaheline seos on modelleeritav lineaarse mudeliga. Seepärast seda nimetataksegi **lineaarseks** korrelatsioonikordajaks. See „tunneb ära“ punktide kogumi, mis on välja venitatud piki sirget. Kui punktikogum järgib mingit mittelineaarset kõverat, siis seose tugevuse hindamiseks lineaarne korrelatsioonikordaja ei sobi.

Lineaarne korrelatsioonikordaja iseloomustab vaid lineaarse seose tugevust.

Mõlemat tüüpi eksitust (erind, mittelineaarsus) on võimalik vältida hajumisdiagrammide uurimisel. Hajumisdiagrammiga tutvumise tähtsuse rõhutamiseks on joonisel 8.8 esitatud nelja erineva punktikogumi hajumisdiagrammid, mida nimetatakse Anscombe'i kvartetiks. Need konstrueeris 1973. aastal inglise statistik Francis Anscombe (Anscombe, 1973). Igas kogumis on 11 punktipaari ja kõikidel juhtudel on lineaarne korrelatsioonikordaja 0,816. Hajumisdiagrammil (a) on tüüpiline korrelatiivne seos, mille tugevuse hindamiseks sobib lineaarne korrelatsioonikordaja. Hajumisdiagrammil (b) on funktsionaalne mittelineaarne seos. Diagrammidel (c) ja (d) on selgesti eristavad erindid.



Joonis 8.8. Anscombe'i kvartett. Kõikide hajumisdiagrammide korral on lineaarne korrelatsioonikordaja 0,816

Lineaarse korrelatsioonikordaja puuduste tõttu kasutatakse ka teisi seosekordajaid (nt Spearmani korrelatsioonikordaja, Kendalli kordaja).

8.6. Astakkorrelatsioon

Küsitluste puhul palutakse tihti reastada mingid suurused näiteks meeldivuse, olulisuse vms järgi. Sellisel juhul kasutatakse järjestusskaalat. Kui soovime analüüsida, kui hästi langevad erinevate vastajarühmade hinnangud kokku, tuleb võrrelda eri rühmade järjestatud tegurite järjenumbreid ehk **astakuid**. Selleks kasutatakse **astakkorrelatsiooni**.

Järjenumbrite ehk astakute korrelatsioonikordaja, mida nimetatakse **Spearmani korrelatsioonikordajaks**, leitakse valemist

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \tag{8.13}$$

kus d_i on erinevates gruppides kõrvuti olevate järjekorranumbrite vahe ja n väärtuspaaride arv.

Spearmani korrelatsioonikordaja

Tabelis 8.5 on valikud A kuni B järjestatud isikute 1 ja 2 poolt täpselt ühtemoodi. Kuna kõik erinevused d_i on nullid, siis ka summa $\sum d_i^2 = 0$ ja Spearmani korrelatsioonikordaja tuleb 1. Tabelis 8.6 on valikud A kuni B järjestatud isikute 1 ja 3 poolt täpselt vastupidi. Spearmani korrelatsioonikordaja on siis

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 40}{5(5^2-1)} = -1.$$

Tabel 8.5. Kaks järjestust on täpselt ühesugused

Valik	Järjestus 1	Järjestus 2	d_i^2
A	1	1	0
B	2	2	0
C	3	3	0
D	4	4	0
E	5	5	0

Tabel 8.6. Kaks järjestust on täpselt vastupidised

Järjestus 1	Järjestus 3	d_i^2
1	5	16
2	4	4
3	3	0
4	2	4
5	1	16

Järelikult on Spearmani korrelatsioonikordaja väärtus vahemikus $[-1, 1]$. Kui järjestus on täpselt ühesugune, siis $r_S = 1$, täpselt vastupidise järjestuse korral $r_S = -1$.

Näide 8.7. Motivatsioonifaktorid ja Spearmani korrelatsioonikordaja

N08Korrelatsioon 8.7

Ettevõtte töötajaid saab motiveerida mitmeti. Töötasu on ainult üks peamistest motivatsioonifaktoritest. Lisaks rahale sisaldab aga motivatsioonisüsteem ka mitterahalist osa. Motivatsioonisüsteemi loomisel tuleks uurida, mida töötajad tähtsaks peavad. Ettevõttes X läbi viidud uuringus paluti töötajatel järjestada motivatsioonifaktorid A kuni I tähtsuse järjekorras. Tippjuhid pidid faktorid reastama nii, nagu arvasid keskastme juhte neid reastavat. Tippjuhtide väljapakutud pingerida võrreldi sellega,

kuidas keskastmejuhid ise samad motivatsioonifaktorid reastasid.

Faktor	Tippjuhid	Keskastmejuhid	d_i	d_i^2
A	1	5	−4	16
E	4	8	−4	16
B	2	2	0	0
D	5	7	−2	4
H	3	6	−3	9
C	7	1	6	36
F	8	4	4	16
I	6	3	3	9

Kokkulangevuse kvantitatiivseks hinnanguks saab kasutada Spearmani korrelatsioonikordajat. Viimase veeru summa $\sum d_i^2 = 106$, valikute arv $n = 8$. Spearmani korrelatsioonikordaja

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 106}{8(8^2-1)} = -0{,}262.$$

Kuna Spearmani korrelatsioonikordaja on negatiivne, siis järelikult tippjuhid ei tea, mismoodi motiveerida keskastmejuhte: tippjuhtide arvamus keskastmejuhtidele olulistest motivatsioonifaktoritest ei pea paika.

Astakkorrelatsiooni kasutatakse ka intervallskaalas mõõdetud kvantitatiivsete tunnuste vahelise seose kirjeldamisel, kui seos on mittelineaarne või on vaja vähendada erindite mõju. Selleks leitakse valimis elementide järjenumbrid ehk astakud mõlema tunnuse alusel ning seejärel arvutatakse Spearmani korrelatsioonikordaja astakute vahel. Kui kahe või enama mõõtmistulemuse väärtused on võrdsed, määratakse võrdsetele väärtustele sama astak, mis on arvuliselt võrdne vastavate astakute keskväärtusega.

Tabeli 8.7 kahes esimeses veerus on intervallskaalas mõõdetud suuruste X ja Y väärtused. Viimases kahes veerus on nende väärtuste astakud oma variatsioonreas. Lineaarne korrelatsioonikordaja tuleb 0,874, aga Spearmani korrelatsioonikordaja on 1. Tegemist on monotoonse seosega: kui X kasvab, siis alati kasvab ka Y (joonis 8.9 (a)).

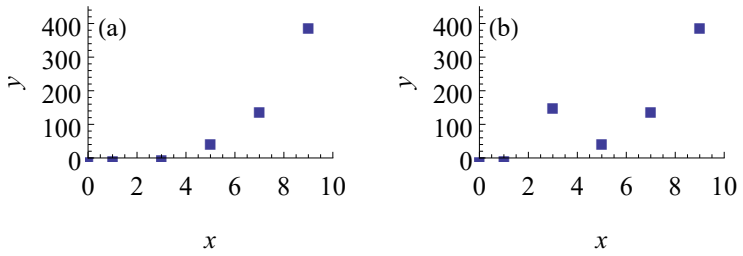
Tabelis 8.8 aga on monotoonne kasvamine rikutud. Kui X väärtus kasvab väärtuselt 3 väärtuseni 5, siis Y samal ajal kahaneb väärtuselt 160 väärtuseni 55 (joonis 8.9 (b)). Spearmani korrelatsioonikordaja tuleb nüüd 0,829.

Tabel 8.7. Monotoonne kasv, vt joonis 8.9 (a)

x_i	y_i	x astak	y astak
0	1	1	1
1	3	2	2
3	8	3	3
5	55	4	4
7	150	5	5
9	400	6	6

Tabel 8.8. Mittemonotoonne kasv, vt joonis 8.9 (b)

x_i	y_i	x astak	y astak
0	1	1	1
1	3	2	2
3	**160**	3	5
5	55	4	3
7	150	5	4
9	400	6	6



Joonis 8.9. (a) monotoonne kasv, (b) mittemonotoonne kasv

Monotoonne seos

Sõltuvust nimetatakse **monotoonseks**, kui ühe tunnuse kasvamine toob kaasa teise tunnuse kasvamise ning ühe tunnuse kahanemine toob kaasa teise tunnuse kahanemise.

- Pearsoni korrelatsioonikordaja mõõdab lineaarse seose tugevust.
- Spearmani korrelatsioonikordaja mõõdab monotoonse seose tugevust:
  - kui tunnuste vahel on kasvav rangelt monotoonne seos, on Spearmani korrelatsioonikordaja väärtus 1;
  - kui tunnuste vahel on kahanev rangelt monotoonne seos, on Spearmani korrelatsioonikordaja väärtus -1.

On võimalik näidata, et kui astakud on kõik erinevad, siis Spearmani korrelatsioonikordaja arvutusvalem langeb kokku lineaarse korrelatsioonikordaja valemiga.

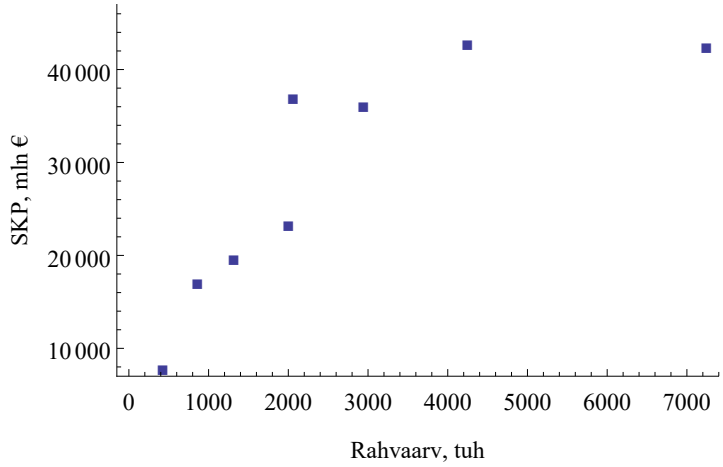
Näide 8.8. Lineaarne ja Spearmani korrelatsioonikordaja

Suurema rahvaarvuga riikides peaks SKP olema ka suurem. Selle hüpoteesi kontrollimiseks analüüsime Euroopa Liidu kaheksa kõige väiksema sisemajanduse koguproduktiga riigi SKP ja rahvaarvu 2014. aastal[a]. Rahvaarvu ja SKP vaheline lineaarne korrelatsioonikordaja on 0,818.

N08Korrelatsioon
N8.8

Spearmani korrelatsioonikordaja leidmiseks on leitud kummagi suuruse astak iga riigi jaoks. Seejärel on leitud astakute vahed ja nende põhjal Spearmani korrelatsioonikordaja, mis tuleb 0,952. See tähendab, et seos ei ole päris lineaarne, küll aga tugevalt monotoonne.

Riik	Rahvaarv, tuh	SKP, mln €	Rahvaarvu astak	SKP astak	Astakute vahe d_i
Horvaatia	4246,8	43084,8	2	1	1
Bulgaaria	7245,7	42750,9	1	2	−1
Sloveenia	2061,1	37303,2	4	3	1
Leedu	2943,5	36444,4	3	4	−1
Läti	2001,5	23580,9	5	5	0
Eesti	1315,8	19962,7	6	6	0
Küpros	858,0	17393,7	7	7	0
Malta	425,4	8106,1	8	8	0



[a] Allikas: Eurostat.

8.7. Ülesanded

Kovariatsioon

8.1. Tabelis on toodud kahe juhusliku suuruse väärtused. Leida nendevaheline kovariatsioon. VASTUS lk 679.

X	1	2	6	7
Y	2	1	3	6

8.2. Aktsia A tulumäära standardhälve on 7% ja aktsia B tulumäära standardhälve 14%. Tulumääradevaheline kovariatsioon on 0,00147.

Nendest kahest aktsiast moodustatakse aktsiaportfell, kus aktsia A osakaal on 60%. Kui suur on portfelli standardhälve? VASTUS lk 679.

8.3. Olgu meil kaks aktsiat A ja B. Aktsia A tulumäärade standardhälve on σ_A ja aktsia B tulumäärade standardhälve on σ_B. Tulumääradevaheline kovariatsioon on cov_{AB}. Moodustame nendest aktsiatest aktsiaportfelli, kus aktsia A osakaal on w_A. Tuletada valem selle osakaalu leidmiseks, kui investori eesmärgiks on moodustada võimalikult väikese riskiga aktsiaportfell. VASTUS lk 679.

8.4. Kui suur peaks ülesandes 8.2 toodud aktsiate korral olema aktsia A osakaal portfellis, et portfelli risk oleks minimaalne? Kasutada eelmises ülesandes leitud valemit. Kui suur on sel juhul portfelli standardhälve? VASTUS lk 679.

8.5. Lineaarse korrelatsioonikordaja väärtus nelja erineva seose korral on:

Korrelatsioonikordaja

a) 0,7,
b) 0,1,
c) $-0{,}2$,
d) $-0{,}9$.

Milline seos on kõige tugevam ja milline kõige nõrgem? VASTUS lk 679.

8.6. Aktsia A tulumäära standardhälve on 15%, aktsia B tulumäära standardhälve on 20% ja nendevaheline korrelatsioonikordaja 0,3. Kui suur on tulumääradevaheline kovariatsioon? VASTUS lk 679.

8.7. 1998. aastal ajakirjas Journal of Financial Economics ilmunud artiklis „Larger board size and decreasing firm value in small firms" analüüsiti ettevõtte juhatuse suuruse seost muude ettevõtet iseloomustavate tunnustega (Eisenberg, Sundgren ja Wells, 1998). Uuring põhines Soome väikeettevõtete hulgas läbiviidud valimvaatlusel. Tabelis on toodud korrelatsioonikordajad juhatuse suuruse (liikmete arv) ja teiste tunnuste vahel. Juhatuse liikmete maksehäired iseloomustavad juhatuse liikmete personaalset finantsdistsipliini: krediitkaardivõlad, maksmata arved, maksmata maksud.

Korrelatsiooni olulisus

	Korrelatsioonikordaja juhatuse liikmete arvuga	Valimi maht
Varad (tuhat Soome marka)	0,074	879
Varad, logaritmitud	0,287	870
Varade muutus, logaritmitud	−0,029	871
Ettevõtte vanus aastates	0,147	879
Juhatuse liikmete maksehäired	−0,109	879

1. Leida, millised korrelatsioonikordajad on statistiliselt olulised nivool 0,05, 0,01, 0,001.

2. Kas võib väita, et vanemates ettevõtetes on juhatuse suurus suurem?
3. Kas võib väita, et suurema juhatuse korral on juhatuse liikmetel maksehäireid rohkem?

VASTUS lk 679.

Astak-korrelatsioon

8.8. Kas juhile vajalikud omadused on avalikus sektoris ja erasektoris ühesugused? Sellele küsimusele otsisid vastust artikli „Do senior managers differ in the public and private sector?“ autorid (Arroba ja Wedgwood-Oppenheim, 1994).

Autorid lähtusid R. M. Belbini (1981) formuleeritud kaheksast rollist, mis on vajalikud efektiivse meeskonnatöö jaoks. Järgnevalt on esitatud nende rollide lühikirjeldused.

- Esimees (*chair*): rahulik, enesekindel, tugeva eesmärgitunnetusega, oskab leida meeskonna tugevaid ja nõrku külgi.
- Töömesilane (*company worker*): konservatiivne, kohusetundlik, organiseerimisvõimeline ja etteaimatav, realiseerib kavandatud ülesanded.
- Vormistaja (*shaper*): dünaamiline, tarmukas.
- Ideede generaator (*plant*): intelligentne, kujutlusvõimeline, pakub välja uusi ideid ja strateegiaid.
- Ressursside uurija (*resource investigator*): entusiastlik, uudishimulik, suhtlemisaldis.
- Kriitik, hindaja (*monitor evaluator*): emotsioonideta, ettenägelik, analüüsib probleemide lahendamisvõimalusi, puudub inspiratsioon.
- Meeskonnas töötaja (*team worker*): sotsiaalse orientatsiooniga, edendab meeskonnavaimu.
- Lõpuleviija (*completer finisher*): püüdlik, kohusetundlik.

Artikli autorid palusid 157 Suurbritannia omavalitsuse juhtivtöötajal reastada eespool toodud omadused tähtsuse järjekorras. Saadud pingerida võrreldi Belbini uuringuga, kus samad omadused olid reastatud 78 peamiselt erasektoris töötava juhi poolt. Mõlemad pingeread on toodud järgnevas tabelis. Leida Spearmani korrelatsioonikordaja.

VASTUS lk 679.

Roll	Omavalitsused	Erasektor
Töömesilane	2	3
Esimees	3	4
Vormistaja	1	1
Ideede generaator	7	7
Ressursside uurija	6	6
Kriitik, hindaja	4	5
Meeskonnas töötaja	5	2
Lõpuleviija	8	8

Järgmiste ülesannete andmed on failis ÜL08Korrelatsioon

ÜL08Korrelatsioon

A.8.1. Tabelis on 91 maailma riigi kohta järgmised andmed 1990. aasta seisuga (UNESCO, 1990):

Korrelatsioonikordaja

SÜND sündide arv 1000 elaniku kohta;
SURM surmade arv 1000 elaniku kohta;
IM.SURM imikute (alla 1-aastaste) surmajuhtumid 1000 sünni kohta;
IGA M meeste keskmine eluiga;
IGA N naiste keskmine eluiga;
RKT rahvuslik kogutoodang ühe elaniku kohta (USD);
GRUPP riikide grupp:

1. Ida-Euroopa;
2. Lõuna-Ameerika ja Mehhiko;
3. Lääne-Euroopa, Põhja-Ameerika, Jaapan, Austraalia, Uus-Meremaa;
4. Kesk-Ida;
5. Aasia;
6. Aafrika;

RIIK riigi nimetus.

1. Leida korrelatsioonikordajad RKT ja ülejäänud tunnuste vahel.
2. Millised suurused on RKT-ga positiivselt seotud, millised negatiivselt?
3. Milline suurus on RKT-ga kõige tugevamini seotud? Kõige nõrgemini?
4. Leida korrelatsioonikordaja RKT ja sündide arvu vahel kahes riikide grupis eraldi: 3. grupp (Lääne-Euroopa, Põhja-Ameerika, Jaapan, Austraalia, Uus-Meremaa) ja 5. grupp (Aasia). Milline on järeldus?
5. Konstrueerida hajumisdiagrammid eelmises punktis nimetatud suuruste vahel: RKT ja sündide arv 3. ja 5. grupis.

VASTUS lk 679.

A.8.2. Tabelis on kuue börsiindeksi päevane sulgemisväärtus 1.01.–31.12.2014[1].

Korrelatsioonimaatriks

- USA
  - S&P 500 ehk Standard & Poor's 500 sisaldab New Yorgi börsil noteeritud 500 ettevõtte aktsiat;
  - DJIA ehk Dow Jones Industrial Average sisaldab 30 USA suurima ettevõtte aktsiat;

[1] Allikad: YAHOO/Finance, `https://uk.finance.yahoo.com`
FRED Economic Data, `http://research.stlouisfed.org/`

  - Nasdaq 100 on USA börsi NASDAQ indeks, mis sisaldab 100 ettevõtte aktsiat;
- Euroopa
  - FTSE 100 on Londoni börsi indeks, sisaldab 100 suurima kapitalisatsiooniga ettevõtte aktsiat;
  - OMXS on Stockholmi börsi indeks;
- Aasia
  - Nikkei 225 on Jaapani börsi indeks.

Indeksite väärtused leitakse ainult tööpäevadel, puhkepäevadel ja riiklikel pühadel väärtused puuduvad.

1. Leida korrelatsioonimaatriks, kus on kõikide indeksite paarikaupa leitud korrelatsioonikordajad.
2. Milliste indeksite vahel on kõige tugevam seos?
3. Milliste indeksite vahel on kõige nõrgem seos?
4. Milliste indeksite vahel on negatiivne seos?
5. Milline indeks on teistega kõige nõrgemini seotud?
6. Millise indeksiga on OMXS kõige tugevamini seotud?

Näpunäide: programmis Excel kasutada korrelatsioonimaatriksi saamiseks vahendit *Correlation* komplektist *Data Analysis*. VASTUS lk 680.

A.8.3. Näites 8.2 analüüsiti mõningate majandussuuruste vahelisi seoseid Eesti maakondades:

- töötuse määr, % kogu tööhõivelisest elanikkonnast;
- keskmine brutokuupalk, eurot;
- sündinud ettevõtete arv aastas;
- hõivatute osatähtsus sekundaarsektoris (töötlev tööstus), %;
- hõivatute osatähtsus tertsiaalsektoris (teenindav sektor), %.

Aanalüüsist oli välja jäetud Harjumaa. Nüüd on tabelisse lisatud ka Harjumaa andmed. Leida korrelatsioonimaatriks ning võrrelda korrelatsioonikordajaid nendega, mis olid leitud ilma Harjumaata. VASTUS lk 680.

Autokorrelatsioon

A.8.4. Tabelis on kolme OMXT börsil noteeritud aktsia päevased tulumäärad 18.12.2014 –18.12.2015. Leida kõigi jaoks autokorrelatsiooni kordaja. Millise aktsia tulumäära muutumine on kõige juhuslikum? VASTUS lk 680.

Korrelatsiooni olulisus

A.8.5. Kasutades andmeid 2009. aasta kohalike omavalitsuse valimiste kohta, analüüsida, kas erakondade kulud valimiskampaaniale ja saadud häälte arv on omavahel korrelatsioonis. Otsustamiseks kasutada nii nivood 5% kui ka nivood 1%. VASTUS lk 680.

A.8.6. Mõnikord kasutatakse turunduses trikki, kus reklaamis näidatakse mingi toote jaoks mitte ühe tüki hinda, vaid hinda kahe, kolme

või nelja toote ostmisel. Näiteks olgu toote müügihind 2,5 €/tk. Kui reklaamis näidata, et 2 tk saab osta viie euro eest või kolm tükki 7,5 euro eest, siis paljud ostjad arvavad, et ostes mitu tükki korraga, teevad nad soodsama tehingu.

Üks pood otsustas seda teooriat kontrollida. Teatud kaupa reklaamiti viie nädala jooksul erinevate koguste maksumustega. Ühel nädalal oli reklaamis ühe tüki hind, teise nädala jooksul näidati reklaamis kahe tüki maksumust, kolmandal nädalal oli reklaamis maksumus kolme tüki korraga ostmisel jne. Igal nädalal registreeriti vastava toote müügimaht nädalas, andmed on toodud tabelis. Kas see eksperiment kinnitab teooriat, et müügimaht suureneb, kui reklaamis näidata summat mitme tüki korraga ostmisel? Vastus lk 680.

Kriitiline korrelatsioonikordaja väärtus

A.8.7. Tabelis on toodud andmed 39 Eesti teenindusettevõtte kohta 2013. aastal. Andmed on võetud Eesti ettevõtete konkurentsivõime edetabelist 2013[2]. Leida

a) korrelatsioonimaatriks;
b) ettevõtete arvule (valimi mahule) vastav korrelatsioonikordaja kriitiline väärtus olulisuse nivool 0,05;
c) milliste suuruste vahel on statistiliselt oluline korrelatsioon.

Vastus lk 680.

Spearmani korrelatsioonikordaja

A.8.8. Bakalaureusetöö „Töötajate motiveerimine väikeettevõttes“ raames läbiviidud uuringus paluti muu hulgas Eesti, Läti ja Soome väikeettevõtete töötajatel järjestada olulisuse järgi erinevad vajadused (Kala, 2004). Vajadused, mida paluti järjestada, olid järgmised:

T turvalisuse vajadus (ohutud töötingimused, tööga kindlustatus, soodustused);
S sotsiaalne vajadus (inimestevaheline sõprus, lähedus, suhtlemine);
F füsioloogiline vajadus (hea töötasu, mugavad töötingimused);
H hinnangu (lugupidamise) vajadus;
E eneseteostuse vajadus (annete ja võimete täielik realiseerimine).

Kasutades Spearmani korrelatsioonikordajat, leida, milliste riikide korral langeb töötajate vajaduste järjestus rohkem kokku: kas Eesti ja Läti või Eesti ja Soome. Vastus lk 680.

A.8.9. Kas keskmine tootlikkus ja keskmine tunnipalk Eesti maakondades on omavahel seotud? Tabelis on andmed 15 Eesti maakonna kohta 2013. aastal[3]:

[2] Allikas: Eesti konkurentsivõimelisim teenindusettevõte 2013. `http://ettevotluskonkurss.ee/edetabelid-2013/`

[3] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabelid EM027: ettevõtete (20 ja enama hõivatuga) lisandväärtus ja tootlikkusnäitajad maakonna järgi ja PA22: keskmine brutotunnipalk maakonna järgi (kvartalid).

tunnitootlikkus müügitulu alusel, eurot;
keskmine brutotunnipalk, eurot.

1. Leida lineaarne korrelatsioonikordaja. Kas seos on statistiliselt oluline?
2. Leida Spearmani korrelatsioonikordaja.
3. Analüüsida, miks lineaarne korrelatsioonikordaja on oluliselt suurem. Näpunäide: uurida hajumisdiagrammi.
4. Leida lineaarne korrelatsioonikordaja ilma erindita. Kas seos on oluline?

VASTUS lk 680.

Peatükk 9

Regressioonanalüüs

Eelmises peatükis vaatlesime korrelatsioonanalüüsi kui vahendit tunnustevahelise seose hindamiseks. Korrelatsioonikordaja võimaldab arvuliselt kirjeldada seose tugevust ning vastav t-test võimaldab hinnata, kas seos on statistiliselt oluline.

Kui kahe suuruse vahel on olemas seos, siis järgmiseks eesmärgiks on selle seose matemaatiline modelleerimine. Järgnevalt vaatamegi, kuidas empiiriliste andmete põhjal leida arvutusteks vajalikku matemaatilist mudelit.

9.1. Matemaatiline mudel, selle üldkuju ja konkreetne kuju

Nähtuste ja protsesside analüüsimisel kasutatakse tihti mitmesuguseid mudeleid.

Mudel on reaalses maailmas esineva objekti analoog, mis asendab seda objekti tunnetusprotsessis.

Mudel

Mudel peab
- välja tooma olulise;
- kõrvale jätma mitteolulise.

Mudel võib olla füüsiline (näiteks mudelauto) või märkmudel. Märkmudeliteks on näiteks skeem, graafik, tabel, matemaatiline mudel.

Matemaatiline mudel on mingit reaalses maailmas eksisteerivat nähtust kirjeldavate matemaatiliste seoste kogum.

Matemaatiline mudel

Üldjuhul pannakse matemaatiline mudel kirja kujul

$$y = f(x_1, x_2, \ldots, a, b, \ldots), \tag{9.1}$$

kus eristatakse järgmisi mudeli **komponente**:

- sõltuv muutuja y ehk funktsioon;
- sõltumatud muutujad $x_1, x_2, \ldots$ ehk argumendid;
- mudeli parameetrid $a, b, \ldots$ ehk konstandid.

Kuju (9.1) on väga üldine. Konkreetse nähtuse analüüsimisel tuleb täpsustada, milliseid matemaatilisi seoseid me modelleerimisel kasutame. Kõige lihtsam matemaatiline mudel on lineaarne mudel. Selle kuju on kõige lihtsam ja lihtne on ka parameetrite tõlgendus.

Lineaarne mudel

Lineaarse mudeli üldkuju on

$$y = ax + b, \tag{9.2}$$

kus parameeter a on lineaarliikme kordaja ja parameeter b vabaliige.

Funktsiooni (9.2) graafikuks on sirge xy-tasandil. Lineaarse mudeli parameetrite tõlgendused on järgmised (vt ka lisa A.8):

Lineaarse mudeli parameetrite tõlgendus

- lineaarliikme kordaja a näitab, kui palju muutub y, kui x suureneb 1 võrra;
- vabaliige b näitab sõltuva muutuja y väärtust, kui $x = 0$.

Valem (9.2) on lineaarse mudeli üldkuju. Selle põhjal ei saa me teha veel mingeid arvutusi ega konstrueerida graafikut. Kui me anname parameetritele mingid väärtused, siis valime kõikvõimalike lineaarsete mudelite hulgast välja ühe konkreetse. Võtame näiteks

$$a = 2, \quad b = 3.$$

Kui me paneme need arvud mudeli üldkujusse (9.2), saame mudeli konkreetse kuju:

$$y = 2x + 3. \tag{9.3}$$

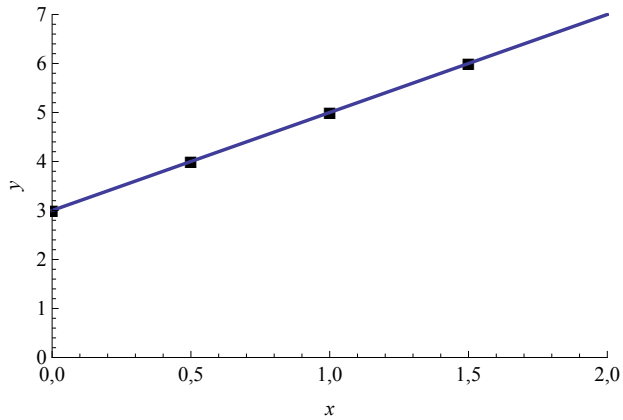
Selle põhjal võime leida valitud x väärtustele vastavad y väärtused (tabel 9.1) ning konstrueerida graafiku (joonis 9.1).

Mudeli üldkuju ja konkreetne kuju

Mudeli kuju, kus parameetrite arvväärtused puuduvad, nimetatakse mudeli **üldkujuks**. Kui parameetritele on omistatud konkreetsed arvväärtused, siis on meil olemas mudeli **konkreetne kuju**.

Tabel 9.1. Arvutused mudeli (9.3) järgi

i	x_i	y_i
1	0	3
2	0,5	4
3	1	5
4	1,5	6



Joonis 9.1. Lineaarse mudeli $y = 2x + 3$ graafik, millele on lisatud tabelis 9.1 arvutatud punktid

Tabel 9.2. Mudeli üldkuju ja konkreetne kuju

Parameetrid määramata		Parameetrid määratud	
Mudeli üldkuju	Parameetrid	Parameetrid	Mudeli konkreetne kuju
$y = ax + b$	a, b	$a = 2,\ b = 3$	$y = 2x + 3$
$y = ax^2 + bx + c$	a, b, c	$a = 2,\ b = 4,\ c = -2$	$y = 2x^2 + 4x - 2$
$y = y_0 e^{rx}$	y_0, r	$y_0 = 4{,}5,\ r = 0{,}1$	$y = 4{,}5e^{0{,}1x}$

Mõned näited erinevate mudelite üldkuju ja konkreetse kuju kohta on tabelis 9.2.

Et mudel kirjeldaks reaalses elus eksisteerivat nähtust ja annaks meile selle kohta teavet, tuleb määrata mudeli matemaatiline kuju ning seejärel vaatlusandmete põhjal mudeli parameetrid. Sellega tegelebki regressioonanalüüs.

9.2. Regressioonmudel

Eelmise alapeatüki tabelis 9.1 arvutasime mudeli (9.3) järgi. Teades x_i väärtust, võisime y_i väärtuse täpselt välja arvutada. Sellisel juhul on Y väärtus täpselt määratud ehk **determineeritud**. Reaalses elus esinevate nähtuste modelleerimisel puutume aga kokku juhuslikkusega.

Tuletame meelde, kuidas juhuslikkus mõjub kogumi keskväärtuse leidmisel. Olgu Eesti meeste keskmine pikkus 179 cm. Kui Jaan on 175 cm pikkune, siis tema erinevus keskmisest (-4 cm) on tingitud paljudest juhuslikest teguritest ja seda võime nimetada pikkuse **juhuslikuks komponendiks**. Kui Mati pikkus on 182 cm, siis tema korral on pikkuse juhuslik komponent 3 cm. Suvalise Eesti mehe pikkuse võime kirja panna nii:

$$\text{PIKKUS} = 179 + \varepsilon,$$

kus PIKKUS on sentimeetrites ning ε on juhuslik komponent.

Kui me soovime aga leida poisslapse pikkust, siis see sõltub vanusest ja juhuslikust komponendist. Näiteks 2–16-aastase poisslapse pikkus[1]

$$\text{PIKKUS} = 80{,}4 + 6 \cdot \text{VANUS} + \varepsilon. \tag{9.4}$$

Selle mudeli järgi on 8-aastaste poiste keskmine pikkus $80{,}4 + 6 \cdot 8 = 128{,}4$ sentimeetrit. Konkreetse 8-aastase naabripoisi pikkus on $128{,}4 + \varepsilon$ sentimeetrit.

Regressioonmudel

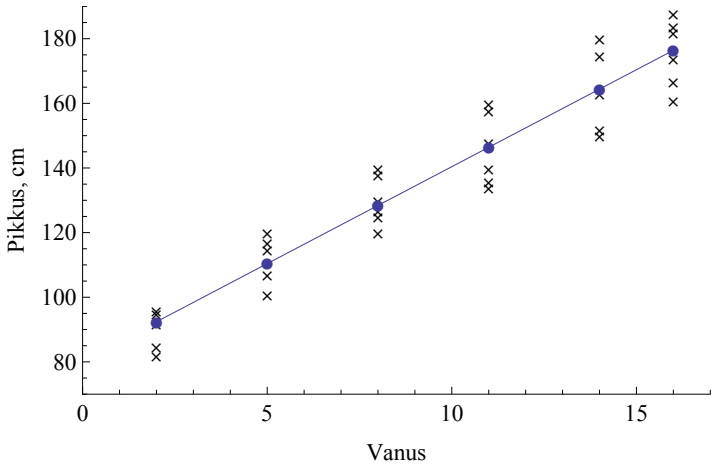
Regressioonmudel koosneb deterministlikust ja juhuslikust komponendist:

$$Y = \text{deterministlik komponent} + \text{juhuslik komponent}, \tag{9.5}$$

kus Y on modelleeritav tunnus.

Deterministlik komponent on see oluline osa, mille mudel peab välja tooma. Selle järgi saame analüüsida reaalse maailma nähtust, mida regressioonmudel kirjeldab. Juhuslik komponent on mitteoluline osa, mis antud uurimisetapil meid ei huvita.

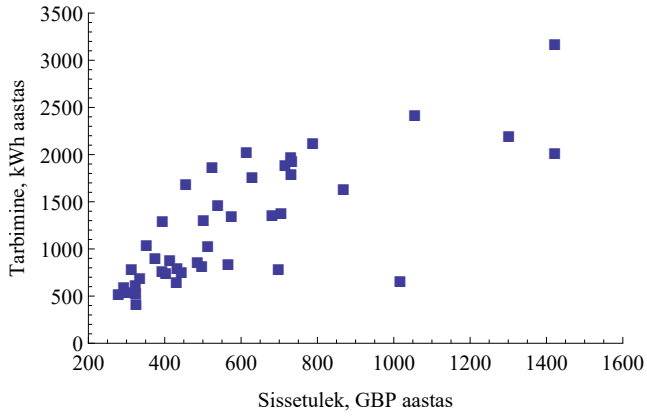
[1] Allikas: `http://www.hopsti.ee`



Joonis 9.2. 2–16-aastaste poiste pikkuse sõltuvus vanusest. Ristikestega on märgitud erinevas vanuses konkreetsete poiste pikkused. Sirgel olevad ringid tähistavad vastavas vanuses poiste keskmist pikkust, mis sõltub vanusest

Näide 9.1. Sissetulek ja elektrienergia tarbimine

H. Houthakker (Houthakker, 1951) analüüsis elektrienergia tarbimist Suurbritannia erinevates linnades 1930. aastate lõpus. Eesmärgiks oli kindlaks teha, kas majapidamiste sissetulek mõjutab elektri tarbimist. Andmed pärinesid 48 linnast. Igas linnas leiti majapidamise keskmine sissetulek (GBP aastas) ja keskmine elektrienergia tarbimine (kWh aastas). Joonisel 9.3 on esitatud vastav hajumisdiagramm, millelt näeme, et sissetuleku ja elektrienergia tarbimise vahel on positiivne seos. Lineaarne korrelatsioonikordaja on $r = 0{,}767$.



Joonis 9.3. Majapidamiste sissetulek ja elektrienergia tarbimine

Esitame nüüd küsimuse, kui suur on elektrienergia tarbimine, kui majapidamise sissetulek aastas on 1200 GBP? Selle leidmiseks on vaja matemaatilist mudelit.

Vaatame joonist 9.3, kus tunnuseks X on majapidamiste sissetulek ja tunnuseks Y elektrienergia tarbimine. On näha, et vaatluspunktid ei asu ühe konkreetsel sirgel nii nagu näiteks joonisel 9.1. Arvestades aga seost (9.5), tekib idee, et modelleerida saame deterministlikku komponenti. Kõrvalekalded sirgest on seletatavad juhusliku komponendiga.

Regressioonanalüüs

Regressioonanalüüs uurib suurustevahelist sõltuvust ja võimalusi selle funktsionaalseks kirjeldamiseks etteantud valemi põhjal. Regressioonanalüüsi käigus leitakse regressioonmudeli deterministlik komponent.

Regressioonanalüüsi puhul eeldame alati, et **juhusliku komponendi keskväärtus on 0**. Kui see ei ole 0, siis eksisteerib järelikult mingi süstemaatiline, mittejuhuslik kõrvalekalle, ja meil tuleb deterministlikku komponenti korrigeerida. Näiteks kui me leiame kõigi Eesti meeste pikkuste erinevused keskmisest pikkusest 179 cm ja nende erinevuste keskmine ei ole null, siis järelikult Eesti meeste keskmine pikkus ei ole 179 cm, vaid mingi muu väärtus.

Kui juhusliku komponendi keskväärtus on 0, siis seosest (9.5) järeldub, et Y **tinglik keskväärtus**

$$E[Y|X] = \text{deterministlik komponent}. \tag{9.6}$$

Tinglik keskväärtus tähendab, et keskväärtus ei ole konstantne, vaid sõltub argumenttunnustest. Joonisel 9.2 esitatud sirge on leitud mudeli (9.4) deterministliku komponendi põhjal ja kirjeldab poisslaste keskmise pikkuse sõltuvust vanusest:

$$E(\text{PIKKUS}|\text{VANUS}) = 80{,}4 + 6 \cdot \text{VANUS}.$$

Regressioonanalüüsi käigus leitakse, kuidas funktsioontunnuse Y tinglik keskväärtus sõltub seda mõjutavatest argumenttunnustest.

Termini „regressioon“ võttis esmakordselt kasutusele Francis Galton[2], kes pani tähele, et pikkadel vanematel on lühemad lapsed kui

[2] Francis Galton (1822–1911), inglise antropoloog.

nad ise ja lühematel vanematel pikemad lapsed, s.t laste pikkus suundub keskmise pikkuse poole. Ta nimetas seda regressiooniks keskmise taseme poole (*regression toward the mean*).

Oluline on, et ka regressioonanalüüsi läbiviimiseks tuleb **valem** ehk funktsiooni matemaatiline kuju **ette anda**. See võib olla lineaarne või mittelineaarne: ruutfunktsioon, eksponentsiaalne funktsioon, astmefunktsioon või mingi keerulisema kujuga funktsioon. Regressioonanalüüsi käigus leitakse eelnevalt kogutud empiiriliste andmete alusel regressioonmudeli deterministliku osa parameetrite arvväärtused, s.t leitakse andmestikule vastav mudeli konkreetne kuju. Lisaks tehakse mitmesuguseid teste, mis võimaldavad hinnata, kas väljavalitud tunnuste vahelist seost saab üldse etteantud valemiga kirjeldada.

Statistilise kogumi elementidel võib olla palju erinevaid tunnuseid. Regressioonanalüüsiks valitakse välja teatud tunnuste komplekt, kus üks tunnus on funktsioontunnus Y ja ülejäänud argumenttunnused $X_1, X_2, \ldots, X_k$, mida nimetatakse ka **regressoriteks**. Tunnuste komplekti valik eeldab seda, et meil peab olema mingi ettekujutus uuritavast nähtusest: millised tunnused on omavahel seotud, mis millest sõltub. Võib ka juhtuda, et kõik esialgu väljavalitud regressorid ei mõjutagi funktsioontunnust, see selgub regressioonanalüüsi käigus.

Kui tunnuste komplekt on välja valitud, viiakse läbi valimvaatlus, mille käigus mõõdetakse kõigi väljavalitud tunnuste väärtused valimisse kuuluvatel objektidel. Kui kasutada objektide eristamiseks indeksit i, kusjuures $i = 1, 2, \ldots, n$, siis iga objekti jaoks saadakse väärtuste komplekt $y_i, x_{1i}, x_{2i}, \ldots, x_{ki}$. Siin on k argumenttunnuste arv ja n valimi maht.

Regressioonmudel võib sisaldada kas üht või mitut argumenttunnust. Näiteks algul võime uurida elektrienergia tarbimise mudelit, kus deterministlik komponent sõltub ainult majapidamise sissetulekust. Hiljem võime muuta mudeli deterministlikku komponenti keerukamaks ja analüüsida ka seda, kuidas elektri tarbimine sõltub elektrienergia hinnast, majapidamises olevate elektriseadmete koguvõimsusest jne. Vastavalt sellele jagatakse regressioonanalüüs kaheks:

Lihtne ja mitmene regressioon

- üks argumenttunnus, $y = f(x)$ — **lihtne regressioon** (*simple regression*);
- mitu argumenttunnust, $y = f(x_1, x_2, \ldots, x_k)$ — **mitmene regressioon** (*multiple regression*).

9.3. Vähimruutude meetod

Kuna kõige lihtsam matemaatiline mudel on lineaarne mudel, siis vastavat regressioonmudelit kasutatakse modelleerimisel kõige sagedamini.

Lineaarne regressioonmudel

Lineaarse regressioonmudeli üldkuju on

$$y = \alpha x + \beta + \varepsilon, \tag{9.7}$$

kus α ja β on mudeli parameetrid ning ε juhuslik liige.

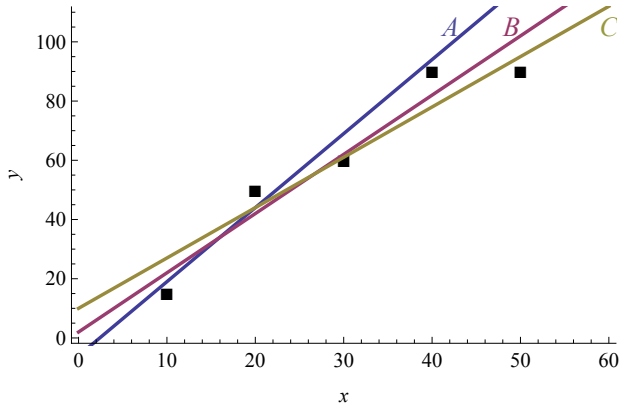
Lineaarne regressioonülesanne

Kui me leiame regressioonanalüüsi käigus parameetrite α ja β arväärtused, siis ongi leitud mudeli (9.7) deterministlik komponent. Kuna funktsiooni $\alpha x + \beta$ graafikuks on sirge, siis otsime kõikvõimalike sirgete hulgast üht konkreetset sirget, mille tõus on α ja algordinaat β. Lineaarse mudeli parameetrite leidmist regressioonanalüüsi abil nimetatakse **lineaarseks regressioonülesandeks**.

Olgu meil mudeliga (9.7) esitatud seose genereeritud 5 erinevat punkti. Punktide koordinaadid xy-tasandil on toodud tabelis 9.3 ja punktid on esitatud joonisel 9.4.

Tabel 9.3. Joonisel 9.4 asuvate punktide koordinaadid

i	1	2	3	4	5
x_i	10	20	30	40	50
y_i	15	50	60	90	90



Joonis 9.4. Milline sirge kirjeldab punktiparve kõige paremini? Sirgete võrrandid on (9.8)

Lineaarne regressioonülesanne seisneb selles, et läbi punktiparve tuleb tõmmata sirge, mis kirjeldab seda punktiparve võimalikult hästi. Aga mis tähendab „võimalikult hästi“? Samu andmeid kasutades võivad erinevad inimesed saada erinevad sirged. Joonisel 9.4 on kolm sirget:

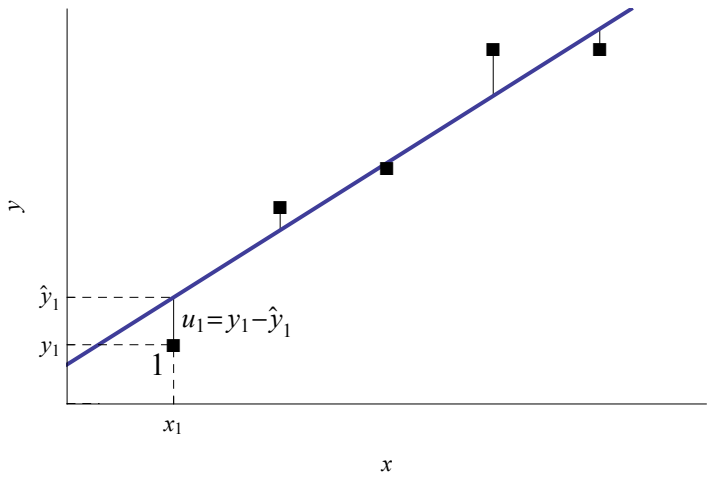
A, B ja C, mille võrrandid on vastavalt

$$\begin{aligned} y_A &= 2{,}5x - 6, \\ y_B &= 2x + 2, \\ y_C &= 1{,}7x + 10. \end{aligned} \tag{9.8}$$

Milline sirge kirjeldab seda punktiparve kõige paremini?

Selleks, et otsustada, milline sirge on kõige parem, on vaja **objektiivset kriteeriumi**. Üks võimalus selle defineerimiseks on, et punktiparve kirjeldab hästi selline sirge, mille korral üksikute punktide hälbed sirgest (joonis 9.5) on võimalikult väikesed, kõik punktid on sirgele nii lähedal kui võimalik. Hälbeid saab arvutada punkti koordinaatide ja sirge võrrandi kaudu. Punkti 1 hälve on selle punkti y-koordinaadi y_1 ja sellele punktile vastava sirgel asuva punkti y-koordinaadi $\hat{y}_1$ vahe:

$$u_1 = y_1 - \hat{y}_1. \tag{9.9}$$



Joonis 9.5. Vertikaalsed kriipsukesed kujutavad punktide hälbeid sirgest. Näidatud on punkti 1 hälbe u_1 arvutamine

Arvestame, et kui mingi punkt asub sirgel $y = ax + b$, siis selle punkti y-koordinaat $\hat{y}_i = ax_i + b$. Valemist (9.9) saame punkti 1 hälbe arvutamiseks valemi

$$u_1 = y_1 - (ax_1 + b). \tag{9.10}$$

Siin kasutame sirge parameetrite tähistamiseks tähti a ja b, mis on mudeli (9.7) parameetrite α ja β hinnangud.

Hälbeid on sama palju kui punkte, n tükki. Kuidas me saame neist ühe arvu, mis oleks kriteeriumiks sirgete võrdlemisel? Hälbeid kokku

liita pole mõtet, sest positiivsed ja negatiivsed hälbed kustutavad üksteist. Kokku võib liita hälvete ruudud. Hälvete ruutude summa järgi saame otsustada, milline sirge on parim. Seda meetodit nimetatakse vähimruutude meetodiks.

Vähimruutude meetod

Vähimruutude meetodi korral minimeeritakse üksikute punktide hälvete u_i ruutude summat:

$$\sum u_i^2 \to \min. \tag{9.11}$$

Nüüd on meil olemas objektiivne karakteristik erinevate sirgete võrdlemiseks ja me saame otsustada, milline sirge on parim. Joonisel 9.4 esitatud punktiparve ja sirgete analüüsimisel saame järgmised summad (punktide koordinaadid (x_i, y_i) on võetud tabelist 9.3 ja sirgete võrrandid on (9.8)):

sirge A korral $\sum u_i^2 = \sum (y_i - (2{,}5x_i - 6))^2 = 990$;
sirge B korral $\sum u_i^2 = \sum (y_i - (2x_i + 2))^2 = 325$;
sirge C korral $\sum u_i^2 = \sum (y_i - (1{,}7x_i + 10))^2 = 350$.

Näeme, et hälvete ruutude summa on kõige väiksem sirge B korral. Järelikult, nende kolme sirge hulgast kirjeldab seda punktiparve kõige paremini sirge B.

Aga kuidas me teame, et sirge B on kõige parem sirge? Äkki on olemas mõni sirge, mille korral hälvete ruutude summa tuleb veel väiksem? Me peaksime võrdlema kõikvõimalikke sirgeid, kuid neid on lõpmata palju. Appi tuleb matemaatikast tuntud funktsiooni minimeerimise reegel: funktsiooni miinimumkoha leidmiseks tuleb funktsiooni tuletis panna võrduma nulliga ja lahendada saadud võrrand. Hälvete ruutude summas kirjutame iga punkti hälbe välja nii nagu valemis (9.10):

$$u_i = y_i - \hat{y}_i = y_i - (ax_i + b). \tag{9.12}$$

Kui parameetrid a ja b on tundmatud, siis võime hälvet (9.12) vaadelda funktsioonina kahest muutujast a ja b. Samamoodi on ka hälvete ruutude summa kahe muutuja funktsioon ning vastavat funktsiooni tähistame $S(a, b) = \sum u_i^2$. Siis võime vähimruutude meetodi tingimuse (9.11) kirjutada kujul

$$S(a,b) = \sum (y_i - (ax_i + b))^2 \to \min. \tag{9.13}$$

Kuna $S(a, b)$ on kahe muutuja funktsioon, siis tuleb miinimumkoha leidmiseks leida osatuletised ja võrdsustada need nulliga. Saame võrrandisüsteemi

$$\begin{cases} \dfrac{\partial S(a,b)}{\partial a} = 0 \\ \dfrac{\partial S(a,b)}{\partial b} = 0. \end{cases} \tag{9.14}$$

Selle võrrandisüsteemi lahendamine annab meile valemid lineaarse mudeli parameetrite leidmiseks (vt lisa A.9).

Lineaarse regressioonmudeli parameetrid

Vähimruutude meetodil leitud valemid lineaarse regressioonmudeli $y = \alpha x + \beta + \varepsilon$ parameetrite hinnangute jaoks:

$$a = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \tag{9.15}$$

$$b = \bar{y} - a\bar{x}, \tag{9.16}$$

kus a on parameetri α hinnang ja b parameetri β hinnang.

Valemites (9.15) ja (9.16) on x_i ja y_i tunnuste X ja Y väärtused erinevate objektide korral ning $\bar{x}$ ja $\bar{y}$ aritmeetilised keskmised. Sirge, mille parameetrid on arvutatud nende valemitega, on parim vastavat punktiparve läbiv sirge: selle sirge korral on hälvete ruutude summa $\sum u_i^2$ kõige väiksem.

Valemite (9.15) ja (9.16) järgi arvutamine on kalkulaatoriga üpris töömahukas, eriti kui andmeid on palju. On küll olemas ka selliseid kalkulaatoreid, millel on kahemõõtmelise statistika funktsioonid, sealhulgas ka need valemid. Enamasti viiakse aga regressioonanalüüs läbi kas tabelarvutuses või siis mingis spetsiaalses statistikapaketis, kus arvutused valemite (9.15) ja (9.16) järgi tehakse automaatselt.

Tabelarvutuses on parameetri a leidmiseks funktsioon **SLOPE** (eesti keeles tõus) ja parameetri b leidmiseks funktsioon **INTERCEPT** (eesti keeles vabaliige). Need funktsioonid arvutavad vastavalt valemite (9.15) ja (9.16) järgi.

Arvutame regressioonsirge parameetrid tabelis 9.3 toodud andmete korral. Selleks leiame kõigepealt aritmeetilised keskmised:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{10 + 20 + 30 + 40 + 50}{5} = \frac{150}{5} = 30,$$
$$\bar{y} = \frac{\sum y_i}{n} = \frac{15 + 50 + 60 + 90 + 90}{5} = \frac{305}{5} = 61$$

ning siis vahed $x_i - \bar{x}$ ja $y_i - \bar{y}$. Tulemused on esitatud tabelis 9.4.

Tabeli 9.4 viienda veeru summeerimisel saame $\sum (x_i - \bar{x})^2 = 1000$ ja viimase veeru summa on $\sum (x_i - \bar{x})(y_i - \bar{y}) = 1900$. Valemist (9.15) saame

$$a = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{1900}{1000} = 1{,}9$$

ja valemi (9.16) järgi arvutades

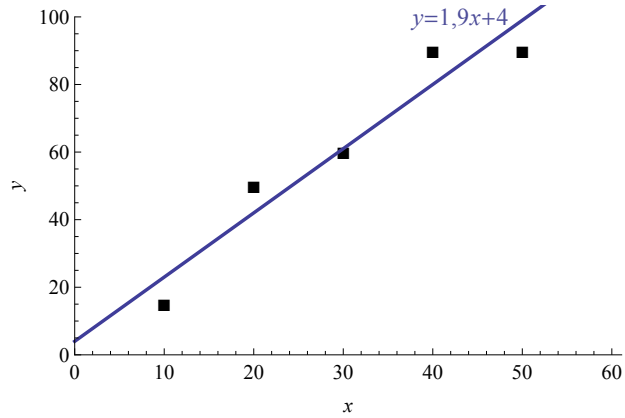
$$b = \bar{y} - a\bar{x} = 61 - 1{,}9 \cdot 30 = 4.$$

Tabel 9.4. Arvutused regressioonsirge parameetrite leidmiseks

i	x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	10	15	-20	400	-46	920
2	20	50	-10	100	-11	110
3	30	60	0	0	-1	0
4	40	90	10	100	29	290
5	50	90	20	400	29	580

Samad tulemused saame tabelarvutuses funktsioone SLOPE ja INTERCEPT kasutades. Regressioonmudeli deterministlik komponent on järelikult (vt ka joonis 9.6)

$$\hat{y} = 1{,}9x + 4. \tag{9.17}$$



Joonis 9.6. Tabelis 9.3 esitatud andmetele vastav regressioonsirge, mis on leitud vähimruutude meetodil

Leiame hälvete ruutude summa selle sirge korral, milleks arvutame iga punkti jaoks sirgel asuva vastava punkti koordinaadi $\hat{y}_i = 1{,}9x_i + 4$, seejärel hälbe $u_i = y_i - \hat{y}_i$ ja selle ruudu (tabel 9.5). Sirge (9.17) korral on hälvete ruutude summa

$$\sum u_i^2 = \sum (y_i - \hat{y}_i)^2 = 310, \tag{9.18}$$

mis on väiksem kui sirge B korral leitud summa 325. Võime olla kindlad, et sirget, mille korral hälvete ruutude summa oleks väiksem, selle punktiparve korral leida ei õnnestu.

Tabel 9.5. Arvutused hälvete ruutude summa leidmiseks

i	x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	10	15	23	−8	64
2	20	50	42	8	64
3	30	60	61	−1	1
4	40	90	80	10	100
5	50	90	99	−9	81

Näide 9.2. Sissetulek ja elektrienergia kasutamine: lineaarne mudel

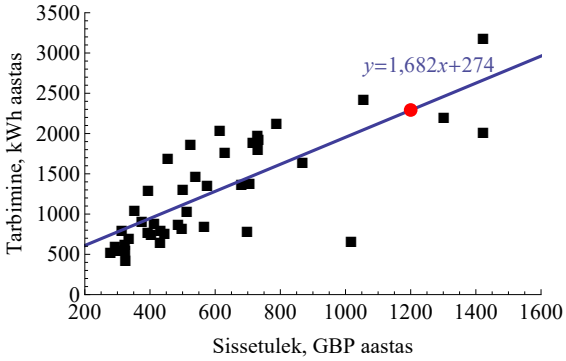
N09Regressioon
N9.2

Näites 9.1 oli toodud hajumisdiagramm, mis kirjeldas majapidamiste sissetuleku ja elektrienergia tarbimise vahelist seost (joonis 9.3). Lõpus oli esitatud küsimus: kui suur on elektrienergia tarbimine, kui majapidamise sissetulek on aastas 1200 GBP? Kuna analüüsi autori H. Houthakkeri kasutatud andmed on saadaval, siis saame läbi viia regressioonanalüüsi. Otsime mudelit kujul

$$y = \alpha x + \beta + \varepsilon, \tag{9.19}$$

kus y on elektrienergia tarbimine, x sissetulek ning ε juhuslik liige. Regressioonanalüüsi viime läbi tabelarvutuses. Mudeli parameetrite hinnanguteks saame

$$a = 1{,}682, \;\; b = 274.$$



Joonis 9.7. Sissetuleku ja elektrienergia tarbimise vaheline seos, sirgel asuv punane punkt vastab tarbimise prognoosile sissetuleku 1200 GBP korral

Järelikult, seost kirjeldav regressioonmudel on

$$y = 1{,}682x + 274 + \varepsilon, \tag{9.20}$$

kus y on elektrienergia tarbimine (kWh aastas) ja x sissetulek (GBP aastas). Võib kirja panna ka ainult mudeli deterministliku komponendi

$$\hat{y} = 1{,}682x + 274. \tag{9.21}$$

Mudeli põhjal saame väita, et kui sissetulek kasvab 1 GBP võrra aastas, siis suureneb elektrienergia tarbimine aastas 1,682 kWh võrra. Kui majapidamise sissetulek on 1200 GBP, siis $x = 1200$. Arvutus mudeli (9.21) põhjal:

$$\hat{y}(1200) = 1{,}682 \cdot 1200 + 274 = 2293. \tag{9.22}$$

Järelikult, aastasissetuleku 1200 GBP korral on prognoositav elektrienergia tarbimine 2293 kWh aastas. See on elektrienergia tarbimise mudeli deterministlik komponent. Sellele lisandub juhuslik komponent ning tegelik tarbimine on

$$y(1200) = 2293 + \varepsilon.$$

Kui regressioonmudeli parameetrid on leitud, saab mudelit kasutada funktsioontunnuse Y väärtuste prognoosimiseks argumenttunnuse X erinevate väärtuste korral. Funktsioontunnuse tegelik väärtus võib olla mudeli abil arvutatud väärtusest erinev, sest lisaks mudelis olevale argumenttunnusele X mõjub tunnusele Y veel muid tegureid, mida mudelis pole arvestatud. Mudeli põhjal arvutatud väärtust $\hat{y}$ võib nimetada ka mudelväärtuseks. Tegelik väärtus y erineb mudelväärtusest juhusliku komponendi ε võrra, mida me ei tea:

$$y = \hat{y} + \varepsilon.$$

Mudelväärtus

Argumenttunnuse väärtusele x_A vastav lineaarse regressioonmudeli deterministliku osa põhjal leitud funktsioontunnuse väärtus on **mudelväärtus**

$$\hat{y}_A = ax_A + b. \tag{9.23}$$

Tabelarvutuses võib mudelväärtuse $\hat{y}$ leidmiseks lineaarse mudeli korral kasutada funktsiooni **TREND**. Selleks tuleb ette anda Y väärtused (*Known_ y's*), X väärtused (*Known_ x's*) ja see X väärtus, mille jaoks me tahame leida Y mudelväärtust (*New_ x's*). Funktsioon leiab lineaarse mudeli parameetrid ja siis kasutab neid mudelväärtuse arvutamisel.

Näide 9.3. Tarbimismudelid

N09Regressioon
N9.3

Saamaks infot leibkonna kulutuste ja tarbimise kohta viib Eesti Statistikaamet läbi leibkonna eelarve uuringuid. Nende alusel on võimalik analüüsida tarbimiskulusid ning toodete ja teenuste pakkujad saavad infot selle kohta, millele tarbijad rohkem kulutavad.
Erinevate toodete ja teenusete tarbimiseks tehakse erinevaid kulutusi. Kuidas on leibkonna kulutused toidule, sidele ja transpordile määratud leibkonna kogukuludega? Vastavate seoste leidmiseks kasutame 2012. aasta andmeid Eesti Statistikaameti andmebaasist tabelist „LE208: Leibkonnaliikme kulutused aastas kuludetsiili järgi“. Tabelis on kulutused pereliikme kohta aastas, eurodes.

Detsiil-vahemik	Kokku	Toit	Side	Transport
1	992,2	342,6	103,2	
2	1585,7	554,1	141,2	52,9
3	2042,7	664,5	151,8	120,1
4	2449,2	763,6	167,7	199,6
5	2863,1	867,2	188,0	220,3
6	3336,7	1002,9	192,5	262,8
7	3892,5	1038,8	218,1	430,4
8	4623,6	1174,4	232,6	509,0
9	5847,6	1244,5	246,5	963,7
10	9865,8	1555,0	365,2	2070,1

Andmed on toodud erinevatesse detsiilvahemikesse kuuluvate perede jaoks. Esimesse detsiilvahemikku kuuluvad need 10% peredest, kelle kogukulud on kõige väiksemad, viimasesse need 10%, kelle kogukulud on kõige suuremad.
Mudeleid otsime kujul

$$y = b + ax + \varepsilon, \tag{9.24}$$

kus y on kulud mingi teenuse või kaubagrupi tarbimiseks, x kulutused kokku ja ε juhuslik liige. Tabelarvutuses leiame parameetrite hinnangud b ja a iga kululiigi jaoks eraldi, kasutades vastavalt funktsioone INTERCEPT ja SLOPE:

$$\begin{aligned} \text{Toit} \quad & b = 434{,}4 && a = 0{,}1297, \\ \text{Side} \quad & b = 97{,}0 && a = 0{,}0277, \\ \text{Transport} \quad & b = -464{,}1 && a = 0{,}2467. \end{aligned}$$

Pannes parameetrite väärtused mudeli üldkujusse (9.24), saame järgmised mudelid:

$$\text{kulutused toidule} \quad y_T = 434{,}4 + 0{,}1297x + \varepsilon, \tag{9.25}$$

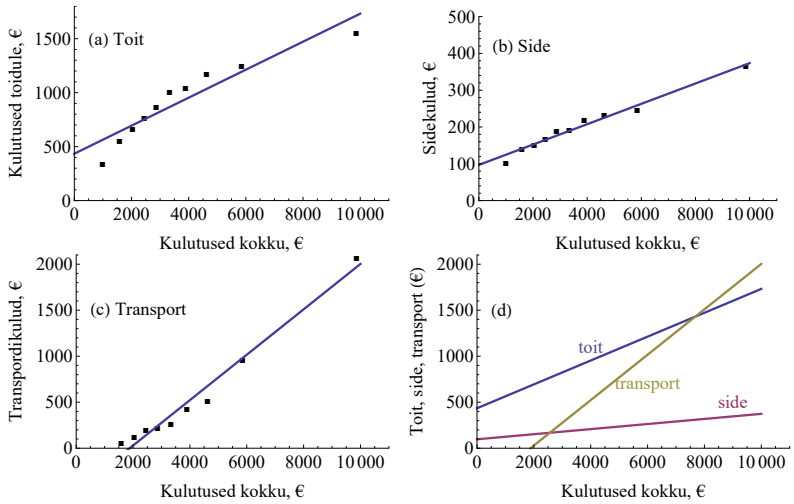
$$\text{kulutused sidele} \quad y_S = 97 + 0{,}0277x + \varepsilon, \tag{9.26}$$

$$\text{kulutused transpordile} \quad y_{TR} = -464{,}1 + 0{,}2467x + \varepsilon. \tag{9.27}$$

Neid mudeleid nimetatakse **tarbimismudeliteks**. Teine võimalus tulemuste esitamiseks on mudelväärtuste kaudu:

$$\hat{y}_T = 434{,}4 + 0{,}1297x,$$
$$\hat{y}_S = 97 + 0{,}0277x,$$
$$\hat{y}_{TR} = -464{,}1 + 0{,}2467x.$$

Vastavad graafikud on esitatud joonisel 9.8.



Joonis 9.8. Tarbimismudelid: (a) toit, (b) side, (c) transport, (d) kolme mudeli võrdlus

Mudelite võrdlemine näitab, et kogukulude suurenedes kasvavad kõige kiiremini kulutused transpordile: selles mudelis on sirge tõus a kõige suurem (vt ka joonis 9.8 (d)). Kui kogukulud x suurenevad 1 euro võrra, siis kulutused transpordile y_{TR} suurenevad 0,2467 euro võrra. Ehk kui kogukulud suurenevad 100 euro võrra, siis kulud transpordile suurenevad 24,67 euro võrra. Kulutused toidule suurenevad aeglasemalt ja kõige aeglasemalt suurenevad kogukulude kasvades sidekulud y_S. See tähendab, et sideteenuste tarbimine sõltub vähe sellest, kas pere saab rohkem või vähem raha kulutada.

Kolme mudeli põhjal võime öelda, et igast lisandunud 100 eurost läheb 24,67 eurot transpordikulude suurendamiseks, 12,97 eurot toidukulude suurendamiseks ja 2,77 eurot sideteenuste tarbimise suurendamiseks. Kui on olemas prognoos majapidamiste sissetulekute muutumise kohta, saame tarbimismudelite abil prognoosida, kui palju raha ühte või teise valdkonda juurde tuleb.
Leiame, kui palju kulub ühe pereliikme toidule, kui kogukulud pereliikme kohta on 2000 eurot aastas. Mudelist (9.25) saame, et

$$\hat{y}_T(2000) = 434{,}4 + 0{,}1297 \cdot 2000 \approx 694$$

eurot aastas. See on ligikaudu 34% kogukuludest. Kui aga kogukulud pereliikme kohta on 6000 eurot aastas, siis kulub toidule

$$\hat{y}_T(6000) = 434{,}4 + 0{,}1297 \cdot 6000 \approx 1213$$

eurot aastas. See on umbes 20% kogukuludest. Järelikult, kogukulude suurenemisel väheneb toidule tehtavate kulutuste osatähtsus.

Nagu nägime, võimaldab regressioonanalüüsi kasutamine saada mitmesugust informatsiooni majanduses valitsevate seoste kohta. Arvutustega saab kiiresti hakkama arvuti, inimesele jääb tulemuste tõlgendamine. Eriti oluline on osata interpreteerida lineaarse mudeli lineaarliikme kordajat (vt ka lisa A.8).

Vaadeldud vähimruutude meetod pole ainuke võimalus regressioonmudeli parameetrite hindamiseks. On ka keerulisemaid hindamismeetodeid, mida kasutatakse siis, kui vähimruutude meetod ei anna usaldusväärseid tulemusi. Mõningad neist:

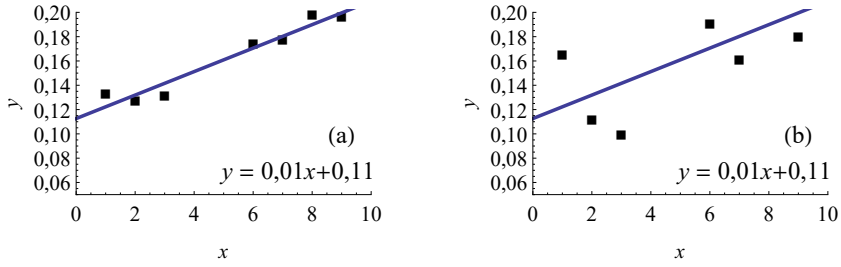
- maksimaalse tõepära meetod MLE (*Maximum Likelihood Estimation*);
- kaalutud vähimruutude meetod WLS (*Weighted Least Squares*);
- üldistatud vähimruutude meetod GLS (*Generalized Least Squares*).

Eristamaks vähimruutude meetodit teistest, nimetatakse seda **harilikuks vähimruutude meetodiks** OLS (*Ordinary Least Squares*).

9.4. Regressioonmudeli kirjeldusvõime ja determinatsioonikordaja

Joonisel 9.9 on toodud kaks erinevat punktiparve ja kummalegi on lisatud lineaarne regressioonjoon. Visuaalse hinnangu järgi võib väita, et joonisel (a) olev regressioonjoon kirjeldab oma punktiparve paremini

kui joonisel (b) olev joon oma punktiparve. Võime ka öelda, et joonisel (a) on mudeli kirjeldusvõime suurem. Regressioonsirge on mõlemal joonisel ühesugune: $\hat{y} = 0{,}01x + 0{,}11$. Et neid situatsioone eristada, tekib vajadus lisada regressioonmudelile selle kirjeldusvõimet iseloomustav suurus. Järgnevalt vaatame, kuidas seda leida.

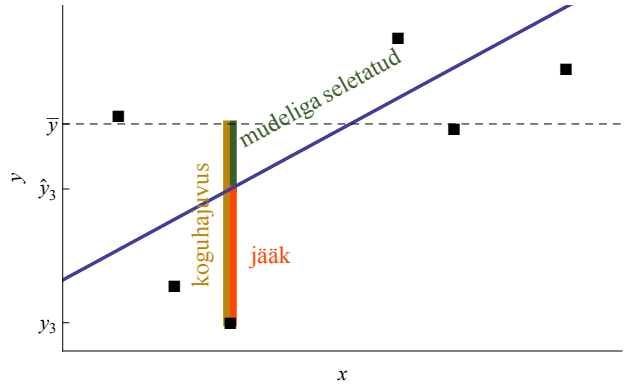


Joonis 9.9. Joonisel (a) olev regressioonsirge kirjeldab oma punktiparve paremini kui joonisel (b) olev sirge oma punktiparve

Olgu meil kaks suurust X ja Y, mille vahel on lineaarse regressioonmudeli abil kirjeldatav statistiline seos. Mudeli leidmiseks on meil kasutada punktipaarid (x_1, y_1), (x_2, y_2), $\ldots$, (x_i, y_i), $\ldots$, (x_n, y_n). Suuruse Y väärtuste varieerumist ehk hajumist ümber aritmeetilise keskmise $\bar{y}$ iseloomustavad erinevused aritmeetilisest keskmisest $(y_i - \bar{y})$. Kui on olemas regressioonjoon, siis võime selle erinevuse jagada kaheks:

$$y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i), \qquad (9.28)$$

kus $\hat{y}_i = ax_i + b$ on empiirilisele punktile vastava regressioonjoonel asuva punkti y-koordinaat (vt joonis 9.10).



Joonis 9.10. Koguhajuvus, mudeliga seletatud hajuvus ja jääkhajuvus punkti $i = 3$ korral. Kriipsjoon vastab suuruse Y aritmeetilisele keskmisele

Valem (9.28) kehtib mingi ühe i-nda punkti jaoks. Et leida karakteristikut terve punktiparve jaoks, tuleb summeerida nende erinevuste

ruudud (vt dispersiooni valem (3.3)). Sellist summat tähistatakse regressioonanalüüsi korral tavaliselt SST-ga (*Total Sum of Squares*):

$$SST = \sum (y_i - \bar{y})^2. \tag{9.29}$$

Koguhajuvus

Eesti keeles nimetatakse sellist ruutude summat **koguhajuvuseks**.

Lähtudes seosest (9.28), on võimalik näidata, et lineaarse regressioonmudeli korral kehtib seos (vt lisa A.10)

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2. \tag{9.30}$$

Paremal pool olev esimene liidetav on suuruse Y hajumise see osa, mis on tingitud regressioonmudeliga kirjeldatud seosest X ja Y vahel. Seda nimetatakse **regressioonhajuvuseks** ja tähistatakse SSR-ga (*Sum of Squares due to Regression*):

Regressioonhajuvus

$$SSR = \sum (\hat{y}_i - \bar{y})^2. \tag{9.31}$$

Valemi (9.30) paremal poolel olevas teises liidetavas on hälbed regressioonjoonest, summeeritud on nende ruudud. See osa koguhajuvusest on regressioonmudeli poolt seletamata ja seda nimetatakse **jääkhajuvuseks**. Ka selle tähistamiseks kasutatakse kolmetähelist lühendit — SSE (*Sum of Squared Errors*):

Jääkhajuvus

$$SSE = \sum (y_i - \hat{y}_i)^2. \tag{9.32}$$

Kui me võrdleme valemit (9.32) valemiga (9.12), näeme, et jääkhajuvus SSE on hälvete u_i ruutude summa. Järelikult, vähimruutude meetodi korral minimeeritakse jääkhajuvust.

Nüüd võime seose (9.30) ümber kirjutada, kasutades tähistusi (9.29), (9.31) ja (9.32).

Koguhajuvus = regressioonhajuvus + jääkhajuvus:

$$SST = SSR + SSE. \tag{9.33}$$

Mudeli kirjeldusvõime on seda suurem, mida suurem on regressioonhajuvus võrreldes koguhajuvusega. Kirjeldusvõime hindamiseks sobib nende suhe, mis on **determinatsioonikordaja** (*R Squared*).

Regressioonmudeli **determinatsioonikordaja** näitab, kui suure osa koguhajuvusest moodustab regressioonhajuvus:

Determinatsioonikordaja

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}. \tag{9.34}$$

Joonise 9.9 vasakpoolsel graafikul (a) toodud punktiparve korral on vastava mudeli determinatsioonikordaja $R^2 = 0{,}94$, graafiku (b) korral aga $R^2 = 0{,}45$.

Näites 9.2 toodud mudeli korral, mis kirjeldas majapidamise sissetuleku ja elektrienergia tarbimise vahelist seost, on determinatsioonikordaja $R^2 = 0{,}59$. See tähendab, et 59% elektritarbimise varieerumisest on põhjustatud sellest, et erinevatel majapidamistel on erinev sissetulek. Ülejäänud 41% tarbimise varieerumisest on tingitud muudest põhjustest.

Valemist (9.34) on näha, et

1) determinatsioonikordaja maksimaalne väärtus on 1 (kui $SSE = 0$). Sellisel juhul jääkhajuvust ei esine, kõik jäägid on nullid;
2) determinatsioonikordaja minimaalne väärtus on 0 (kui $SSR = 0$). See tähendab, et mudeliga pole midagi ära seletatud;
3) kuna summade SSR ja SST ühikud on ühesugused (tunnuse Y ühik ruudus), on determinatsioonikordaja ühikuta suurus.

Determinatsioonikordaja on ühikuta suhtarv, mille väärtus jääb 0 ja 1 vahele:

$$0 \leq R^2 \leq 1.$$

Empiiriliste andmete korral pole determinatsioonikordaja kunagi täpselt 0 või täpselt 1. Võib olla näiteks 0,001 või 0,999.

Tabelarvutuses on determinatsioonikordaja leidmiseks lihtsa regressioonmudeli korral funktsioon **RSQ**.

Lineaarse seose korral iseloomustas seose tugevust ka lineaarne korrelatsioonikordaja. On võimalik näidata, et lineaarse korrelatsioonikordaja ja lineaarse mudeli determinatsioonikordaja vahel on lihtne seos.

Lineaarse mudeli determinatsioonikordaja R^2 võrdub lineaarse korrelatsioonikordaja r ruuduga:

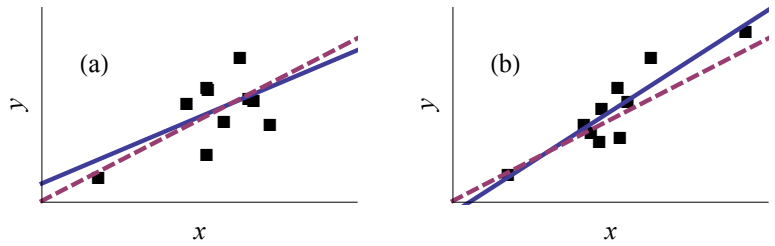
$$R^2 = r^2. \tag{9.35}$$

Mittelineaarse mudeli korral see võrdus ei kehti.

9.5. Regressioonsirge parameetrite usalduspiirid

Regressionanalüüsi läbiviimisel kasutame valimit ja valemite (9.15)–(9.16) põhjal saab leida lihtsa regressioonmudeli parameetrite hinnan-

gud. Mingi teise valimi puhul saame teistsugused hinnangud (joonis 9.11).



Joonis 9.11. Graafikul (a) on ühe valimi põhjal saadud regressioonsirge (pidev joon) ja graafikul (b) on teise valimi põhjal saadud regressioonsirge. Kriipsjoon on tegelik, üldkogumile vastav regressioonsirge

Seepärast tuleb eristada üldkogumis kehtivat tegelikku regressioonmudelit

$$\hat{y} = \alpha x + \beta,$$

mida me ei tea, ja valimi põhjal hinnatud mudelit

$$\hat{y} = ax + b,$$

kus α ja β on regressioonsirge parameetrite tõelised väärtused ning a ja b nende hinnangud. Korrektse analüüsi korral tuleb leida hinnangute standardvead ja usalduspiirid.

Ühe tunnuse korral kasutasime keskväärtuse usalduspiiride leidmiseks valimi standardhälvet (6.5). Selles valemis jagati hälvete ruutude summa vabadusastmete arvuga, mis oli $n-1$. Lineaarses mudelis (9.2) on kaks parameetrit, mille hinnangud leitakse valemitest (9.15)–(9.16). Kui me vaatame jääkhajuvust SSE, siis summas (9.32) on küll n liidetavat, kuid sõltumatuid on kahe kitsenduse (9.15) ja (9.16) tõttu $n-2$. Seetõttu on SSE arvutamisel vabadusastmete arv $n-2$. Jääkhajuvuse SSE põhjal leitakse jääkstandardhälve ehk lineaarse **regressioonmudeli standardviga**:

Mudeli standardviga

$$se = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}. \qquad (9.36)$$

Mudeli **standardviga** iseloomustab funktsioontunnuse Y väärtuste hajumist ümber regressioonsirge, nii nagu ühe tunnuse korral standardhälve iseloomustab üksikväärtuste hajumist ümber keskväärtuse.

Parameetrite standardvead, usalduspiirid

Mudeli standardvea põhjal saab leida **parameetrite standardvead**. Valikvaatluse korral alluvad regressioonsirge parameetrite hinnangud t-jaotusele, mis võimaldab leida ka vastavad **usalduspiirid**.

Lineaarliikme kordaja a standardviga:

$$se(a) = \frac{se}{\sqrt{\sum(x_i - \bar{x})^2}}. \tag{9.37}$$

Usaldatavusele β vastava usaldusvahemiku poollaius parameetri a jaoks

$$\Delta a = t_{\alpha/2}(\nu)\ se(a), \tag{9.38}$$

kus $t_{\alpha/2}(\nu)$ on vabadusastmete arvule $\nu = n - 2$ vastava t-jaotuse täiendkvantiil ja $\alpha = 1 - \beta$.

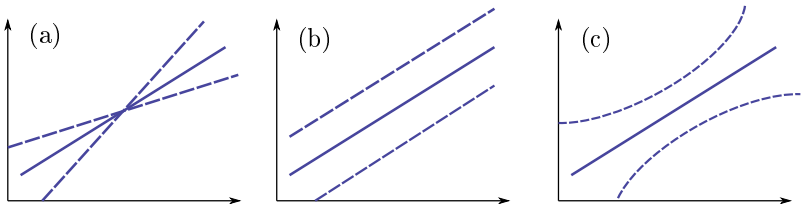
Vabaliikme b standardviga

$$se(b) = se\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}. \tag{9.39}$$

ja usaldusvahemiku poollaius

$$\Delta b = t_{\alpha/2}(\nu)\ se(b). \tag{9.40}$$

Lineaarliikme a ja vabaliikme b määramatus usaldusvahemiku piires põhjustab selle, et valimi põhjal leitud regressioonsirge asend võib olla erinev. Joonisel 9.12 (a) on kujutatud lineaarliikme määramatusest tingitud regressioonsirge määramatus, sirge võib olla kriipsjoontega piiratud koonuses. Joonisel 9.12 (b) on esitatud vabaliikme määramatusest tingitud regressioonsirge määramatus, sirge võib asuda kriipsjoontega piiratud koridoris. Mõlema parameetri määramatuse koosmõju on esitatud joonisel 9.12 (c).



Joonis 9.12. (a) Lineaarliikme, (b) vabaliikme ja (c) mõlema parameetri määramatusest tingitud regressioonsirge asendi määramatus

Tabelarvutuses kasutatakse mudeli standarvea leidmiseks funktsiooni **STEYX**. Parameetrite standardvigade leidmiseks funktsioon puudub. Programmis Excel on aga parameetrite standardvigade ning usalduspiiride leidmiseks otstarbekas kasutada vahendit *Regression* analüüsikomplektist *Data Analysis* (vt ka lisa C.9). See vahend

leiab parameetrite standarvead ja nende usalduspiirid etteantud usaldatavusega ning usaldatavusega 0,95 (vt tabel 9.6). Lisaks leitakse hüpoteeside kontrollimiseks vajalikud t-statistikud ning neile vastavad olulisuse tõenäosused (vt ptk 9.12). Analoogne tabel väljastatakse ka teistes statistilise analüüsi pakettides.

Tabel 9.6. Tarkvara genereeritava regressioonanalüüsi aruande struktuur

	Para-meetri hinnang	Standard-viga	t-statistik	Olulisuse tõenäosus p	Usaldusvahemik: Alumine piir	Usaldusvahemik: Ülemine piir
Vabaliige						
Lineaarliikme kordaja						

Näide 9.4. Autotootja BMW kulufunktsioon

N09Regressioon
N9.4

Autotootja BMW 2013. aasta finantsaruandes on viimase 10 aasta põhilised näitajad, nende hulgas ka toodetud autode arv ning kulud (BMW, 2013).

Aasta	Autode arv	Kulud, mln €
2004	1 250 345	35 389
2005	1 323 119	37 138
2006	1 366 838	37 370
2007	1 541 503	41 435
2008	1 439 918	39 169
2009	1 258 417	34 150
2010	1 481 253	38 303
2011	1 738 160	45 082
2012	1 861 826	48 208
2013	2 006 366	49 821

Leiame lineaarse kulufunktsiooni kujul

$$C = aq + b,$$

kus C on kulud miljonites eurodes ning q toodetud autode arv aastas ehk tootmismaht. Parameeter a on siis muutuvkulu tooteühiku kohta ehk piirkulu ja parameeter b püsikulu. Vastavate suuruste hinnangute ning usalduspiiride leidmiseks kasutame Excelis vahendit *Regression* komplektist *Data Analysis*. Aruande esimene osa sisaldab korrelatsioonikordajat (*Multiple R*) ja

determinatsioonikordajat (*R Square*) ning mudeli standardviga (*Standard Error*).

Regression Statistics	
Multiple R	0,991
R Square	0,982
Adjusted R Square	0,979
Standard Error	776,7
Observations	10

Aruande teist osa ANOVA vaatame lähemalt peatükis 9.11. Aruande viimane, kõige tähtsam osa esitab infot mudeli parameetrite kohta.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	9304	1535	6,06	0,000302	5764	12844
Autod	0,0205	0,000993	20,7	3,16E-08	0,0182	0,0228

Vabaliikme ehk püsikulude punkthinnang on 9304 miljonit eurot standardveaga 1535 miljonit eurot. Püsikulude vahemikhinnang usaldatavusega 95% on 5764 kuni 12 844 miljonit eurot.
Lineaarliikme kordaja ehk piirkulu punkthinnang on 0,0205 miljonit eurot auto kohta standardveaga 0,000993 miljonit eurot auto kohta. Piirkulu vahemikhinnang on 0,0182 kuni 0,0228 miljonit eurot auto kohta. Pannes parameetrite punkthinnangud mudeli üldkujusse, saame BMW kulufunktsiooniks

$$\hat{C} = 0{,}0205q + 9304,$$

kus C on kulud miljonites eurodes ning q aastas toodetud autode arv. Suurendades aasta toodangut 1 auto võrra, suurenevad BMW kulud 0,0205 miljoni euro võrra ehk 20,5 tuhat eurot. Determinatsioonikordaja näitab, et autode tootmismahu muutus kirjeldab ära 98,2% kulude muutusest.

9.6. Mudeli kasutamine prognoosimiseks

Regressioonmudelit võib kasutada prognoosimiseks. Võttes ette argumendi väärtuse x_A, võime prognoosida sellele väärtusele vastavat

1) funktsioontunnuse üksikväärtust $\hat{y}_A$;
2) funktsioontunnuse keskväärtust $\hat{\mu}_A = E(\hat{y}_A|x_A)$.

Üksikväärtuse prognoos on regressioonmudeli mudelväärtus

$$\hat{y}_A = ax_A + b. \tag{9.41}$$

Üksikväärtuse prognoosi standardviga

$$se(\hat{y}_A) = se\sqrt{1 + \frac{1}{n} + \frac{(x_A - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \tag{9.42}$$

ja usaldusvahemiku poollaius usaldatavuse $\beta = 1 - \alpha$ korral

$$\Delta\hat{y}_A = t_{\alpha/2}(\nu)\ se(\hat{y}_A). \tag{9.43}$$

Keskväärtuse prognoos on

$$\hat{\mu}_A = ax_A + b \tag{9.44}$$

ja selle standardviga

$$se(\hat{\mu}_A) = se\sqrt{\frac{1}{n} + \frac{(x_A - \bar{x})^2}{\sum(x_i - \bar{x})^2}}. \tag{9.45}$$

Keskväärtuse prognoosi usaldusvahemiku poollaius

$$\Delta\hat{\mu}_A = t_{\alpha/2}(\nu)\ se(\hat{\mu}_A). \tag{9.46}$$

Mõlema usaldusvahemiku poollaiuse valemis on t-jaotuse vabadusastmete arv $\nu = n - 2$, kus n on valimi maht. Rõhutada tuleb, et need valemid kehtivad vaid lihtsa lineaarse regressioonmudeli korral.

Nagu valemitest (9.41) ja (9.44) näeme, on üksikväärtuse ja keskväärtuse prognoosi punkthinnangud ühesugused. Erinev on prognoosi standardviga. Valemite (9.42) ja (9.45) võrdlus näitab, et keskväärtuse prognoosi standardviga on väiksem. See on loomulik, sest väärtusega x_A üksikobjekti väärtuse $\hat{y}_A$ prognoosimisel on määramatus suurem kui kõigi väärtusega x_A objektide keskväärtuse prognoosimisel.

Mõlema prognoosi standardvea valemis on ruutjuure all murd, mille lugeja on $(x_A - \bar{x})^2$. See avaldis on seda suurem, mida kaugemal asub argumenttunnuse väärtus x_A keskmisest $\bar{x}$. Järelikult on mõlema prognoosi standardviga seda suurem, mida kaugemal on prognoosi aluseks olev väärtus x_A argumenttunnuse keskmisest.

Prognoosi standardvea leidmiseks tabelarvutuses funktsioon puudub ning arvutada tuleb vastavalt valemile (9.42) või (9.45). Mudeli standardvea *se* leidmiseks saab kasutada funktsiooni **STEYX**. Hälvete ruutude summa $\sum(x_i - \bar{x})^2$ leidmiseks võib kasutada funktsiooni **DEVSQ**.

N09Regressioon
N9.5

Näide 9.5. Toidule tehtavate kulutuste prognoosimine

Näites 9.3 leidsime, et toidule tehtavate kulutuste sõltuvust kogukuludest kirjeldab mudel

$$\hat{y}_T = 434{,}4 + 0{,}1297x,$$

kus y on kulutused toidule pereliikme kohta aastas (eurot) ning x kulud kokku pereliikme kohta aastas (eurot).
Kui palju kulutab pereliikme kohta toidule pere, kes kokku kulutab aastas 2000 eurot pereliikme kohta? See on üksikväärtuse prognoosimine. Prognoosi punkthinnang

$$\hat{y}_T(2000) = 434{,}4 + 0{,}1297 \cdot 2000 \approx 694.$$

Standardvea leidmiseks kasutame valemit (9.42). Mudeli standardvea se leiame tabelarvutuses funktsiooni STEYX abil ja see on 130,9 eurot. Aritmeetiline keskmine $\bar{x} = 3749{,}9$ eurot ja hälvete ruutude summa $\sum(x_i - \bar{x})^2 = 60440436$ eurot^2. Viimase leidmiseks kasutame tabelarvutuses funktsiooni DEVSQ. Üksikväärtuse prognoosi standardviga valemist (9.42)

$$se\left(\hat{y}(2000)\right) = 130{,}9 \cdot \sqrt{1 + \frac{1}{10} + \frac{(2000 - 3749{,}9)^2}{60440436}} \approx 140{,}43.$$

Usaldatavuse 0,95 korral $t_{0{,}025}(8) = 2{,}306$, mille leidmiseks on tabelarvutuses funktsioon T.INV.2T. Usaldusvahemiku poollaius

$$\Delta\hat{y}(2000) = 2{,}306 \cdot 140{,}43 \approx 324.$$

Usaldatavusega 0,95 on sellise pere korral toidule tehtavad kulud

$$\hat{y}_T(2000) = 694 \pm 324$$

eurot pereliikme kohta aastas ehk vahemikus 370 kuni 1018 eurot.
Kui me võtame aga vaatluse alla kõik pered, kes kulutavad aastas 2000 eurot pereliikme kohta ja soovime leida nende perede toidule tehtavate kulutuste keskmist, siis see on keskväärtuse prognoosimine. Selle prognoosi punkthinnang on samuti

$$\hat{\mu}_T(2000) = 434{,}4 + 0{,}1297 \cdot 2000 \approx 694.$$

Keskväärtuse prognoosi standardviga valemist (9.45)

$$se(\hat{\mu}_T(2000)) = 130{,}9 \cdot \sqrt{\frac{1}{10} + \frac{(2000 - 3749{,}9)^2}{60440436}} \approx 50{,}81$$

ning usaldusvahemiku poollaius

$$\Delta\hat{\mu}_T(2000) = 2{,}306 \cdot 50{,}81 \approx 117.$$

Keskmiselt kulutavad sellised pered toidule

$$\hat{\mu}_T(2000) = 694 \pm 117$$

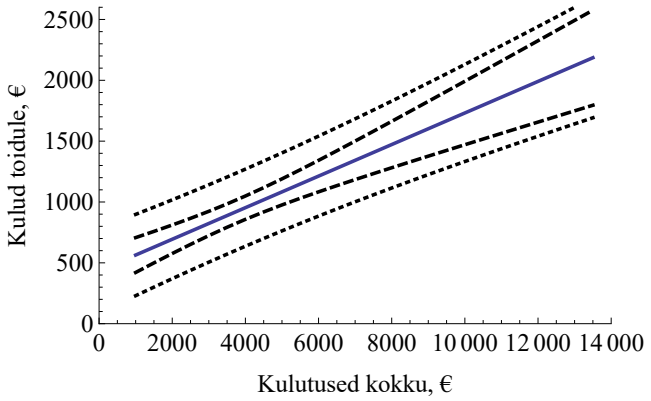
eurot pereliikme kohta aastas ehk vahemikus 577 kuni 811 eurot. Kui teha analoogsed arvutused pere jaoks, kes kokku kulutab 6000 eurot aastas pereliikme kohta, siis ühe pere korral on toidule tehtavad kulutused

$$\hat{y}_T(6000) = 1213 \pm 328 \quad \text{eurot}$$

ja kõikide selliste perede keskmine

$$\hat{\mu}_T(6000) = 1213 \pm 129 \quad \text{eurot.}$$

Kui me leiame usaldusvahemiku alumised ja ülemised piirid erinevate kogukulude väärtuste jaoks, võime need veakoridoridena joonisele kanda. Joonisel 9.13 on kriipsjoontega kujutatud keskväärtuse prognoosi usalduspiirid ning punktiirjoontega üksikväärtuse prognoosi usalduspiirid. Mõlemad usaldusvahemikud on kõige kitsamad kohas, kus asub kogukulude keskmine $\bar{x} = 3750$ eurot.



Joonis 9.13. Toidule tehtavate kulutuste prognoos. Kriipsjooned näitavad keskväärtuse prognoosi usalduspiire, punktiirjooned üksikväärtuse prognoosi usalduspiire

9.7. Mittelineaarne regressioon

Lineaarne mudel on kõige lihtsam mudel ja selle parameetrite leidmine on arvutuslikult kõige lihtsam. Lineaarse mudeli parameetrid on ka lihtsalt tõlgendatavad. Paraku esineb tegelikkuses tihti mittelineaarseid seoseid, mille modelleerimiseks tuleb kasutada mittelineaarseid regressioonmudeleid. Mittelineaarsust on võimalik avastada hajumisdiagrammi uurides ja tihti on see ka majandusteoreetiliselt põhjendatud. Mittelineaarsuse esinemine annab meile analüüsitava nähtuse kohta olulist lisainformatsiooni.

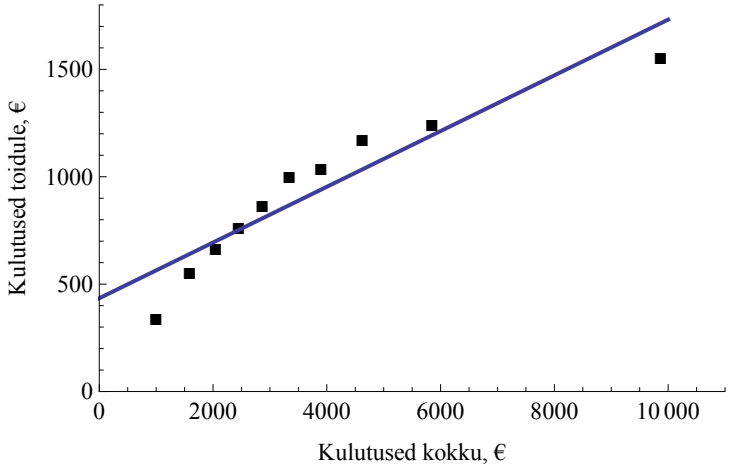
Näide 9.6. Toidule tehtavate kulutuste mittelineaarne sõltuvus kogukuludest

N09Regressioon
N9.6

Näites 9.3 leidsime kolm tarbimismudelit. Kasutasime lineaarset mudelit ja toidule tehtavate kulutuste sõltuvus kogukuludest oli

$$\hat{y}_T = 434{,}4 + 0{,}1297x, \tag{9.47}$$

kus y_T on kulutused toidule pereliikme kohta aastas (eurot) ja x kogukulud pereliikme kohta aastas (eurot). Selle mudeli determinatsioonikordaja on $R^2 = 0{,}881$.



Joonis 9.14. Kulud toidule, lineaarne mudel

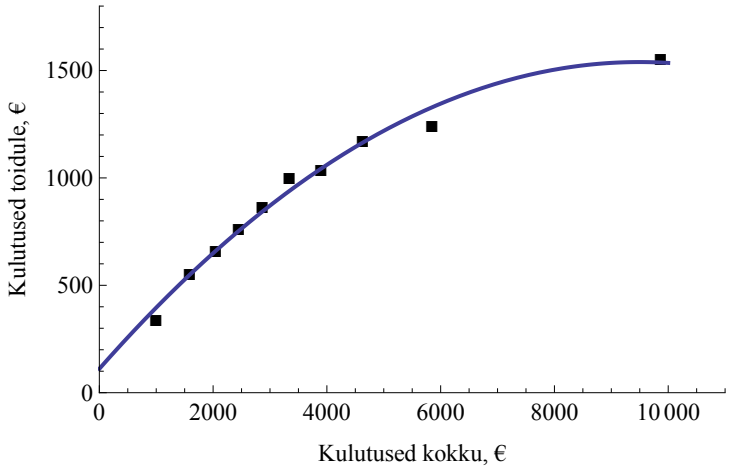
Jooniselt 9.14 näeme, et punktide kõrvalekaldumine sirgest ei ole juhuslik. Keskel on punktid regressioonjoonest kõrgemal, otstes madalamal. Sobivam on seda punktiparve lähendada kõverjoonega, näiteks ruutparabooliga:

$$y_T = ax^2 + bx + c.$$

Paraboolne regressioonanalüüs annab mudeliks

$$\hat{y}_T = 111{,}2 + 0{,}3x - 2 \cdot 10^{-5}x^2, \ R^2 = 0{,}987. \qquad (9.48)$$

Jooniselt 9.15 on näha, et ruutparabool läbib seda punktiparve oluliselt paremini kui sirge.



Joonis 9.15. Kulud toidule, mudeliks ruutpolünoom

Determinatsioonikordajate võrdlemine näitab, et mudeli (9.48) kirjeldusvõime on oluliselt suurem kui lineaarsel mudelil (9.47). Ruutliikme ees olev kordaja on küll väga väike, kuid ruutliige avaldab märgatavat mõju, sest x^2 väärtused on suured.
Mudeli mittelineaarsus tähendab siin seda, et väikeste kogukulude korral kasvavad kulud toidule järsemalt, suuremate kogukulude korral aga toidukulude kasv aeglustub. Kui pere saab oma kulusid näiteks 100 euro võrra suurendada, siis vaesemal perel kasvavad kulud toidule rohkem kui rikkamal perel. Teeme näiteks vastavad arvutused.
Matemaatikast on teada, et funktsiooni muutumise kiirust mingis punktis näitab funktsiooni tuletis selles punktis. Leiame funktsiooni (9.48) tuletise:

$$\hat{y}'_T = 0{,}3 - 4 \cdot 10^{-5}x.$$

Kui kulud pereliikme kohta on 2000 eurot aastas, siis

$$\hat{y}'_T(2000) = 0{,}3 - 4 \cdot 10^{-5} \cdot 2000 \approx 0{,}22.$$

See tähendab, et sellise pere korral kogukulude suurenemisel 1 euro võrra kasvavad kulud toidule 0,22 euro võrra. Kui kulud pereliikme kohta on aga 6000 eurot aastas, siis

$$\hat{y}'_T(6000) = 0{,}3 - 4 \cdot 10^{-5} \cdot 6000 \approx 0{,}06.$$

Sellise pere puhul kogukulude suurenemisel 1 euro võrra kasvavad kulud toidule 0,06 euro võrra. Lineaarne mudel (9.47) annab aga mõlemal juhul toidule tehtavate kulutuste ühesuguse suurenemise *ca* 0,13 euro võrra. Ilmselt kirjeldab paraboolne mudel perede tarbimisharjumusi paremini.

Nägime, et regressiooonmudeli parandamine mitte lihtsalt ei andnud parema kirjeldusvõimega mudelit, vaid andis ka olulist teavet tarbijate käitumise kohta. Kui mittelineaarse mudeli kirjeldusvõime (determinatsioonikordaja) on suurem, siis tuleb kasutada seda mudelit, mis tähendab, et kasv (või kahanemine) ei ole konstantne.

Paraboolne mudel

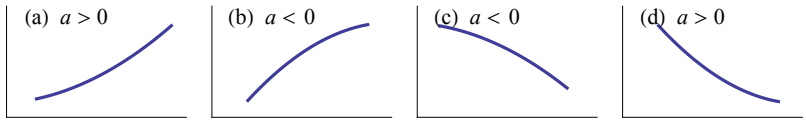
Matemaatikast on teada, et ruutpolünoomi $y = ax^2 + bx + c$ graafik on parabool. Ruutliikme ees oleva kordaja märk määrab ära, kas parabool on ülespoole avatud ($a > 0$) või allapoole avatud ($a < 0$). Sõltuvalt ruutliikme ees olevast märgist ja sellest, kas andmetega haaratud piirkond on parabooli kasvavas või kahanevas osas, sõltub ka ruutliikme tõlgendus.

Kasvamine võib olla

- kiirenev — ruutliikme kordaja on positiivne (joonis 9.16 (a));
- aeglustuv — ruutliikme kordaja on negatiivne (joonis 9.16 (b)).

Kahanemine võib olla

- kiirenev — ruutliikme kordaja on negatiivne (joonis 9.16 (c));
- aeglustuv — ruutliikme kordaja on positiivne (joonis 9.16 (d)).



Joonis 9.16. Kiirenev (a) ja aeglustuv (b) kasvamine, kiirenev (c) ja aeglustuv (d) kahanemine. a on ruutliikme ees olev kordaja

Näide 9.7. Sööda kulu optimeerimine kalakasvanduses

N09Regressioon
N9.7

Söödakoefitsient (*feed conversion ratio*, FCR) iseloomustab loomade võimet tarbitud sööda arvelt oma keha massi kasvatada:

$$FCR = \frac{R}{G},$$

kus R on mingi perioodi jooksul tarbitud sööda kogus ning G looma biomassi suurenemine, mõlemad kilogrammides. Mida väiksem on FCR, seda efektiivsemalt loom sööta omastab ja seda vähem sööta kulub. Näiteks kana korral $FCR = 2$, mis

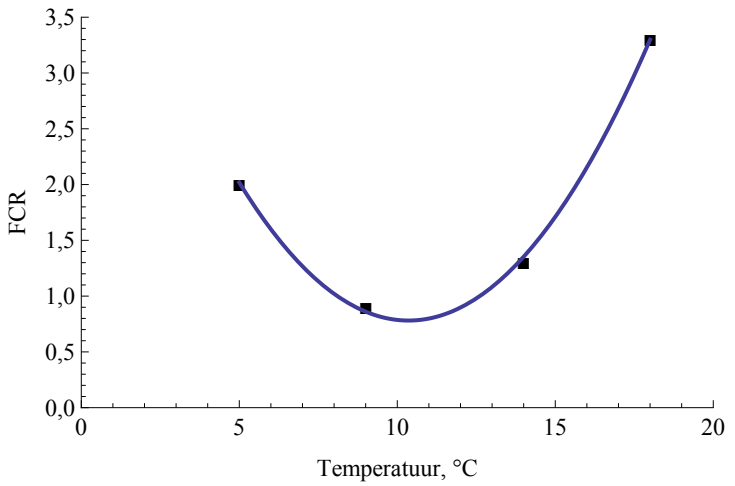
tähendab, et kana massi suurendamiseks 1 kg võrra kulub 2 kg sööta. Sea puhul $FCR = 6$, järelikult sea kehakaalu suurendamiseks 1 kg võrra kulub 6 kg sööta.
Kaladel sõltub söödakoefitsient veetemperatuurist ning kalakasvatajatel on soovitatav hoida basseinides optimaalset veetemperatuuri. Ühes uuringus (Handeland, Imsland ja Stefansson, 2008) kasvatati lõhesid neljas erinevas basseinis, kus temperatuur oli vastavalt 5 °C, 9 °C, 14 °C ja 18 °C. Määrati, kui palju kalad keskmiselt kaalus juurde võtsid, kui neile anti kindel kogus sööta, ning arvutati söödakoefitsient. Tulemused 170–300-grammiste lõhede jaoks on esitatud tabelis.

Temperatuur, °C	FCR
5	2,0
9	0,9
14	1,3
18	3,3

Kandes vaatluspunktid graafikule, näeme, et mudeliks sobib ruutparabool. Parabooli võrrand:

$$\widehat{FCR} = 0{,}0431t^2 - 0{,}8929t + 5{,}4046, \quad R^2 = 0{,}999, \qquad (9.49)$$

kus t on temperatuur Celsiuse kraadides.



Joonis 9.17. Kalade söödakoefitsiendi FCR sõltuvus temperatuurist

Et kulud söödale oleksid minimaalsed, peab söödakoefitsient olema võimalikult väike. Leiame funktsiooni (9.49) miinimumkoha.

Selleks võtame temperatuuri järgi tuletise, paneme selle võrduma nulliga ning leiame temperatuuri t:

$$\begin{aligned} \widehat{FCR}' &= 0{,}0862t - 0{,}8929 \\ 0{,}0862t - 0{,}8929 &= 0 \\ t &= \frac{0{,}8929}{0{,}0862} \\ t &\approx 10{,}4. \end{aligned}$$

Optimaalne basseinivee temperatuur on 10,4 °C. Leiame ka söödakoefitsiendi väärtuse selle temperatuuri korral. Selleks paneme optimaalse temperatuuri väärtuse mudelisse (9.49):

$$\widehat{FCR} = 0{,}0431 \cdot 10{,}4^2 - 0{,}8929 \cdot 10{,}4 + 5{,}4046 \approx 0{,}78.$$

Söödakoefitsiendi ühest väiksem väärtus tähendab, et kalade biomassi juurdekasv on suurem kui söödakulu. See on võimalik seepärast, et kala sisaldab palju vett, sööt on aga kuiv.

Kui graafikult on näha, et empiiriliste punktide hälbed leitud regressioonjoonest ei ole juhuslikud, vaid esineb teatud süstemaatilisus (nagu joonisel 9.14), on soovitatav muuta regressioonmudeli kuju. Tuleks leida sellise kujuga regressioonjoon, mille korral empiirilised punktid asuvad juhuslikult kord ühel, kord teisel pool joont.

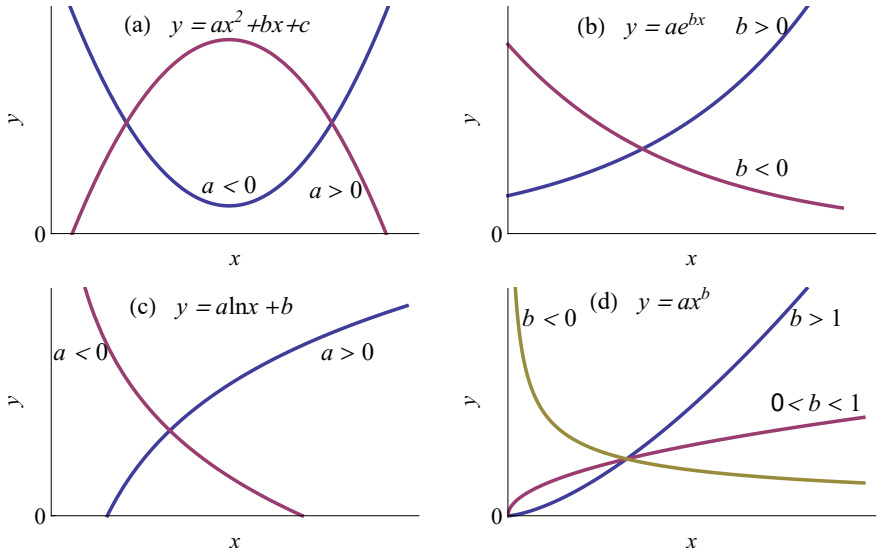
Tabel 9.7. Kõige sagedamini kasutatavad mittelineaarsed regressioonmudelid

Ruutfunktsioon	$y = ax^2 + bx + c$	Joonis 9.18 (a)
Eksponentsiaalne	$y = ae^{bx}$	Joonis 9.18 (b)
Logaritmiline	$y = a \ln x + b$	Joonis 9.18 (c)
Astmefunktsioon	$y = ax^b$	Joonis 9.18 (d)

Lisaks ruutfunktsioonile on võimalik kasutada ka kuupfunktsiooni $y = ax^3 + bx^2 + cx + d$ ning kõrgema astme polünoome, üldiselt n-astme polünoomi $y = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$. Mingi funktsiooni lähendamine n-astme polünoomiga on matemaatikas tuntud meetod. Silmas tuleb pidada aga seda, et me soovime modelleerimise käigus saada informatsiooni reaalsete nähtuste kohta ning kõrgemat järku polünoomide tõlgendamine on raskendatud.

Mudeli kehtivuspiirkond

Oluline on rõhutada seda, et mudelil on teatud **kehtivuspiirkond**. Vaatlusandmetega haaratud piirkonna jaoks on meil empiiriline tõestus, et seal mudel kehtib. Selleks tõestuseks on mudeli kirjeldusvõime näitaja determinatsioonikordaja. Me ei saa aga kindlad olla, et mudel



Joonis 9.18. Sagedamini kasutatavad mittelineaarsed regressioonmudelid

kehtib ka väljaspool seda piirkonda. Näites 9.7 nägime jooniselt 9.17, et on olemas miinimum, sest ilmnes nii kahanemine kui ka kasvamine. Siis võib parabooli võrrandist miinimumkoha leida. Kuid näites 9.6 ei tohi pikendada joonisel 9.15 toodud ruutpolünoomi graafikut kaugemale kui kogukulud 10 000 eurot, sest siis hakkab ruutparabool langema. Pole loogiline, et kogukulude kasvades kulud toidule kahanevad.

Sobiva kuju leidmiseks on põhimõtteliselt kaks võimalust.

1. Teooriast on teada, millise kujuga seos uuritavate suuruste vahel eksisteerib, on teada mudeli üldkuju. Sellisel juhul on vaja leida vaid konkreetsele andmestikule vastavad parameetrid, s.t mudeli konkreetne kuju. Näiteks seos inflatsiooni ja töötuse määra vahel on erinevates riikides ühesugune, erinevad on vaid mudeli parameetrid.
2. Seose kuju pole teada. Sellisel juhul proovitakse erineva kujuga mudeleid ning determinatsioonikordaja järgi otsustatakse, millise neist kirjeldab seda seost kõige paremini (determinatsioonikordaja on kõige suurem). Järelikult võimaldab regressioonanalüüs täiendada ka majandusteooriat, pakkudes välja empiirilise mudeli uuritava seose kirjeldamiseks.

N09Regressioon
N9.8

Näide 9.8. Sissetulek ja elektrienergia tarbimine: logaritmiline mudel

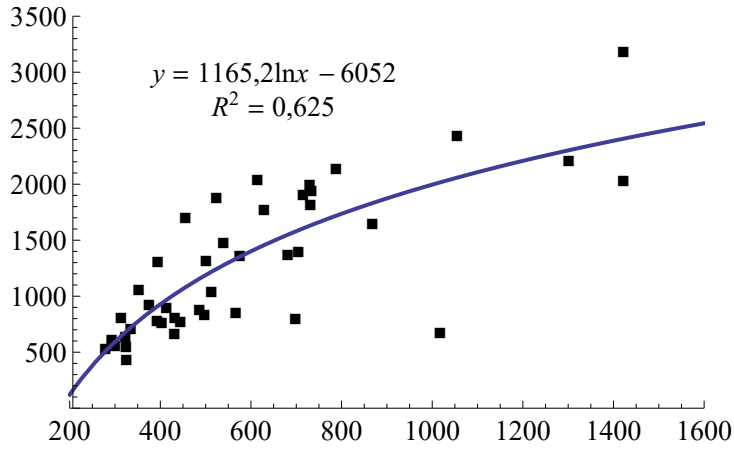
Hüvise nõudluse funktsionaalset sõltuvust sissetulekust kirjeldab Engeli[a] kõver (*Engel curve*). Sõltuvalt hüvise tüübist võib Engeli kõver olla erineva kujuga (Kaldaru, 1995, lk 56). Vajalike hüviste korral aeglustub sissetuleku suurenedes nõudluse kasv. Näites 9.2 modelleerisime elektrienergia tarbimise y sõltuvust sissetulekust x lineaarse mudeliga ja mudeliks saime $\hat{y} = 1{,}68x + 274$. Lineaarse mudeli korral on aga kasv konstantne. Aeglustuva kasvuga on näiteks logaritmiline mudel (joonis 9.18 (c)). Seepärast kasutame nüüd elektrienergia tarbimise ja sissetuleku vahelise seose modelleerimiseks logaritmilist mudelit

$$y = a \ln x + b, \tag{9.50}$$

kus y on elektrienergia tarbimine (kWh aastas) ja x majapidamise sissetulek (GBP aastas). Regressioonanalüüsi läbiviimisel saame mudeliks (vt joonis 9.19)

$$\hat{y} = 1165{,}2 \ln x - 6052, \quad R^2 = 0{,}625. \tag{9.51}$$

Determinatsioonikordaja on 0,625, mis on suurem kui lineaarse mudeli determinatsioonikordaja 0,59. Järelikult sobib logaritmiline mudel selle seose kirjeldamiseks paremini.



Joonis 9.19. Elektrienergia kasutamine, logaritmiline mudel

Logaritmilise kasvu mudel on ruutparaboolist parem seetõttu, et selle graafik ei hakka langema. Prognoosime selle mudeli järgi elektri tarbimist, kui sissetulek on 1200 GBP aastas. Arvutus mudeli (9.51) järgi:

$$\hat{y}(1200) = 1165{,}2 \cdot \ln 1200 - 6052 \approx 2209.$$

Järelikult, kui majapidamise sissetulek on 1200 GBP aastas, siis prognoositav elektrienergia tarbimine on 2209 kWh aastas. See on väiksem kui lineaarse mudeli (9.22) põhjal saadud 2293 kWh aastas.

[a] Ernst Engel (1821–1896), saksa statistik.

Mittelineaarse regressiooni puhul leitakse mudeli parameetrid samal põhimõttel, mis lineaarse regressiooni korral: kasutatakse vähimruutude meetodit (9.11). Kuid mittelineaarsel juhul ei ole võimalik leida parameetrite jaoks arvutusvalemeid. Parameetrite leidmiseks kasutatakse numbrilist meetodit. Selle käigus antakse parameetritele mingid arvväärtused, leitakse hälbed regressioonjoonest ja hälvete ruutude summa. Seejärel muudetakse parameetreid, leitakse uuesti hälvete ruutude summa ja vaadatakse, kas see väheneb. Kui jah, muudetakse parameetreid samas suunas ja leitakse jällegi hälvete ruutude summa. Kui summa aga suurenes, muudetakse parameetreid teises suunas. Nii jätkatakse samm-sammult, kuni hälvete ruutude summa enam ei väheni. See algoritm on vastavas tarkvaras programmeeritud ning kasutaja sellele mõtlema ei pea.

Tabelarvutuses on mittelineaarse regressioonmudeli leidmiseks kõige lihtsam kasutada hajumisdiagrammi, millele on lisatud vastava kujuga trendijoon (*trendline*). Selleks tuleb empiiriliste andmete põhjal koostatud hajumisdiagrammil märkida ära vastav punktikogum ning hiire parempoolse klahviga avatavast objektmenüüst valida „Lisa trendijoon“ (*Add Trendline*). Seejärel tuleb valida sobiva kujuga mudel ning lisavalikutest määrata, et diagrammil kuvataks nii mudel (*Display Equation on Chart*) kui ka determinatsioonikordaja (*R-squared*).

Tabelis 9.7 toodud mittelineaarsete mudelite parameetrite määramiseks on võimalik kasutada ka lineaarset regressioonanalüüsi. Selleks tuleb mudel eelnevalt lineariseerida ja teisendada andmeid. Lähemalt vaatame seda alapeatükis 9.17.

9.8. Jääkide analüüs

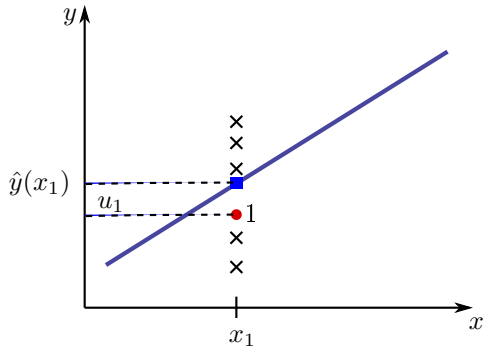
Regressioonmudeli analüüsimise üks meetodeid on jääkide analüüs. See võimaldab otsustada, kas mudeli kuju on sobivalt valitud, ning analüüsida võimalike ebatüüpiliste objektide olemasolu.

Valimisse kuuluva i-nda objekti tegeliku väärtuse y_i ja mudelväärtuse $\hat{y}_i$ vahe u_i on hälve ehk **regressioonijääk** (*residual*):

$$u_i = y_i - \hat{y}_i. \tag{9.52}$$

Regressioonijääk

Rõhutada tuleb seda, et regressioonijääk u_i ja juhuslik komponent ε_i ei ole üks ja seesama. Juhusliku komponendi väärtust me ei tea, regressioonijäägid saame aga valimis olevate objektide jaoks leida valemist (9.52). Regressioonmudel ei kirjelda mitte ainult valimisse kuuluvaid objekte, vaid tervet üldkogumit. Üldkogumis võib leiduda aga objekte, millel on ühesugune X väärtus, kuid erinev funktsioontunnuse Y väärtus (joonis 9.20).



Joonis 9.20. Punane punkt 1 vastab valimis olevale objektile, mille jaoks saame leida regressioonijäägi u_1. Sama argumendi väärtus x_1 võib kogumis olla aga mitmel objektil. Nende jaoks on funktsioontunnuse väärtus $\hat{y}(x_1)+\varepsilon$

Tuletame meelde, et regressioonmudeli deterministlik osa ehk mudelväärtus määrab ära funktsioontunnuse tingliku keskväärtuse (9.6). Regressioonmudel tuleb alati kirja panna terve kogumi jaoks, seepärast peab mudelväärtusele lisama tundmatu juhusliku komponendi ε:

$$y_i = \hat{y}_i + \varepsilon_i.$$

Regressioonijääke u_i, mida me teame, saame aga analüüsida. Analüüsitakse näiteks regressioonijääkide paiknemist ja selleks kasutatakse vastavat diagrammi.

N09Regressioon
N9.9

Näide 9.9. Kulud toidule ja jääkide analüüs

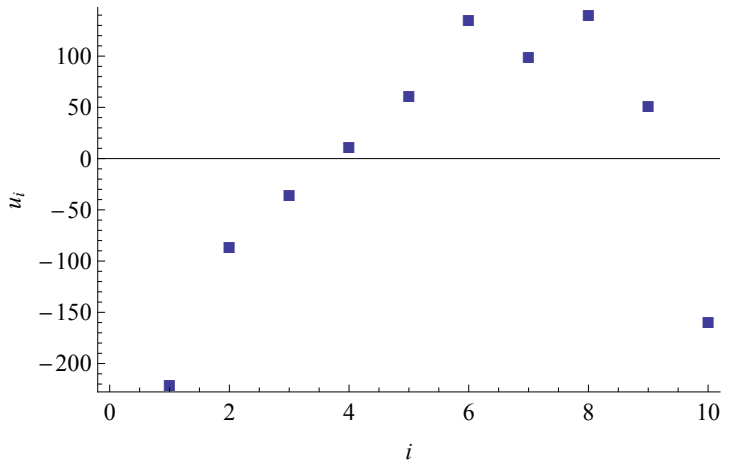
Näites 9.3 modelleerisime toiduainetele tehtavate kulutuste sõltuvust kogukuludest lineaarse mudeliga

$$\hat{y}_T = 434{,}4 + 0{,}1297x,$$

kus y_T on kulud toiduainetele pereliikme kohta kuus ja x kogukulud. Selle mudeli põhjal on arvutatud tabeli kolmas veerg. Tabeli viimases veerus on leitud vastavad jäägid.

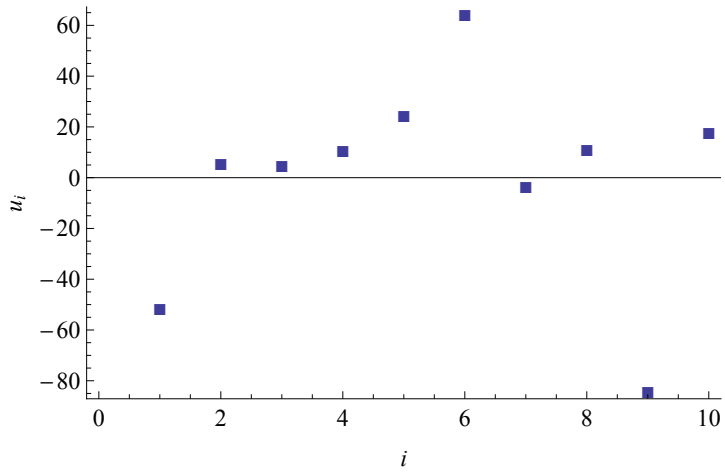
i	y_i	$\hat{y}_i$	$u_i = y_i - \hat{y}_i$
1	342,6	563,1	−220,5
2	554,1	640,0	−85,9
3	664,5	699,3	−34,8
4	763,6	752,0	11,6
5	867,2	805,7	61,5
6	1002,9	867,2	135,7
7	1038,8	939,3	99,5
8	1174,4	1 034,1	140,3
9	1244,5	1 192,9	51,6
10	1555,0	1 714,1	−159,1

Joonisel 9.21 olev hajumisdiagramm on koostatud tabeli esimese ja viimase veeru põhjal: horisontaalteljel on vaatluse järjenumber i ja vertikaalteljel jääk u_i. Näeme, et jäägid ei ole juhuslikud, esineb süstemaatiline kõrvalekalle: keskmiste vaatluste korral on kõik jäägid positiivsed, äärmiste vaatluste korral negatiivsed. Seepärast lineaarne mudel antud seose kirjeldamiseks ei sobigi.



Joonis 9.21. Kulud toidule, lineaarse mudeli jäägid

Näites 9.6 leidsime toidule tehtavate kulutuste jaoks paraboolse mudeli. Joonisel 9.22 on jääkide diagramm paraboolse mudeli (9.48) korral. Jääkide paigutus on juhuslikum kui joonisel 9.21.



Joonis 9.22. Kulud toidule, paraboolse mudeli jäägid

Standardiseeritud jääk

Erinevate mudelite korral on jääkide suurus erinev. Et oleks lihtsam hinnata, milliste objektide korral on jääk väga suur, leitakse **standardiseeritud jäägid** (*Standardized Residuals*), mis alluvad standardiseeritud normaaljaotusele. Arvestades standardiseeritud skaalale üleminekuks kasutatavat valemit (3.9) ja seda, et jääkide aritmeetiline keskmine $\bar{u} = 0$, saadakse standardiseeritud jääkide jaoks valem

$$u_i^{std} = \frac{u_i}{s}, \tag{9.53}$$

kus jääkide valimi standardhälve

$$s = \sqrt{\frac{1}{n-1}\sum u_i^2}. \tag{9.54}$$

Siin on n valimi maht ning standardhälbe leidmisel arvestame jällegi seda, et $\bar{u} = 0$.

Jääkide diagramm

Jääkide diagramm (*Residual Plot*) on hajumisdiagramm, kus vertikaalteljel on regressioonijäägid (või standardiseeritud jäägid), horisontaalteljel võivad olla

- argumenttunnuse väärtused x_i;
- funktsioontunnuse mudelväärtused $\hat{y}_i$;
- vaatluste järjekorranumbrid.

Programmis Excel saab vahendi *Regression* kasutamisel lisada regressioonanalüüsi aruandele jäägid, standardiseeritud jäägid ning jääkide diagrammi, kui aknas *Regression* märkida ära *Residuals*, *Standardized Residuals*, *Residuals Plots* (vt lisa C.9). Jääkide diagrammi horisontaalteljel on Excelis argumenttunnuse X väärtused.

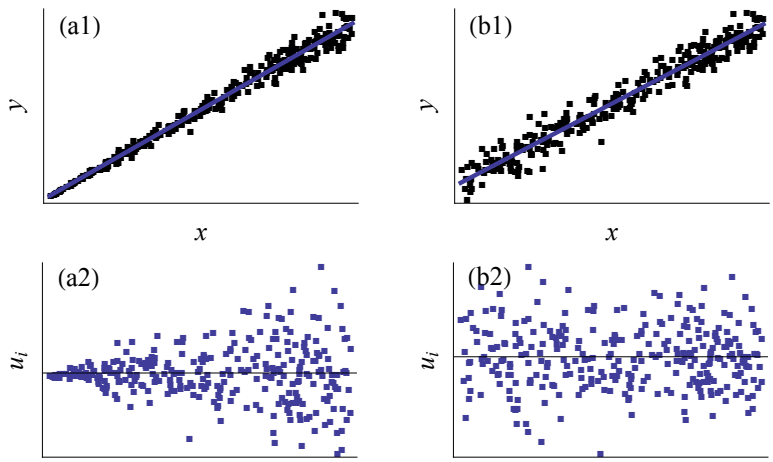
Vähimruutude meetodi kasutamisel eeldatakse jääkide kohta järgmist:
1) jäägid on normaaljaotusega, mille keskväärtus on 0;
2) jääkide hajuvus on konstantne;
3) jäägid ei sõltu argumenttunnuse väärtustest.

Vähimruutude meetodi eeldused

Korrektseks analüüsiks tuleb kontrollida nende eelduste täitumist. Üheks võimaluseks on jääkide diagrammi analüüsimine. Kui need eeldused on täidetud, peaks jääkide diagrammil olev punktiparv kujutama endast ühtlase laiusega riba, kusjuures

- positiivsete ja negatiivsete jääkide arv peab olema ligikaudu võrdne;
- standardiseeritud jääkide jaotus peab vastama standardiseeritud normaaljaotusele (keskväärtus 0, standardhälve 1) ning erinevatesse vahemikesse jäävad osakaalud peavad olema vastavalt:
  - vahemikus $(-1, 1)$ ligikaudu 68%;
  - vahemikus $(-2, 2)$ ligikaudu 95%;
  - vahemikus $(-3, 3)$ ligikaudu 99%;
- jääkide varieeruvus peab olema konstantne.

Joonis 9.23 illustreerib jääkide varieeruvust. Vasakul pool on jääkide varieeruvus väikeste x väärtuste korral väike ja varieeruvus suureneb x suurenemisel. Paremal pool on jäägid konstantse varieeruvusega.



Joonis 9.23. Graafikul (a1) ja vastaval jääkide diagrammil (a2) ei ole jääkide varieeruvus konstantne, graafikul (b2) ja vastaval jääkide diagrammil on jäägid konstantse varieeruvusega

Jääkide analüüs võimaldab ka hinnata, kui üks või mitu vaatluspunkti erinevad mudelist oluliselt. Standardiseeritud jääkidest peab ligikaudu 95% jääma vahemikku $(-2, 2)$. Kõik need standardiseeritud jäägid, mis jäävad sellest vahemikust välja, nõuavad suuremat tähelepanu. Tuleb analüüsida, kas on tegemist juhusega või vastava vaatluse erilise iseloomuga.

Ebatüüpiline vaatlus ja erind

Kui standardiseeritud jäägi absoluutväärtus
$|u_i^{std}| > 2$, on tegemist **ebatüübilise** vaatlusega;
$|u_i^{std}| > 3$, on tegemist **erindiga** (*outlier*).

Ebatüüpiliste vaatluste väljaselgitamine võimaldab neid lähemalt uurida. Eesmärgiks on välja selgitada põhjused, miks üks või teine vaatlus on ebatüüpiline. Alapeatükis 9.22 modelleeritakse autotööstuse tootmisfunktsiooni ning standardiseeritud jääkide analüüs näitab, et üks autotootja on ebatüüpiline. Mis selle tootjaga juhtus, võib viidatud alapeatükist lugeda.

N09Regressioon
N9.10

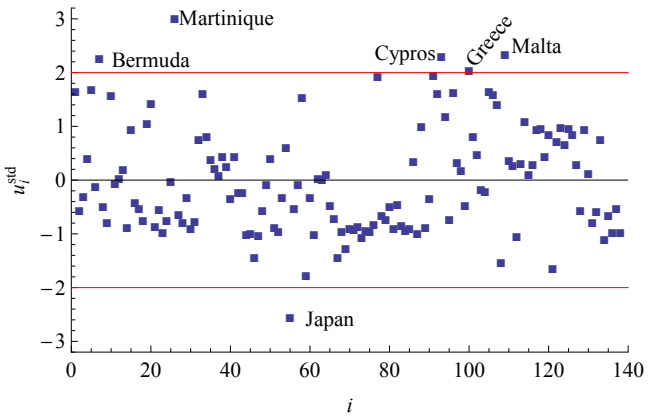
Näide 9.10. SKP ja lauatelefonide arv

Kasutades Rahvusvahelise Telekommunikatsiooni Ühingu ITU[a] andmeid 1998. aastast, leiti lineaarne mudel, kuidas lauatelefonide arv erinevates riikides sõltub vastava riigi sisemajanduse koguproduktist elaniku kohta. Riike oli kokku 138. Mudeliks saadi

$$\hat{y} = 0{,}0017x + 10{,}2, \quad R^2 = 0{,}75,$$

kus x on SKP elaniku kohta (USD) ja y lauatelefonide arv 100 elaniku kohta. Riikides, kus SKP elaniku kohta on 1000 dollari võrra suurem, on telefoniliinide arv 100 elaniku kohta 1,7 võrra suurem.

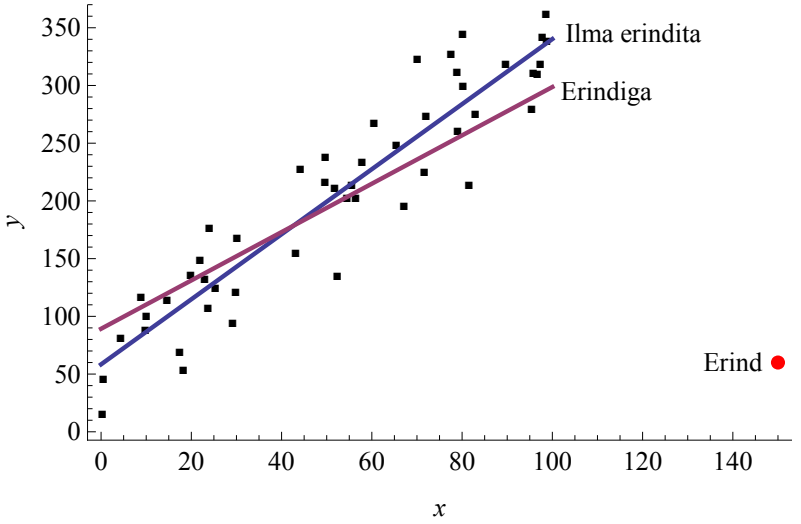
Jääkide analüüs näitab, et kuus riiki on ebatüüpilised: Bermuuda, Martinique, Küpros, Kreeka, Malta ja Jaapan (vt joonis 9.24). Nende riikide korral on standardiseeritud jäägi absoluutväärtus suurem kui 2. Nendest esimese viie korral on jääk positiivne, mis tähendab, et tegelik lauatelefonide arv on suurem kui mudel prognoosis. Jaapani korral on aga jääk negatiivne, järelikult seal oli tegelik lauatelefonide arv mudelväärtusest väiksem. Üheks põhjuseks võib olla see, et kuni 1989. aastani tegutses Jaapani telekommunikatsiooniturul monopoolne ettevõte Nippon Telegraph and Telephone Corp. ja lauatelefoni liitumismaks ning kõnetariif oli väga kõrge.



Joonis 9.24. SKP ja lauatelefonide arv, standardiseeritud jääkide diagramm. Lisatud on standardiseeritud jääkidele -2 ja $+2$ vastavad horisontaalsed jooned. Horisontaalteljel on valimis oleva riigi järjenumber

[a] *International Telecommunication Union*, `http://www.itu.int`

Joonisel 9.25 on 50 punktist koosnev punktiparv, kus üks punkt asub teistest oluliselt eemal — see on erind. Kui regressioonanalüüsis kasutada kõiki punkte, saadakse mudeliks $\hat{y} = 2{,}1x + 89$ (erindiga). Kui aga viia läbi regressioonanalüüs, kus erind on välja jäetud, saadakse mudeliks $\hat{y} = 2{,}8x + 58$ (ilma erindita). Nagu näeme, võib üks erind oluliselt mõjutada parameetrite hinnanguid.



Joonis 9.25. Kaks regressioonjoont: erindiga ja ilma erindita

Vähimruutude meetodil leitud mudeli parameetrid on tundlikud erindite suhtes.

Kui on alust arvata, et erind ei ole tingitud juhusest, tuleks leida ka regressioonmudel ilma erindita. Erind ei pruugi mingitel põhjustel sobida teiste valimis olevate objektide hulka.

9.9. Mitmene regressioon

Reaalses elus mõjutab mingit suurust enamasti palju erinevaid seletavaid tunnuseid:

- käive võib sõltuda müüdava toote hinnast, turunduskuludest, tarbija sissetulekust, konkureerivate kaupade hinnast jne;
- tööjõu tootlikkus võib sõltuda tootmisvahendite vanusest, põhivarasse tehtud investeeringutest ühe töötaja kohta, töötajate haridustasemest jne;
- kinnisvaratehingute arv võib sõltuda kinnisvara hinnast, eluasemelaenude intressimäärast, üldisest majandusolukorrast.

Reaalsust võimalikult hästi kirjeldavate seoste saamiseks tuleb kasutada mitme argumenttunnusega mudeleid, kus üks suurus Y sõltub k argumendist X_l:

$$y = f(x_1, x_2, \ldots, x_k).$$

Mitme argumenttunnusega regressioonmudeli hindamine on **mitmene regressioon**. Kõige sagedamini kasutatakse lineaarset regressioonmudelit:

$$y = b + a_1x_1 + a_2x_2 + ... + a_kx_k + \varepsilon, \tag{9.55}$$

mida võib nimetada mitmeseks lineaarseks regressioonmudeliks ja kus

y on sõltuv ehk funktsioontunnus;
$x_1, x_2, \ldots, x_k$ on argumenttunnused ehk sõltumatud tunnused ehk regressorid;
$b, a_1, \ldots, a_k$ on mudeli parameetrid;
ε on juhuslik liige.

Lineaarse mudeli korral võib parameetreid $a_1, \ldots, a_k$ nimetada **kordajateks** (*coefficient*): a_1 on tunnuse X_1 kordaja, a_2 on tunnuse X_2 kordaja jne. Parameeter b on vabaliige (*intercept*) ehk konstantne liige. Parameetrite tõlgendus mudelis (9.55) on analoogne tõlgendusega ühe argumendiga lineaarse mudeli korral:

Mitmese lineaarse mudeli parameetrite tõlgendus

- b võrdub y väärtusega, kui kõik argumendid on nullid;
- a_1 näitab, kui palju muutub y, kui x_1 suureneb 1 võrra ja teised argumendid jäävad samaks;
- a_2 näitab, kui palju muutub y, kui x_2 suureneb 1 võrra ja teised argumendid jäävad samaks;

- jne.

Tuleb tähele panna, et parameetrite tõlgendamisel lubatakse suureneda ainult ühel argumenttunnusel korraga. See tähendab, et kordaja a_l ($l = 1, \ldots, k$) näitab ainult argumenttunnuse x_l suurenemisest põhjustatud funktsioontunnuse muutust. Majandusteoorias tuntakse sellist tingimust nimetuse all *ceteris paribus* (muidu võrdsetel tingimustel). Matemaatikas vastab sellele osatuletis. Kui y on funktsioon kujul (9.55), siis mudeli parameeter a_l on osatuletis tunnuse x_l järgi:

$$a_l = \frac{\partial y}{\partial x_l}.$$

Näide 9.11. Ravimipoe käibe mudel

N09regressioon
N9.11

Walgreens on USA suurim ravimimüügi kett, millel on üle 8000 ravimipoe kõigis USA osariikides. Kasutades 27 juhuslikult väljavalitud poe andmeid, saadi poe aastakäivet kirjeldav regressioonmudel[a]:

$$\hat{y} = -18{,}9 + 16{,}2x_1 + 0{,}175x_2 + 11{,}5x_3 + 13{,}6x_4 - 5{,}31x_5, \qquad R^2 = 0{,}993,$$

kus

y on netokäive aastas, tuhat \$;
x_1 on poe pindala, tuhat ruutjalga;
x_2 on varude maksumus, tuhat \$;
x_3 on reklaamikulud aastas, tuhat \$;
x_4 on piirkonnas elavate perede arv, tuhat;
x_5 on piirkonnas tegutsevate konkurentide arv.

Determinatsioonikordaja näitab, et selle mudeliga on kirjeldatud 99,3% käibe varieerumisest. Tõlgendame mudeli parameetreid:

- poel, mille pindala (x_1) on tuhande ruutjala võrra suurem, on 16,2 tuhande dollari võrra suurem netokäive aastas;
- poel, kus varude maksumus (x_2) on tuhande dollari võrra suurem, on 0,175 tuhande dollari võrra suurem netokäive aastas;
- reklaamikulude (x_3) suurendamine tuhande dollari võrra suurendab käivet 11,5 tuhande dollari võrra;
- poel, mille teeninduspiirkonnas elab tuhat peret rohkem (x_4), on 13,6 tuhande dollari võrra suurem netokäive aastas;
- piirkonnas tegutsevate konkurentide arvu (x_5) suurenemine 1 võrra vähendab netokäivet 5,31 tuhat dollarit aastas.

Mudeli parameetrite märgid vastavad äritegevuse loogikale: konkurentide arvu suurenemine vähendab käivet, kõik ülejäänud tunnused on positiivse mõjuga.

[a] A Statistical Study of Walgreens, `http://www.oocities.org/tye45/seo.htm`

Seda, millised tulevad mudeli parameetrite arvväärtused, on eelnevalt raske hinnata. Küll aga võib püstitada hüpoteesi parameetrite märgi kohta: millised tunnused mõjutavad funktsioontunnust positiivselt, millised negatiivselt. Regressioonmudeli parameetrite märgid peavad olema loogilised.

Mitmese regressioonülesande korral kasutatakse mudeli parameetrite leidmiseks samuti vähimruutude meetodit: minimeeritakse hälvete ruutude summat. Valemid parameetrite leidmiseks kahe argumenttunnuse korral on esitatud lisas A.11. Nagu valemitest (A.67) ja (A.68) näha, sõltub tunnuse X_1 kordaja a_1 nii tunnuse X_1 kui ka tunnuse X_2 väärtustest. Samamoodi sõltub tunnuse X_2 kordaja a_2 mõlema argumenttunnuse väärtusest.

Kui meil on mudelis k argumenttunnust, siis tuleb leida $k+1$ parameetrit: iga tunnuse kordaja pluss vabaliige. Üldjuhul kasutatakse parameetrite valemite leidmiseks maatriksarvutust, mis võimaldab arvutusi kompaktselt läbi viia. Ka üldisel juhul k argumenttunnuse korral sõltub iga tunnuse ees olev kordaja kõigi tunnuste väärtustest valimisse kuuluvatel objektidel.

N09Regressioon
N9.12,16,17

Näide 9.12. Tööjõu pakkumine

1966. aastal viidi USA-s läbi tööjõu pakkumise uuring. Valimisse võeti 6000 leibkonda, kus perekonnapeaks oli mees ja aastane sissetulek alla 15 000 dollari. Registreeriti järgmised tunnused: küsitletava töötundide arv aastas, tunnitasu, abikaasa aastatulu, teiste pereliikmete aastatulu, mittetöine tulu, perekonna varaline seis (pangaarved jms), küsitletava vanus, ülalpeetavate arv, rass ja haridustase. Nende andmete alusel grupeeriti küsitletud 39 gruppi ja leiti vastavate suuruste keskmised iga grupi jaoks. Eesmärgiks oli leida, kas tunnitasu suurenemine mõjutab positiivselt tööjõu pakkumist (keskmist töötundide arvu aastas) ja kuidas mõjutavad seda teised suurused. (Greenberg ja Kosters, 1970)

Kui uuriti, kuidas töötatud tundide arv aastas (TUNNID) sõltub tunnitasust (TTASU, \$), saadi mudeliks

$$\widehat{\text{TUNNID}} = 1913 + 80{,}9 \cdot \text{TTASU}, \quad R^2 = 0{,}333. \tag{9.56}$$

Mudel näitab, et tunnitasu tõustes ühe dollari võrra suurenes tööjõu pakkumine 80,9 tunni võrra aastas. Kuid mudeli kirjeldusvõime on väike, determinatsioonikordaja ainult 0,333.
Seejärel lisati mudelisse ka majapidamise keskmine varade suurus (VARAD, \$) ja inimese vanus ning otsiti mudelit kujul

$$\text{TUNNID} = b + a_1 \cdot \text{TTASU} + a_2 \cdot \text{VARAD} + a_3 \cdot \text{VANUS} + \varepsilon.$$

Peale regressioonanalüüsi läbiviimist saadi mudeliks

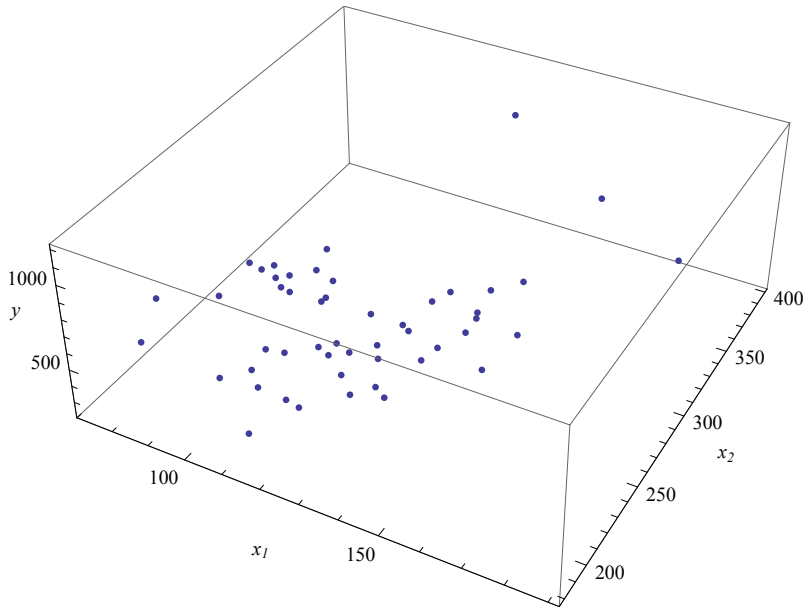
$$\begin{aligned} \widehat{\text{TUNNID}} = {} & 2444{,}8 - 47{,}6 \cdot \text{TTASU} + 0{,}02641 \cdot \text{VARAD} - \\ & - 8{,}66 \cdot \text{VANUS}, \\ & R^2 = 0{,}715. \end{aligned} \tag{9.57}$$

Paneme tähele, et selles mudelis on tunnitasu ees olev kordaja negatiivne — suurema tunnitasu korral töötatud tundide arv väheneb. See on loogiline, sest suurema tunnitasu korral tuleb teatud elustandardi saavutamiseks teha vähem tööd.
Aga kuidas me teame, et usaldusväärsem on mudel (9.57)? Determinatsioonikordaja näitab, et selle mudeli kirjeldusvõime on oluliselt suurem kui mudelil (9.56). Kuid selle järgi otsustamine pole õige. Tuleb kasutada korrigeeritud determinatsioonikordajat, millega tutvume järgmises alapeatükis.

Näitest 9.12 nägime, et kui me jätame mudelist välja olulised tunnused, siis võime mudelisse võetud tunnuste jaoks saada valed parameetrite hinnangud ja teha valesid järeldusi. Ka siis, kui meie eesmärgiks on uurida ainult ühe argumenttunnuse mõju, peame mudelisse võtma kõik olulised tunnused, mis võivad funktsioontunnust mõjutada. Aga millised tunnused on olulised? Seda vaatame alapeatükis 9.12.

Kaht ja rohkemat argumenttunnust sisaldava regressioonmudeli graafiline esitamine on komplitseeritud. Kahe argumenttunnusega lineaarsele mudelile vastab kahemõõtmeline tasand kolmemõõtmelises ruumis. Kui argumenttunnuseid on k tükki, siis vastab lineaarsele mudelile k-mõõtmeline tasand $k+1$-mõõtmelises ruumis.

Joonisel 9.26 on kujutatud 50 punkti kolmemõõtmelises teljestikus.

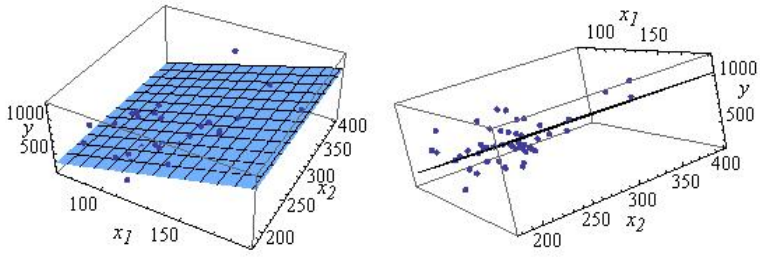


Joonis 9.26. 50 punkti kolmememõõtmelises teljestikus (x_1, x_2, y)

Nende punktide koordinaatide põhjal läbiviidud regressioonanalüüs annab meile tasandi võrrandi:

$$\hat{y} = -13{,}4 + 5{,}7x_1 - 0{,}095x_2.$$

Joonisel 9.27 on see tasand kahe erineva nurga alt.



Joonis 9.27. Tasand $\hat{y} = -13{,}4 + 5{,}7x_1 - 0{,}095x_2$ kahe erineva nurga alt

Programmis Excel tuleb mitmese lineaarse regressioonmudeli parameetrite leidmiseks kasutada vahendit *Regression* komplektist *Data Analysis* (vt lisa C.9). Tunnuse X piirkonnana (*Input X Range*) märgitakse ära kõikide mudelisse võetavate argumenttunnuste väärtused, soovitatavalt koos veergude pealkirjadega (*Labels*). Vastavad veerud peavad asuma kõrvuti. Teine võimalus on kasutada funktsiooni **LINEST**, mille kohta võib täpsemat juhendit vaadata lisast C.10.

9.10. Korrigeeritud determinatsioonikordaja

Näites 9.12 oli toodud kaks lineaarset regressioonmudelit: ühe argumenttunnusega mudel (9.56) ja kolme argumenttunnusega mudel (9.57). Nende mudelite võrdlemiseks ei tohi kasutada determinatsioonikordajat R^2, sest see võib anda valeinformatsiooni. Nimelt suureneb determinatsioonikordaja alati, kui mudelisse lisada uusi tunnuseid. Suureneb ka siis, kui need tunnused pole uuritava seose seisukohalt olulised. Suvalise juhusliku suuruse lisamine mudelisse suurendab determinatsioonikordajat. Nii võib mudelisse lisada kõikvõimalikke suurusi, mis pole uuritava nähtusega üldse seotud, kuigi determinatsioonikordaja järgi mudeli kirjeldusvõime üha paraneb.

Korrigeeritud determinatsioonikordaja

Selle puuduse kõrvaldamiseks on konstrueeritud **modifitseeritud** ehk **korrigeeritud determinatsioonikordaja** (*Adjusted R Squared*), mis arvestab ka argumenttunnuste arvu:

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}, \tag{9.58}$$

kus n on valimi maht ja k mudelis olevate argumenttunnuste arv. Valemi tuletuskäik on toodud lisas A.12.

Näites 9.12 leitud mudeli (9.56) korral $R^2 = 0{,}333$, $n = 39$ ja $k = 1$. Arvutus valemi (9.58) põhjal:

$$R_a^2 = 1 - (1 - 0{,}333)\frac{39-1}{39-1-1} \approx 0{,}315.$$

Mudeli (9.57) korral $R^2 = 0{,}715$, $n = 39$ ja $k = 3$. Arvutus valemi (9.58) põhjal:

$$R_a^2 = 1 - (1 - 0{,}715)\frac{39-1}{39-3-1} \approx 0{,}691.$$

Näeme, et tunnuste VARAD ja VANUS lisamine mudelisse (9.56) parandas mudelit, sest korrigeeritud determinatsioonikordaja suurenes.

Korrigeeritud determinatsioonikordajat R_a^2 ei saa samamoodi interpreteerida nagu determinatsioonikordajat R^2. Kui R^2 näitab mudeli kirjeldusvõimet, siis korrigeeritud determinatsioonikordaja R_a^2 on vaid näitaja, mille abil saab võrrelda erinevat arvu regressoreid sisaldavaid mudeleid.

Korrigeeritud determinatsioonikordaja kasutamine

Erinevat arvu argumenttunnuseid sisaldavate mudelite võrdlemiseks tuleb kasutada **korrigeeritud determinatsioonikordajat**.

- Kui tunnuse lisamisel mudelisse korrigeeritud determinatsioonikordaja suureneb, on selle tunnuse lisamine õigustatud.
- Kui tunnuse lisamisel korrigeeritud determinatsioonikordaja väheneb, ei ole selle tunnuse lisamine õigustatud.

Korrigeeritud determinatsioonikordaja on väiksem kui determinatsioonikordaja ja võib olla ka negatiivne, kui determinatsioonikordaja on väga väike. Näiteks kui $R^2 = 0{,}2$, $k = 3$ ja $n = 10$, siis valemist (9.58) saame, et $R_a^2 = -0{,}2$.

Põhimõtteliselt võib sõltumatuid tunnuseid olla kuitahes palju, praktikas püütakse siiski vältida liiga paljude muutujate kasutamist.

N09Regressioon
N9.13,20

Näide 9.13. Käibe mudel

Ettevõtte analüütikud soovivad leida mudelit ühe toote müügikäibe prognoosimiseks. Arvatakse, et käive (tuhat eurot) võiks sõltuda toote hinnast P (eurodes), turunduskuludest M (tuhat eurot), majanduskasvu indeksist E ja tooteühiku kulude indeksist C. On olemas andmed nende suuruste kohta 20 perioodil. Hinnati nelja erinevat regressioonmudelit. Esimeses mudelis võeti argumendiks ainult majanduskasvu indeks E, teises mudelis lisati sellele turunduskulud M, seejärel lisati ühiku kulude indeks C ja lõpuks hind P. Iga mudeli korral kirjutati välja determinatsioonikordaja R^2 ning korrigeeritud determinatsioonikordaja R_a^2. Mudelite hindamise tulemused on esitatud tabelis.

	Mudel	R^2	R_a^2
1	$\hat{Y} = -3989 + 50{,}13E$	0,716	0,700
2	$\hat{Y} = -3527 + 42{,}9E + 332{,}9M$	0,885	0,872
3	$\hat{Y} = -2642 + 40{,}10E + 469{,}7M - 6{,}29C$	0,906	0,888
4	$\hat{Y} = -2639 + 39{,}46E + 478{,}5M - 4{,}35C - 68{,}31P$	0,908	0,883

Nagu näeme, suurenes determinatsioonikordaja R^2 iga kord, kui mudelisse lisati mõni uus tunnus. Korrigeeritud determinatsioonikordaja R_a^2 suurenes, kui esimesse mudelisse lisati tunnus M, seejärel ka tunnuse C lisamisel. Tunnuse P lisamisel aga korrigeeritud determinatsioonikordaja vähenes. Järelikult pole selle tunnuse lisamine enam õigustatud ning parimaks mudeliks on

kolmas mudel, mis sisaldab majanduskasvu indeksit E, turunduskulusid M ning ühiku kulu indeksit C:

$$\hat{Y} = -2642 + 40{,}10E + 469{,}7M - 6{,}29C.$$

Nende suurustega on käibe Y muutumisest ära seletatud 90,6%. Kui majanduskasvu indeks E suureneb 1 punkti võrra, siis käive suureneb 40,1 tuhat eurot. Turunduskulude suurendamine tuhande euro võrra suurendab käivet 469,67 tuhat eurot. Ühiku kulu indeksi C suurenemine 1 punkti võrra vähendab käivet 6,29 tuhande euro võrra.

9.11. Regressioonmudeli statistiline olulisus

Olgu meil valimi põhjal leitud lineaarne regressioonmudel funktsioontunnusega Y ja argumenttunnustega X_l. Mingi i-nda objekti jaoks võime siis kirja panna:

$$y_i = b + a_1x_{1i} + a_2x_{2i} + ... + a_kx_{ki} + \varepsilon_i. \tag{9.59}$$

Millal võime öelda, et see mudel kirjeldab piisavalt hästi tunnuse Y käitumist? Determinatsioonikordaja R^2 näitab küll mudeli kirjeldusvõimet, kuid kui suur peab R^2 olema, et võiksime mudeliga rahule jääda? Majandusnähtuste kirjeldamisel kasutatakse tihti mudeleid, mille determinatsioonikordaja on näiteks 0,2 ja 0,4 vahel, mõnikord isegi veel väiksem.

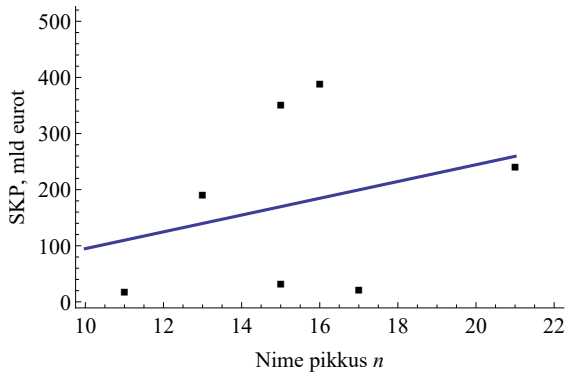
Vaatame näiteks kahte tunnust: Y ja X. Ka siis, kui nende tunnuste vahel puudub igasugune seos, annab vähimruutude meetodi kasutamine mudeli $y = ax+b+\varepsilon$ korral parameetrile a nullist erineva hinnangu.

Näide 9.14. Peaministri nime pikkus ja SKP

Tabelis on toodud 6 riigi peaministri nimed ja vastava riigi SKP (miljard eurot) aastal 2012. Kas on olemas lineaarne seos peaministri nime pikkuse ja riigi SKP vahel?

N09Regressioon
N9.14,15

Riik	Peaminister	Tähtede arv nimes	SKP, mld eurot
Eesti	Andrus Ansip	11	16,0
Läti	Valdis Dombrovskis	17	20,2
Leedu	Andrius Kubilius	15	30,7
Soome	Jyrki Katainen	13	189,4
Rootsi	Fredrik Reinfeldt	16	387,9
Taani	Helle Thorning-Schmidt	21	239,2
Norra	Jens Stoltenberg	15	349,1



Joonis 9.28. Peaministri nimi ja riigi SKP

Otsime seost kujul $SKP = b + an$, kus n on tähtede arv nimes. Regressioonanalüüs annab tulemuseks järgmise mudeli:

$$\widehat{SKP} = -54{,}68 + 14{,}96n, \quad R^2 = 0{,}089.$$

Kas me võime saadud mudeli põhjal väita, et kui riigi peaministri nimi on pikem, on ka riigi SKP suurem? Et iga täht peaministri nimes suurendab riigi SKP-d ligikaudu 15 miljardi euro võrra? Determinatsioonikordaja on küll väike, aga kui väike see olla võib, selle jaoks meil kriteerium puudub.

Et eristada, kas mudel (9.59) väljendab juhuslikku seost või mitte, on vaja teatud kriteeriumi. Kriteeriumi saamiseks kasutatakse regressioonanalüüsi tulemuste dispersioonanalüüsi ehk regressiooni ANOVA-t. Tuletame meelde, et rohkem kui kahe valimi keskmiste võrdlemisel kasutati samuti dispersioonanalüüsi. Selle korral otsiti vastust küsimusele, kas rühmakeskmiste erinevus on põhjustatud uuritava faktori mõjust või valimite juhuslikkusest (vt alapeatükk 7.14).

Regressiooni dispersioonanalüüsi ANOVA korral otsitakse vastust küsimusele, kas funktsioontunnuse Y varieerumine on tingitud argumenttunnuste X_l varieerumisest, mis mõjutab tunnust Y regressioonseose kaudu, või on tunnuse Y varieerumine juhuslik.

Seda, millised suurused kirjeldavad tunnuse Y koguhajuvust, regressioonhajuvust ja jääkhajuvust, vaatasime alapeatükis 9.4, kui tuletasime valemi determinatsioonikordaja jaoks. Seosest tingitud varieerumist kirjeldab regressioonhajuvus SSR (valem (9.31)) ja jääkhajuvust SSE (valem (9.32)). Regressiooni dispersioonanalüüsiks jagatakse need vastava vabadusastmete arvuga ja saadakse **keskruudud**. Summa SSR vabadusastmete arv on k ning summa SSE vabadusastmete arv $n-k-1$, kus k on argumenttunnuste arv.

Regressiooni keskruut (*Mean Square due to Regression*)

$$MSR = \frac{SSR}{k}, \tag{9.60}$$

Regressiooni keskruut MSR

kus regressioonhajuvus $SSR = \sum(\hat{y}_i - \bar{y})^2$ ja k on argumenttunnuste arv regressioonmudelis. Jääkhajuvuse keskruut (*Mean Square due to Errors*)

$$MSE = \frac{SSE}{n-k-1}, \tag{9.61}$$

Jääkhajuvuse keskruut MSE

kus jääkhajuvus $SSR = \sum(y_i - \hat{y}_i)^2$ ja n on valimi maht.

Regressiooni dispersioonanalüüsil lähtutakse järgmistest kaalutlustest:

1) keskruut MSE on regressiooni jääkliikmete u_i dispersiooni σ_u^2 nihketa hinnang;
2) kui mudelis (9.59) on kõik kordajad $a_1, \ldots, a_k$ nullid, siis $MSR = MSE$ ja suhe $MSR/MSE = 1$;
3) kui kõik kordajad ei ole nullid, siis $MSR > MSE$ ning suhe $MSR/MSE > 1$.

Kuna regressioonanalüüsi korral on tegemist valikvaatlusega, võib valimi põhjal leitud suhe MSR/MSE olla ühest suurem ka siis, kui kogumis regressioonseos puudub ja mudeli kordajate tegelikud väärtused on kõik nullid. Kui palju peab see suhe olema ühest suurem, et võiksime öelda „Mudel on statistiliselt oluline“? Otsustamiseks on meil vaja kriteeriumi. Matemaatilisest statistikast on teada, et keskruutude suhe allub F-jaotusele vabadusastmete arvuga $n-1$ ning $n-k-1$:

$$\frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)} \sim F(k, n-k-1).$$

Järelikult saame testimiseks vajaliku kriitilise väärtuse F-jaotusest.

Regressioonmudeli statistilise olulisuse testimine

Regressioonmudeli $y = b+a_1x_1+a_2x_2+...+a_kx_k+\varepsilon$ **statistilise olulisuse** testimine.
Hüpoteesipaar:
H_0: $a_1 = a_2 = \ldots = a_k = 0$, mudel ei ole statistiliselt oluline;
H_1: üks või mitu kordajat a_j on nullist erinevad, mudel on statistiliselt oluline.
Teststatistik

$$F = \frac{MSR}{MSE} \sim F(k; n-k-1), \tag{9.62}$$

kus n on valimi maht ja k argumenttunnuste arv. Tegemist on ühepoolse hüpoteesiga, sest kui regressioonmudel on statistiliselt oluline, siis $MSR > MSE$.
Otsustamine: H_0 tagasi lükata

olulisuse tõenäosuse p kasutamisel,	kui $p < \alpha$;
kriitilise väärtuse kasutamisel,	kui $F > F_{kr}(\alpha)$.

Siin on α olulisuse nivoo, millele vastav kriitiline väärtus $F_{kr}(\alpha)$ on vabadusastmetega k ja $n-k-1$ F-jaotuse α-täiendkvantiil.

Tabelid kriitilise väärtuse $F_{kr}(\alpha)$ määramiseks mõningate vabadusastmete arvu korral on esitatud lisas B.2. Näiteks, kui meil on valim mahuga $n = 23$ ja regressioonmudeli $y = a_0+a_1x_1+a_2x_2+\varepsilon$ hindamisel tuleb $F = 3{,}8$, siis vastav kriitiline väärtus olulisuse nivool $\alpha = 0{,}05$ on 3,49 (vabadusastmete arvud on $k = 2$ ja $n-k-1 = 23-2-1 = 20$). Kuna $3{,}8 > 3{,}49$, on nullhüpotees ümber lükatud ja mudel statistiliselt oluline. Kui valimi maht oleks väiksem, näiteks $n = 13$, aga F väärtus sama, siis kriitiline väärtus on 4,1 ja kehtib nullhüpotees, sest $3{,}8 < 4{,}1$. Mida suurem on valimi maht, seda väiksem on F-testi kriitiline väärtus ja nullhüpoteesi ümberlükkamiseks piisab väiksemast F-statistiku väärtusest.

Tabelarvutuses leiab statistiku F kriitilised väärtused funktsioon **F.INV.RT**, kus *Probability* on olulisuse tõenäosus, *Deg_freedom1* lugeja vabadusastmete arv k ja *Deg_freedom2* nimetaja vabadusastmete arv $n-k-1$.

Alapeatükis 9.3 leiti tabelis 9.3 toodud andmete alusel lineaarse regressioonmudeli parameetrid ja mudeliks saadi $\hat{y}_i = 1{,}9x_i + 4$ (valem (9.11)). Teeme selle mudeli jaoks regressiooni dispersioonanalüüsi arvutused. Jääkhajuvuse $SSE = \sum(y_i - \hat{y}_i)^2$ ja regressioonhajuvuse $SSR = \sum(\hat{y}_i - \bar{y})^2$ leidmiseks leiame vastavad vahed ja vahede ruudud ning arvestame, et aritmeetiline keskmine $\bar{y} = 61$ (tabel 9.8).

Tabel 9.8. Arvutused jääkhajuvuse ja regressioonhajuvuse leidmiseks


N09Regressioon
T9.8

i	x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$
1	10	15	23	−8	64	−38	1444
2	20	50	42	8	64	−19	361
3	30	60	61	−1	1	0	0
4	40	90	80	10	100	19	361
5	50	90	99	−9	81	38	1444

Tabeli kuuenda veeru $(y_i - \hat{y}_i)^2$ summeerimisel saame jääkhajuvuse

$$SSE = \sum (y_i - \hat{y}_i)^2 = 310 \tag{9.63}$$

ja viimase veeru sumeerimisel regressioonhajuvuse

$$SSR = \sum (\hat{y}_i - \bar{y})^2 = 3610. \tag{9.64}$$

Paneme tähele, et jääkhajuvus (9.63) on sama, mis hälvete ruutude summa (9.18).

Kuna mudelis (9.11) on üks argumenttunnus, siis $k = 1$ ning valemist (9.60) saame

$$MSR = \frac{SSR}{k} = \frac{3610}{1} = 3610.$$

Keskruudu MSE leidmiseks arvestame, et punktide arv $n = 5$ ning $n - k - 1 = 5 - 1 - 1 = 3$. Valemist (9.61) saame

$$MSE = \frac{SSE}{n - k - 1} = \frac{310}{3} \approx 103{,}33.$$

Valemist (9.62) leiame F-statistiku:

$$F = \frac{MSR}{MSE} = \frac{3610}{103{,}33} \approx 34{,}94.$$

Vastav kriitiline väärtus olulisuse nivool 0,05 on F-jaotuse täiendkvantiil $F_{0,05}(1, 3) = 10{,}13$, kasutatakse ühepoolset hüpoteesi. Täiendkvantiili võib võtta kas lisa B.2 tabelist või arvutada tabelarvutuses funktsiooniga F.INV.RT(0,05; 1; 3) ≈ 10,13. Kuna 34,94 > 10,13, on nullhüpotees ümber lükatud ja regressioonmudel on statistiliselt oluline.

Sama järelduseni jõuame, kui leiame F-statistikule vastava olulisuse tõenäosuse ja võrdleme seda olulisuse nivooga 0,05. Tabelarvutuses saab F-testi olulisuse tõenäosuse leida funktsiooniga F.DIST.RT. Kuna F.DIST.RT(34,94; 1; 3) = 0,00967 < 0,05, tuleb nullhüpotees tagasi lükata: mudel on statistiliselt oluline.

Tabel 9.9. Arvutatud ANOVA tabel (tabeli 9.8 põhjal)

ANOVA

Varieeruvuse allikas	Vabadus-astmete arv df	Hälvete ruutude summa SS	Keskruut MS	F-statistik
Regressioonhajuvus	1	3610	3610	34,94
Jääkhajuvus	3	310	103,33	
Koguhajuvus	4	3920		

Need arvutustulemused koondatakse ühte tabelisse 9.9, mida nimetatakse regressioonanalüüsi ANOVA tabeliks. Lisatakse ka koguhajuvus $SST = SSR + SST$, mis antud juhul on $3610 + 310 = 3920$.

Selline standardne regressioonanalüüsi ANOVA tabel väljastatatakse regressioonanalüüsi läbiviimisel statistikapakettides ja ka programmis Excel (*Data Analysis, Regression*). F-testi kriitilist väärtust tavaliselt ei väljastata, selle asemel leitakse F-statistikule vastav olulisuse tõenäosus p (*Significance F*), mida tuleb võrrelda olulisuse nivooga α. Tabelis 9.10 on esitatud standardse ANOVA tabeli struktuur regressioonanalüüsi korral.

Regessioon-analüüsi ANOVA tabel

Tabel 9.10. Regressioonanalüüsi ANOVA tabeli struktuur

Varieeruvuse allikas	Vabadus-astmete arv df	Hälvete ruutude summa SS	Keskruut MS	F-statistik	Olulisuse tõenäosus
Regressioon	k	$SSR = \sum(\hat{y}_i - \bar{y})^2$	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$	p
Jääk-hajuvus	$n-k-1$	$SSE = \sum(y_i - \hat{y})^2$	$MSE = \frac{SSE}{n-k-1}$		
Kogu-hajuvus	$n-1$	$SST = \sum(y_i - \bar{y})^2 = SSR + SSE$			

Näide 9.15. Peaministri nime pikkus ja SKP: regressioonmudeli olulisuse testimine

N09Regressioon
N9.14,15

Näites 9.14 toodud andmete põhjal väljastatakse Excelis järgmine regressioonanalüüsi ANOVA tabel:

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	13356	13356	0,489	0,516
Residual	5	136670	27334		
Total	6	150026			

Regressioonmudeli olulisuse kontrollimiseks kasutame olulisuse nivood $\alpha = 0{,}05$. Kuna olulisuse tõenäosus $p = 0{,}516 > 0{,}05$, tuleb jääda nullhüpoteesi juurde ja mudel ei ole statistiliselt oluline. Järelikult, riigi peaministri nime pikkuse ja SKP vahel seos puudub.

Analüüsime, mida tähendab nullhüpoteesi kehtimine regressioonmudeli statistilise olulisuse kontrollimisel. Nullhüpoteesi kehtimise korral, kui mudelis (9.59) $a_1 = a_2 = \ldots = a_k = 0$, saame valimis oleva i-nda objekti jaoks kirja panna seose

$$y_i = b + u_i, \tag{9.65}$$

kus u_i on regressiooni jääkliige. See tähendab, et tunnuse Y väärtused varieeruvad juhuslikult ümber konstandi b. Lihtne on näidata, et see konstant on tunnuse Y aritmeetiline keskmine. Selleks summeerime y_i väärtused üle kõigi objektide:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (b + u_i) = \sum_{i=1}^{n} b + \sum_{i=1}^{n} u_i.$$

Vähimruutude meetodi kasutamisel jääkliikmete summa $\sum_{i=1}^{n} u_i = 0$. Saame, et

$$\begin{aligned} \sum_{i=1}^{n} y_i &= \sum_{i=1}^{n} b = nb \\ \frac{1}{n} \sum_{i=1}^{n} y_i &= b \\ \bar{y} &= b. \end{aligned}$$

Järelikult nullhüpoteesi kehtimise korral varieeruvad tunnuse Y väärtused juhuslikult ümber aritmeetilise keskmise $\bar{y}$.

N09Regressioon
N9.12,16,17

Näide 9.16. Tööjõu pakkumine: regressioonmudeli statistilise olulisuse kontroll

Näites 9.12 analüüsiti, kuidas tööjõu pakkumine (töötundide arv aastas) sõltub tunnitasust, varade suurusest ja töötaja vanusest. Mudelit otsiti kujul

$$\text{TUNNID} = b + a_1 \cdot \text{TTASU} + a_2 \cdot \text{VARAD} + a_3 \cdot \text{VANUS} + \varepsilon. \quad (9.66)$$

Excelis tehtud regressioonanalüüsi tulemuseks saame järgmise aruande (esitame siin kõik kolm osa):

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,846
R Square	0,715
Adjusted R Square	0,691
Standard Error	35,61
Observations	39

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	111392,6	37130,9	29,28	1,18E-09
Residual	35	44390,6	1268,3		
Total	38	155783,2			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2444,8	93,62	26,11	1,54E-24	2254,7	2634,9
TTASU	−47,6	23,01	−2,070	0,0459	−94,319	−0,908
VARAD	0,02641	0,00393	6,720	8,8E-08	0,0184	0,0344
VANUS	−8,66	1,706	−5,078	1,27E-05	−12,13	−5,20

F-testi olulisuse tõenäosus p (*Significance F*) on $1{,}18 \cdot 10^{-9} < 0{,}05$, järelikult on nullhüpotees tagasi lükatud ja mudel on statistiliselt oluline nivool 0,05. Vähemalt üks mudeli kordajatest, kas a_1, a_2 või a_3, on nullist erinev.

Determinatsioonikordaja võib avaldada F-statistiku kaudu:

$$R^2 = \frac{1}{1 + \frac{1}{F}\frac{n-k-1}{k}}. \quad (9.67)$$

Selle valemi ja F-statistiku kriitiliste väärtuste põhjal on võimalik leida kriitilised väärtused determinatsioonikordajale. Näiteks lihtsa reg-

ressioonmudeli $y = ax + b + \varepsilon$ korral, kui valimi maht $n = 5$, on F kriitiline väärtus 10,13 ($k = 1$, $n - k - 1 = 3$). Valemist (9.67) saame

$$R^2_{kr} = \frac{1}{1 + \frac{1}{10{,}13} \cdot \frac{3}{1}} \approx 0{,}772.$$

Järelikult peab sellisel juhul mudeli determinatsioonikordaja olema suurem kui 0,772 ning siis on mudel statistiliselt oluline. Suurema valimi korral on determinatsioonikordaja kriitiline väärtus väiksem. Näiteks valimi mahu $n = 22$ korral piisab determinatsioonikordaja väärtusest 0,179, et ühe argumenttunnusega mudel oleks statistiliselt oluline.

Statistiku F võib avaldada ka determinatsioonikordaja R^2 kaudu:

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}. \tag{9.68}$$

Seda valemit võib kasutada juhul, kui kasutatav tarkvara F-statistikut regressioonanalüüsil ei leia, kuid on soov mudeli statistilist olulisust testida.

9.12. Mudeli parameetrite statistiline olulisus

Kui lineaarse regressioonmudeli statistilise olulisuse kontrollimisel saadakse, et nullhüpotees on ümber lükatud, siis vähemalt üks regressioonmudeli kordajatest $a_1, a_2, \ldots, a_k$ on nullist erinev. Järgnevalt tuleb kontrollida, kas **üksikud** kordajad erinevad statistiliselt oluliselt nullist. Nii selgitatakse välja, ega argumenttunnuste hulgas pole ülearuseid. Kui selgub, et mingi tunnuse kordaja a_j on mitteoluline, ei ole vastava tunnuse X_j mudelisse lülitamine põhjendatud ning see tuleks välja jätta. Vabaliikme a_0 statistilist olulisust tavaliselt ei kontrollita, sest see jäetakse mudelisse ka siis, kui see on statistiliselt mitteoluline. Mudeleid, kus vabaliige puudub, vaatame alapeatükis 9.16.

Parameetrite statistilise olulisuse kontrollimisel lähtutakse nende standardvigadest $se(a_j)$. Üldjuhul, k seletava tunnuse korral, on parameetri a_j standardvea arvutusvalem komplitseeritud ja selle esitamiseks kasutatakse maatriksesitust (vt näiteks (Aarma ja Vensel, 2005, lk 178)). Tarkvara genereeritud regressioonanalüüsi väljundtabelis on parameetrite standardvead veerus *Standard Error*.

Mudeli parameetrite olulisuse kontrollimiseks kasutatakse kahepoolset t-testi, mis viiakse läbi iga parameetri a_j jaoks eraldi. Testimiseks kasutatav t-statistik allub t-jaotusele vabadusastmete arvuga $n - k - 1$.

Regressioonmudeli parameetrite statistilise olulisuse testimine

Regressioonmudeli $y = b+a_1x_1+a_2x_2+...+a_kx_k+\varepsilon$ **parameetri statistilise olulisuse** testimine.
Iga parameetri $a_1, a_2, \ldots, a_k$ korral testitakse hüpoteesipaari:
H_0: $a_j = 0$, parameeter ei ole statistiliselt oluline;
H_1: $a_j \neq 0$, parameeter on statistiliselt oluline.

Teststatistik on

$$t = \frac{a_j}{se(a_j)} \sim t(n-k-1), \tag{9.69}$$

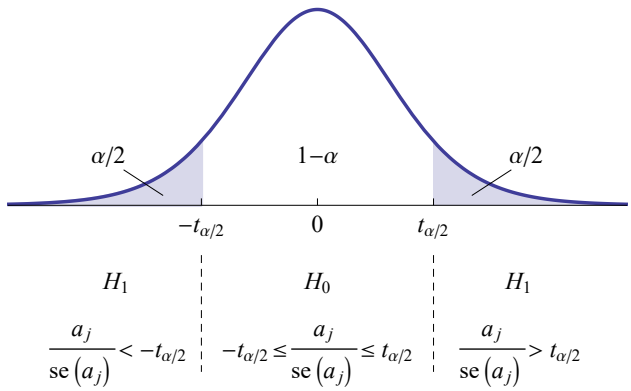
kus $se(a_j)$ on parameetri a_j standardviga, n valimi maht ja k argumenttunnuste arv.

Otsustamine: H_0 tagasi lükata

olulisuse tõenäosuse p kasutamisel,	kui $p < \alpha$;
kriitilise väärtuse kasutamisel,	kui $\lvert t\rvert > t_{\alpha/2}$.

Siin on α olulisuse nivoo, millele vastav kriitiline väärtus $t_{\alpha/2}$ leitakse t-jaotusest vabadusastmete arvuga $n-k-1$.

Kui tegelikus mudelis tunnus X_j puudub, siis selle kordaja on null ja kehtib nullhüpotees. Juhuvalimi põhjal leitud regressioonmudeli korral jääb sel juhul suhe $a_j/se(a_j)$ vahemikku $[-t_{\alpha/2}, t_{\alpha/2}]$ tõenäosusega $1-\alpha$ (joonis 9.29). Sellisel juhul võetakse nullhüpotees vastu. Kui aga $|a_j/se(a_j)| > t_{\alpha/2}$ (positiivse a_j korral $a_j/se(a_j) > t_{\alpha/2}$ või negatiivse a_j korral $a_j/se(a_j) < -t_{\alpha/2}$), siis on vähetõenäoline, et vastav tunnus tegelikus mudelis puudub. Sellisel juhul lükatakse nullhüpotees tagasi ja võetakse vastu sisukas hüpotees H_1.



Joonis 9.29. Millal võetakse vastu nullhüpotees H_0 ja millal sisukas hüpotees H_1

Kui mõne kordaja puhultuleb vastu võtta nullhüpotees, siis pole vastava argumenttunnuse mudelisse kaasamine õigustatud. Sellisel juhul viiakse läbi uue regressioonmudeli hindamine, millest mitteoluline tunnus on välja jäetud.

Tabelid kriitilise väärtuse $t_{\alpha/2}$ määramiseks mõningate vabadusastmete arvude korral on toodud lisas B.1. Näiteks kui meil on valim mahuga $n = 10$ ja regressioonmudeli $y = a_0+a_1x_1+a_2x_2+\varepsilon$ hindamisel tuleb parameetri a_1 korral statistik $t = 2{,}9$, siis vastav kriitiline väärtus olulisuse nivool $\alpha = 0{,}05$ on 2,36: $\alpha/2 = 0{,}025$ annab meile veeru ja vabadusastmete arv $n - k - 1 = 10 - 2 - 1 = 7$ rea. Kuna $2{,}9 > 2{,}36$, on nullhüpotees ümber lükatud ja parameeter a_1 statistiliselt oluline.

Tabelarvutuses saab kahepoolse hüpoteesi korral leida t-statistiku kriitilised väärtused $t_{\alpha/2}$ funktsiooni T.INV.2T abil. Näiteks kui vabadusastmete arv on 7, siis kahepoolse t-testi kriitiline väärtus olulisuse nivool 0,05 on T.INV.2T(0,05;7)= 2,3646. Olulisuse nivool 0,01 saame aga kriitiliseks väärtuseks T.INV.2T(0,01;7)= 3,5.

Näide 9.17. Tööjõu pakkumise mudeli parameetrite olulisus

N09Regressioon
N9.12,16,17

Näites 9.16 testisime tööjõu pakkumise mudeli statistilist olulisust. Valimi maht oli 39 ja argumenttunnuste arv 3. Teme nüüd arvutused parameetrite statistilise olulisuse testimiseks. Paneme tähele, et t-testi kriitiline väärtus on kõigi parameetrite jaoks ühesugune, sest vabadusastmete arv on ühesugune: $39-3-1 = 35$, ja olulisuse nivoo võtame ka sama, $\alpha = 0{,}05$. Vastava kriitilise väärtuse leiame tabelarvutusest: T.INV.2T(0,05; 35)= 2,03. Esitame siin uuesti programmis Excel väljastatud kordajate tabeli.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	2444,8	93,62	26,11	1,54E-24
TTASU	−47,6	23,01	−2,070	0,0459
VARAD	0,02641	0,00393	6,720	8,8E-08
VANUS	−8,66	1,706	−5,078	1,27E-05

Leiame iga tunnuse kordaja jaoks valemi (9.69) põhjal t-statistiku väärtuse ja võrdleme selle absoluutväärtust kriitilise väärtusega:

$$\text{TTASU} \quad \frac{-47{,}6}{23{,}01} \approx -2{,}07, \quad |-2{,}07| > 2{,}03, \quad \text{võtta vastu } H_1;$$

$$\text{VARAD} \quad \frac{0{,}02641}{0{,}00393} \approx 6{,}72, \quad |6{,}72| > 2{,}03, \quad \text{võtta vastu } H_1;$$

VANUS $\dfrac{-8{,}66}{1{,}706} \approx -5{,}078, \quad |-5{,}078| > 2{,}03, \quad$ võtta vastu H_1.

Praktikas pole vaja neid arvutusi teha, t-statistik on leitud Exceli väljastatud regressioonanalüüsi tabelis veerus *t Stat*. Lisaks on veerus *P-value* leitud ka t-statistikule vastav olulisuse tõenäosus p ning teine võimalus hüpoteesi kontrollimiseks on selle võrdlemine olulisuse nivooga α. Kui võtame $\alpha = 0{,}05$, siis

TTASU	$p = 0{,}0459 < 0{,}05,$	võtta vastu H_1;
VARAD	$p = 8{,}8 \cdot 10^{-8} < 0{,}05,$	võtta vastu H_1;
VANUS	$p = 1{,}27 \cdot 10^{-5} < 0{,}05,$	võtta vastu H_1.

Mida teha aga siis, kui mõne tunnuse t-testi korral tuleb vastu võtta nullhüpotees ning vastav tunnus ei ole statistiliselt oluline? Sellisel juhul ei ole selle tunnuse mõju funktsioontunnusele tõestatud ning vastav argumenttunnus tuleb mudelist välja jätta.

Näide 9.18. Sigaretid

N09Regressioon
N9.18,23

USA Riiklik Kaubanduskomisjon (FTC, *Federal Trade Commission*) on mitmel korral määranud USA-s toodetavate sigarettide tõrva- ja nikotiinisisaldust ning sigareti tõmbamisel õhku eralduva süsinikoksiidi (CO) kogust. Kas süsinikoksiidi kogus sõltub tõrva ja nikotiini hulgast sigaretis? Selle seose uurimiseks tuleb püstitada lineaarne mudel:

$$y = b + a_T x_T + a_N x_N + \varepsilon, \tag{9.70}$$

kus

y on lenduva süsinikoksiidi kogus (mg),
x_T on sigaretis oleva tõrva kogus (mg),
x_N on sigaretis oleva nikotiini kogus (mg).

Kasutame 25 USA-s toodetava sigareti andmeid, mis on võetud FTC aruandest (Mendenhall ja Sincich, 1993). Excelis tehtud regressioonanalüüsi tulemused:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,958
R Square	0,919
Adjusted R Square	0,911
Standard Error	1,41
Observations	25

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	495,3	247,6	124,1	1,04E-12
Residual	22	43,9	2,0		
Total	24	539,2			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	3,090	0,844	3,662	0,0014	1,3397	4,8395
xT	0,962	0,237	4,067	0,0005	0,4717	1,4533
xN	−2,646	3,787	−0,699	0,4920	−10,5005	5,2079

Nagu ANOVA tabelist näeme, on mudel tervikuna statistiliselt oluline: F-testi olulisuse tõenäosus on $1{,}04 \cdot 10^{-12} < 0{,}05$ ja tuleb vastu võtta sisukas hüpotees. Koefitsientide veeru põhjal võime kirja panna mudeli:

$$\hat{y} = 3{,}09 + 0{,}962 x_T - 2{,}646 x_N. \tag{9.71}$$

Mudeli tõlgendamisel aga tekib probleem: nikotiini hulga suurenemisel süsinikoksiidi kogus väheneb, mis pole loogiline.
Kontrollime t-testiga argumenttunnuste statistilist olulisust. Hüpoteesipaar iga parameetri a_j korral:

H_0: $a_j = 0$;
H_1: $a_j \neq 0$.

Otsustamiseks kasutame t-statistikute olulisuse tõenäosust (veerg *P-value*):

xT $p = 0{,}0005 < 0{,}05$, võtta vastu H_1, on statistiliselt oluline;
xN $p = 0{,}4920 > 0{,}05$, võtta vastu H_0, ei ole statistiliselt oluline.

Nikotiini kordaja ei ole nivool 0,05 statistiliselt oluline. Järelikult tuleb nikotiini kogus mudelist välja jätta. Uus mudel sisaldab ainult üht sõltumatut tunnust, tõrvasisaldust x_T:

$$y = b + a_T x_T + \varepsilon. \tag{9.72}$$

Regressioonanalüüs annab järgmise tulemuse:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,957
R Square	0,917
Adjusted R Square	0,913
Standard Error	1,397
Observations	25

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	494,3	494,3	253,4	6,55E-14
Residual	23	44,9	1,95		
Total	24	539,2			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2,743	0,675	4,06	0,00048	1,347	4,140
xT	0,801	0,050	15,92	6,55E-14	0,697	0,905

Mudelisse jäänud tõrvasisalduse kordaja on statistiliselt oluline: xT $p = 6{,}55 \cdot 10^{-14} < 0{,}05$, võtta vastu H_1, on oluline.

Veidi paranes ka korrigeeritud determinatsioonikordaja (*Adjusted R Square*): nüüd on see 0,913, mudeli (9.71) korral oli aga 0,911. Lõplik mudel on

$$\hat{y} = 2{,}74 + 0{,}801 x_T, \quad R^2 = 0{,}917,$$

kus y on süsinikoksiidi sisaldus sigaretisuitsus (mg) ja x_T sigareti tõrvasisaldus (mg). Mudel kirjeldab 91,7% süsinikoksiidi koguse varieerumisest erinevate sigarettide suitsus. Tõrvasisalduse tõus 1 mg võrra suurendab lenduvas suitsus oleva süsinikoksiidi sisaldust 0,801 mg võrra.

Mitteoluliste tunnuste eemaldamine

Kui regressioonmudeli parameetrite testimisel selgub, et mõne tunnuse kordaja on statistiliselt mitteoluline (võetakse vastu H_0), tuleb teha uus regressioonmudeli hindamine, millest vastav tunnus on välja jäetud. Korrektses regressioonmudelis peavad kõikide tunnuste kordajad olema statistiliselt olulised uurija valitud nivool (tavaliselt 0,05 või 0,1). Vabaliige võib olla statistiliselt mitteoluline (vt ka alapeatükk 9.16).

Vabaliige peaks olema statistiliselt oluline siis, kui sel on kindel sisu ja me tahame selle väärtust kasutada. Näiteks lineaarses kulufunktsioonis $C(q) = aq + C_F$ on vabaliige C_F püsikulud. Püsikulude määramiseks peab vabaliige olema statistiliselt oluline.

Mudeli parameetrite statistilise olulisuse testimine võimaldab välja selekteerida need argumenttunnused, mis funktsioontunnust reaalselt mõjutavad. Andmete kogumise eel peab uurija püstitama hüpoteesi,

millised suurused võivad mõjutada analüüsitavat suurust Y. Seejärel kogutakse valikvaatluse abil vastavad andmed. Regressioonanalüüsi käigus tehakse hüpoteesiga püstitatud argumenttunnuste hulgast valik: proovitakse läbi erinevate tunnuste komplektidega mudelid ja lõplikku mudelisse jäetakse vaid need argumenttunnused, mis on statistiliselt olulised uurija võetud nivool α. Tavaliselt on selleks 0,05, kuid mõningatel juhtudel kasutatakse ka nivood 0,1. Seda siis, kui näiteks teooriast on teada, et antud suurus peab mudelis olema, kuid nivool 0,05 see statistiliselt oluline ei ole.

Näide 9.19. Tööjõu pakkumine ja väike valim

N09Regressioon
N9.19

Näites 9.17 testisime tööjõu pakkumise mudeli argumenttunnuste TTASU, VARAD ja VANUS statistilist olulisust ning saime, et kõik tunnused olid olulised. Valimi maht oli 39. Viime uuesti läbi mudeli

$$\text{TUNNID} = b + a_1 \cdot \text{TTASU} + a_2 \cdot \text{VARAD} + a_3 \cdot \text{VANUS} + \varepsilon$$

hindamise, kuid kasutame seekord väiksemat valimit. Selleks teeme 39 objekti hulgast juhuvalimi mahuga 10. Mudeli hindamisel Excelis saame parameetrite jaoks järgmise tulemuse:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	2418	286	8,46	0,000149
TTASU	−29,3	58,3	−0,502	0,634
VARAD	0,0260	0,0137	1,90	0,106
VANUS	−9,46	5,46	−1,73	0,134

Nagu näeme veerust *P-value*, on nüüd tunnustele TTASU, VARAD ja VANUS vastavad olulisuse tõenäosused suuremad kui olulisuse nivoo 0,05 ning kõigi korral võtame vastu nullhüpoteesi: tunnus ei ole statistiliselt oluline. Järelikult ei õnnestu väiksema valimi korral tõestada, et need tunnused mõjutavad töötatud tundide arvu TUNNID.
Võrdleme seda tabelit valimi mahu 39 korral saadud tabeliga näitest 9.17:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	2444,8	93,62	26,11	1,54E-24
TTASU	−47,6	23,01	−2,070	0,0459
VARAD	0,0264	0,0039	6,720	8,8E-08
VANUS	−8,66	1,7060	−5,078	1,27E-05

Näeme, et suurema valimi korral on standardvead väiksemad ja sealt tulenevalt on t-statistikud kaugemal nullist. Lisaks mõju-

tab hüpoteesi kontrollimise tulemust see, et t-statistiku kriitiline väärtus on suurema valimi korral väiksem: valimi mahu 39 korral 2,03, kuid valimi mahu 10 korral 2,45.

Kui regressioonmudeli argumenttunnuse statistilise olulisuse kontrollimisel selgub, et tuleb vastu võtta nullhüpotees, siis sellel võib olla kaks põhjust:

1) vastav tunnus ei mõjuta funktsioontunnust;
2) valim on liiga väike ning mõju pole võimalik tõestada.

Analüüsime veel näites 9.18 toodud Exceli regressioonanalüüsi aruannet, mis saadi mudeli (9.70) hindamisel.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	3,090	0,844	3,662	0,0014	1,3397	4,8395
xT	0,962	0,237	4,067	0,0005	0,4717	1,4533
xN	−2,646	3,787	−0,699	0,4920	−10,5005	5,2079

Tabeli viimases kahes veerus on toodud usaldatavusele 95% vastava usaldusvahemiku alumine (*Lower*) ja ülemine (*Upper*) piir iga parameetri jaoks. Paneme tähele, et selle parameetri jaoks, mis ei ole statistiliselt oluline (tunnus xN), on alumine ja ülemine piir vastandmärgilised. See tähendab, et vastav usaldusvahemik sisaldab arvu 0 (vt ka joonist 9.30). See vastab t-testi tulemusele: nullhüpoteesi kehtimine tähendab, et $a_N = 0$.

Joonis 9.30. Arv 0 jääb nikotiini koguse usaldusvahemiku sisse. Järelikult, nikotiini koguse kordaja võib olla null, mis tähendab, et see tunnus pole oluline

Parameetrite statistiline olulisus ja usalduspiirid

Nullhüpoteesi vastuvõtmine regressioonmudeli parameetri statistilise olulisuse kontrollimisel t-testi abil tähendab, et parameetri usaldusvahemik sisaldab nullpunkti.

Järelikult võime parameetrite statistilise olulisuse üle otsustada ka usaldusvahemike alusel: need parameetrid, mille usaldusvahemik katab nullpunkti, on statistiliselt mitteolulised.

9.13. F-test ja t-testid

Ühe argumenttunnusega lineaarse regressioonmudeli korral on mudeli statistilise olulisuse kontrollimine F-testiga ekvivalentne lineaarliikme kordaja t-testiga: kui F-testi korral tuleb jääda nullhüpoteesi juurde, siis ka lineaarliikme kordaja a testimine t-testiga annab sama olulisuse nivoo korral tulemuseks nullhüpoteesi ning kui F-testi korral on nullhüpotees ümber lükatud, on ka t-testil nullhüpotees ümber lükatud.

Mitmese regressiooni korral sellist ekvivalentsust ei ole ning põhimõtteliselt võib esineda kuus erinevat situatsiooni, mis on esitatud tabelis 9.11.

Tabel 9.11. F-test ja t-testid, erinevad situatsioonid

	Mudel		Mudeli kordajad	
	F-testi tulemus	Järeldus	t-testide tulemused	Järeldused
1	H_1	On oluline	Kõigil H_1	Kõik on olulised
2	H_1	On oluline	Mõnel H_1, mõnel H_0	Mõni on oluline, mõni mitte
3	H_1	On oluline	Kõigil H_0	Ükski pole oluline
4	H_0	Ei ole oluline	Kõigil H_1	Kõik on olulised
5	H_0	Ei ole oluline	Mõnel H_1, mõnel H_0	Mõni on oluline, mõni mitte
6	H_0	Ei ole oluline	Kõigil H_0	Ükski pole oluline

Situatsioonide 1 ja 6 korral on mõlemat tüüpi testide tulemused täielikult kooskõlas. Situatsioonide 2–5 esinemine on seletatav sellega, et t-testide abil kontrollitakse parameetrite statistilist olulisust individuaalselt, F-test aga testib parameetreid korraga. Võib esineda olukordi, kus tunnuste grupp tervikuna on sõltuva tunnuse muutumisel määrav, kuid tunnused üksikuna ei ole. Samamoodi võib juhtuda, et mõni tunnus üksikult võetuna on oluline, kuid grupeerituna teiste tunnustega ei seleta mudel piisaval määral sõltuva tunnuse varieerumist ning F-testil tuleb vastu võtta nullhüpotees (situatsioon 5). Situatsioonid 2 ja 3 tekivad tüüpiliselt multikollineaarsuse esinemisel, kui seletavad tunnused on üksteisega tugevasti seotud (vt alapeatükk 9.15). Kõige ebaharilikum on situatsioon 4, kuid põhimõtteliselt võib ka see esineda.

9.14. Tunnuste valik

Regressioonmudeli koostamiseks tuleb püstitada hüpotees: millised tunnused funktsioontunnust mõjutavad, koguda nende kohta andmeid (teha valikvaatlus) ja seejärel hakata hindama erineva tunnuste komplektiga mudeleid. Alles hindamise käigus saame teada, millised

tunnused on statistiliselt olulised. Kui potentsiaalseid tunnuseid on palju, peab olema mingi süsteem, mille alusel tunnuseid hinnatavasse mudelisse valida.

Sammsammulise (*step by step*) ehk iteratiivse meetodi korral käib tunnuste lisamine või eemaldamine ühekaupa. Iga tunnuse lisamine (eemaldamine) võib mõjutada teiste tunnuste kordajaid, t-statistikuid ja järelikult ka olulisuse hüpoteesi kontrollimise tulemusi. Kasutusel on kaks alternatiivset lähenemisviisi:

- edaspidise valiku korral liigutakse väiksema tunnuste arvuga mudelist suurema poole, s.t tunnuseid lisatakse;
- tagurpidise valiku korral alustatakse suure arvu tunnustega ning ülearused eemaldatakse ükshaaval.

Edaspidine valik

Edaspidise valiku (*forward selection*) korral tuleb potentsiaalsed seletavad tunnused X_l reastada tähtsuse järjekorras. Selleks võib kasutada nende lineaarseid korrelatsioonikordajaid funktsioontunnusega r_{X_lY}.

1. Leitakse korrelatsioonimaatriks.
2. See X_l, mille korrelatsioonikordaja absoluutväärtus on kõige suurem, lisatakse mudelisse esimesena.
3. Kui see on statistiliselt oluline (t-test), jäetakse mudelisse.
4. Lisatakse tähtsuselt järgmine tunnus, mille korrelatsioonikordaja absoluutväärtus on suuruselt järgmine ja kontrollitakse selle statistilist olulisust.
5. Jätkatakse seni, kuni järgmine tunnus ei ole enam statistiliselt oluline või kui korrigeeritud determinatsioonikordaja väheneb.

Alati ei pruugi selline lähenemine niisama lihtsalt edeneda, sest mõni tunnus, mis algul oli statistiliselt oluline, võib järgmise tunnuse lisamisel muutuda mitteoluliseks. Siis tuleb lihtsalt proovida erinevaid tunnuste komplekte. Peab arvestama ka sellega, et potentsiaalsete seletavate tunnuste hulgas võib olla omavahel tugevalt seotud tunnuseid, mida ei saa korraga mudelisse panna (multikollineaarsus, vt alapeatükk 9.15).

N09Regressioon
N9.13,20

Näide 9.20. Käibe mudel ja edaspidine tunnuste valik

Näites 9.13 oli eesmärgiks leida ettevõtte käivet Y kirjeldav lineaarne mudel, kui potentsiaalseteks seletavateks tunnusteks olid toote hind P, turunduskulud M, majanduskasvu indeks E ja tooteühiku kulude indeks C. Tunnuste järjestamiseks tähtsuse alusel kasutame korrelatsioonimaatriksit.

	Y	P	M	E	C
Y	1				
P	0,137	1			
M	0,636	0,656	1		
E	0,846	−0,073	0,285	1	
C	0,237	0,855	0,738	0,0155	1

Vaatame korrelatsioonimaatriksi Y veergu, kus on seletavate tunnuste ja funktsioontunnuse Y vahelised korrelatsioonikordajad. Järjestame tunnused korrelatsioonikordajate absoluutväärtuste kasvamise järjekorras.

	Tunnus	Korrelatsioonikordaja
1	E	0,846
2	M	0,636
3	C	0,237
4	P	0,137

Esimesena paneme mudelisse majanduskasvu indeksi E ja leiame selle mudeli parameetrite hinnangud. See olgu mudel 1. Seejärel lisame turunduskulud M ja leiame mudeli 2. Niimoodi jätkates saame neli mudelit. Järgnevas tabelis on esitatud nende nelja mudeli kordajad, sulgudes on kordajate olulisuse tõenäosused. Viimasel kahel real on determinatsioonikordajad ja korrigeeritud determinatsioonikordajad.

	Mudel 1	Mudel 2	Mudel 3	Mudel 4
Vabaliige	−3989	−3527	−2642	−2639
E	50,13*** $(2{,}6 \cdot 10^{-6})$	42,88*** $(1{,}7 \cdot 10^{-7})$	40,10*** $(5{,}1 \cdot 10^{-7})$	39,46*** $(1{,}7 \cdot 10^{-06})$
M		332,9*** $(1{,}1 \cdot 10^{-4})$	469,7*** $(1{,}7 \cdot 10^{-4})$	478,5*** $(2{,}4 \cdot 10^{-4})$
C			−6,29* (0,083)	−4,35 (0,38)
P				68,31 (0,57)
R^2	0,716	0,885	0,906	0,908
R_a^2	0,700	0,872	0,888	0,883

* oluline nivool 0,1;
*** oluline nivool 0,01.

Kui me valime olulisuse nivooks 0,1, siis parimaks mudeliks on mudel 3: korrigeeritud determinatsioonikordaja R_a^2 on kõige suurem ning kõik tunnused on olulised valitud nivool 0,1. Käibe muutumist kirjeldab mudel

$$\hat{Y} = -2642 + 40{,}10E + 469{,}7M - 6{,}29C, \quad R^2 = 0{,}906,$$

kus E on majanduskasvu indeks, M turunduskulud ja C tooteühiku kulude indeks.

Tagurpidine valik

Tagurpidise valiku (*backward selection*) korral alustatakse mudelist, milles on kõik potentsiaalsed seletavad tunnused.

1. Kui on tunnuseid, mis ei ole statistiliselt olulised ettevõetud nivool, siis eemaldatakse tunnus, mille olulisuse tõenäosus on kõige suurem.
2. Eemaldatakse järgmine tunnus, mille olulisuse tõenäosus on järelejäänud tunnuste hulgas kõige suurem.
3. Jätkatakse, kuni kõik mudelisse jäänud tunnused on valitud nivool statistiliselt olulised.

Mitmetes statistikapakettides on selline protseduur automatiseeritud. Kasutaja määrab potentsiaalsed seletavad tunnused, mis kõik lisatakse esialgsesse mudelisse. Seejärel eemaldab tarkvara ükshaaval kõik mitteolulised tunnused ja jõuab mudelini, kus kõik tunnused on ette antud nivool statistiliselt olulised.

Näide 9.21. Linnaliinibusside kasutamine ja tagurpidine tunnuste valik

N09Regressioon
N9.21

Kasutame 40 USA linna andmeid aastast 1988 määramaks, millest sõltub reisijate arv linnaliinibussides (Stock ja Watson, 2003). Kasutatavad tunnused on järgmised:

REISIJAD — reisijate arv (tuhat reisijat tunnis);
HIND — bussipileti hind (dollarit);
BENSIIN — galloni bensiini hind (dollarit);
TULU — keskmine sissetulek elaniku kohta (dollarit);
RHV — linna rahvaarv (tuhat);
TIHEDUS — rahvastiku tihedus (tuhat elanikku ruutmiili kohta);
PINDALA — linna pindala (ruutmiili).

Sõltuvaks tunnuseks on REISIJAD ning esimesse mudelisse paneme kõik ülejäänud tunnused. Olulisuse nivooks võtame 0,05. Esimese mudeli hindamise aruanne Excelis (esitame vaid parameetrite tabeli ilma usalduspiirideta):

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	2745	2642	1,039	0,306
HIND	−239	452	−0,528	0,601
BENSIIN	522	2658	0,196	**0,845**
TULU	−0,195	0,0649	−3,001	0,0051
RHV	1,711	0,231	7,397	1,69E-08
TIHEDUS	0,116	0,0596	1,954	0,059
PINDALA	−1,155	1,80	−0,641	0,526

Kõige suurema olulisuse tõenäosusega on tunnus BENSIIN ($p = 0{,}845$). Viime läbi uue mudeli hindamise ilma selle tunnuseta.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	3216	1090	2,949	0,00573
HIND	−226	440	−0,512	**0,612**
TULU	−0,1957	0,0638	−3,069	0,00420
RHV	1,717	0,226	7,581	8,33E-09
TIHEDUS	0,1182	0,0580	2,037	0,049
PINDALA	−1,20	1,77	−0,677	0,503

Selles mudelis on kõige suurema olulisuse tõenäosusega tunnus HIND, mille eemaldame järgmisena ja viime läbi uue mudeli hindamise.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	3021	1011	2,99	0,00512
TULU	−0,1939	0,0630	−3,08	0,00404
RHV	1,731	0,222	7,79	3,81E-09
TIHEDUS	0,1159	0,0572	2,03	0,0505
PINDALA	−1,41	1,70	−0,83	**0,413**

Järgmisena eemaldame tunnuse PINDALA.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	2816	976	2,88	0,00659
TULU	−0,2013	0,0621	−3,24	0,00257
RHV	1,577	0,121	13,07	3,10E-15
TIHEDUS	0,1534	0,0349	4,40	9,34E-05

Viimases mudelis on kõik tunnused olulised nivool 0,05. Bussireisijate arvu mudel on

$$\widehat{\text{REISIJAD}} = 2816 - 0{,}2013\,\text{TULU} + 1{,}577\,\text{RHV} + \\ + 0{,}1534\,\text{TIHEDUS}, \quad R^2 = 0{,}919,$$

kus REISIJAD on bussireisijate arv (tuhat reisijat tunnis), TULU elanike keskmine sissetulek (dollarit), RHV linna rahvaarv (tuhat) ja TIHEDUS rahvastiku tihedus linnas (tuhat elanikku ruutmiili kohta). Linnades, mille rahvaarv on suurem, on ka bussireisijate arv suurem, mis on loogiline. Lisaks mõjutab ka rahvastiku tihedus bussireisijate arvu positiivselt, sest tiheda asustusega linnades on autokasutamine raskendatud. Linnades, kus elanike sissetulek on suurem, on bussireisijate arv väiksem tõenäoliselt sellepärast, et kasutatakse rohkem autosid.

Sellise mehaanilise sammsammulise meetodi puudus on, et nii kontrollitakse läbi vaid lineaarsed liikmed. Kui mudelis peaks olema näiteks mõne tunnuse ruutliige, siis seda ei avastata. Seepärast tuleks lisaks uurida ka jääkliikmete graafikuid, et avastada võimalikku mittelineaarsust.

9.15. Multikollineaarsus

Mudelis olevaid seletavaid tunnuseid nimetatakse tihti sõltumatuteks tunnusteks, mis tähendab, et need ei tohi omavahel olla seotud. Majandusnähtuste modelleerimisel on aga raske leida täiesti sõltumatuid tunnuseid. **Multikollineaarsus** (*multicollinearity*) on regressioonmudelisse lülitatavate sõltumatute tunnuste omavaheline tugev korrelatsioon. Sellisel juhul on raske eristada üksikute tunnuste mõju funktsioontunnusele.

Perfektne

Perfektne multikollineaarsus esineb siis, kui sõltumatud tunnused on omavahel lineaarselt seotud. Näiteks soovime hinnata mudelit

$$y = b + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + \varepsilon, \tag{9.73}$$

kuid tunnuste X_1 ja X_2 vahel on lineaarne seos

$$x_2 = cx_1 + d.$$

Lineaarne seos võib olla ka kolme või enama seletava tunnuse vahel, näiteks

$$x_3 = c_1x_1 + c_2x_2 + d.$$

Sellisel juhul on regressioonmudeli (9.73) parameetrite leidmine võimatu, sest vähimruutude meetodi kasutamisel tuleb võrrandisüsteemi determinant 0 ja tekib jagamine nulliga. Tarkvara võib sellisel juhul anda ka veateate.

Ligikaudne

Ligikaudne multikollineaarsus esineb siis, kui sõltumatute muutujate vahel on tugev korrelatsioon, kuid nad pole omavahel täpselt

lineaarselt seotud. Seda saab hinnata, kui vaatame seletavate tunnuste omavahelisi korrelatsioonikordajaid.

Näide 9.22. SKP elaniku kohta ja hõivatute osakaal erinevates majandussektorites


N09Regressioon
N9.22

Uurime, kas riigi elatustase sõltub sellest, kuidas töötajad on jagunenud kolme majandussektori vahel: tööstus, teenindus ja põllumajandus. Kasutame selleks 26 OECD riigi andmeid aastast 2012[a] ja hindame lineaarset mudelit

$$\mathrm{GDPPC} = a_1\,\mathrm{AGR} + a_2\,\mathrm{SER} + a_3\,\mathrm{IND} + b + \varepsilon, \qquad (9.74)$$

kus

GDPPC on SKP elaniku kohta (*GDP Per Capita*), tuhat dollarit;
AGR on põllumajanduses hõivatute osakaal;
SER on teeninduses hõivatute osakaal;
IND on tööstuses hõivatute osakaal.

Mudeli hindamine annab järgmise tulemuse:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,687
R Square	0,472
Adjusted R Square	0,400
Standard Error	12,065
Observations	26

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	2861,5	953,8	6,553	0,00247
Residual	22	3202,2	145,6		
Total	25	6063,7			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	213228,6	379626,2	0,5617	0,5800
AGR	−213315	379620,1	−0,5619	0,5799
SER	−213158	379626,3	−0,5615	0,5801
IND	−213256	379624,6	−0,5618	0,5800

Nagu ANOVA tabelist näeme, on mudel tervikuna statistiliselt oluline, F-testi olulisuse tõenäosus 0,00247 $<$ 0,05. Kuid üksiki tunnus pole statistiliselt oluline, kõik olulisuse tõenäosused (veerg *P-value*) on suuremad kui olulisuse nivoo 0,05. Lisaks

sellele on kordajate märgid ebaloogilised: kõik kordajad on negatiivsed, mis tähendab, et hõivatute osakaalu suurenemine igas sektoris vähendab SKP-d elaniku kohta.
Tegemist on perfektse multikollineaarsusega, sest hõivatute osakaal kokku on 100%:

$$\mathrm{AGR} + \mathrm{SER} + \mathrm{IND} = 100\%.$$

Uurime korrelatsioonimaatriksit.

	GDPPC	*AGR*	*SER*	*IND*
GDPPC	1			
AGR	−0,568	1		
SER	0,665	−0,693	1	
IND	−0,416	0,071	−0,768	1

Kõige tugevamini on sõltuva tunnusega seotud teeninduses hõivatute osakaal SER. Hindame mudelit

$$\mathrm{GDPPC} = a\,\mathrm{SER} + b + \varepsilon.$$

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,665
R Square	0,442
Adjusted R Square	0,419
Standard Error	11,872
Observations	26

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2681,0	2681,0	19,02	0,00021
Residual	24	3382,7	140,9		
Total	25	6063,7			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	−50,9	20,2	−2,52	0,019
SER	125,7	28,8	4,36	0,00021

Tunnus SER on statistiliselt oluline, sest olulisuse tõenäosus $p = 0{,}00021 < 0{,}05$.
Korrelatsioonimaatriksist on näha, et järgmine tunnus, mida võiks proovida mudelisse lisada, on põllumajanduses hõivatute osakaal AGR. Hindame mudelit

$$\mathrm{GDPPC} = a_1\,\mathrm{SER} + a_2\,\mathrm{AGR} + b + \varepsilon.$$

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,681
R Square	0,464
Adjusted R Square	0,418
Standard Error	11,884
Observations	26

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	2815,6	1407,79	9,97	0,00076
Residual	23	3248,1	141,22		
Total	25	6063,7			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	−28,51	30,5	−0,93338	0,36
AGR	−60,89	62,4	−0,97629	0,34
SER	98,63	40,0	2,46445	0,022

Tunnus AGR ei ole statistiliselt oluline, sest selle olulisuse tõenäosus $p = 0{,}34 > 0{,}05$. Ka korrigeeritud determinatsioonikordaja vähenes: eelmise mudeli korral oli see 0,419, nüüd 0,418. Järelikult, mudel tunnustega SER ja AGR ei sobi ning parimaks mudeliks on see, kus on vaid teeninduses hõivatute osakaal:

$$\widehat{\mathrm{GDPPC}} = 125{,}7\,\mathrm{SER} - 50{,}9, \quad R^2 = 0{,}442.$$

Nendes riikides, kus teeninduses hõivatute osakaal on 1% võrra suurem, on SKP elaniku kohta suurem 125,7 dollari võrra.

[a] Allikas: `http://www.oecd-ilibrary.org`

Näide 9.23. Sigaretid ja multikollineaarsus

N09Regressioon
N9.18,23

Näites 9.18 modelleeriti sigaretisuitsus sisalduva süsinikoksiidi koguse y sõltuvust sigaretis sisalduva tõrva ja nikotiini hulgast. Esialgses mudelis oli nikotiinis sisaldus xN statistiliselt mitteoluline, samuti oli ebaloogiline selle kordaja märk.
Uurime võimalikku multikollineaarsust, milleks leiame korrelatsioonimaatriksi:

	xT	xN	y
xT	1		
xN	**0,977**	1	
y	0,957	0,926	1

Sõltumatute muutujate tõrvakogus xT ja nikotiinikogus xN vaheline korrelatsioonikordaja on suurem kui nende ja süsinikoksiidi koguse y vaheline korrelatsioonikordaja — esineb multikollineaarsus. See ongi põhjus, miks nikotiini kordaja märk oli ebaloogiline ja kordaja ise statistiliselt mitteoluline. Selliseid tunnuseid, mille vahel on tugev seos, ei tohi korraga seletavateks tunnusteks võtta.

Multikollineaarsuse tõttu võib sageli tekkida olukord, kus mudel tervikuna on statistiliselt oluline, kuid tõlgendamise võimalused pole kooskõlas teoreetiliste seisukohtadega, mudel on ebaloogiline.

Multikollineaarsuse tunnused

Mõned **multikollineaarsuse tunnused**:

- mõne sõltumatute tunnuste paari omavaheline korrelatsioon on tugevam kui korrelatsioon sõltuva muutujaga;
- mudeli parameetritel on väga suured standardvead ja väga laiad usaldusvahemikud;
- mudeli ühe või mitme parameetri märk on ebaloogiline.

Multikollineaarsuse avastamiseks viiakse lisaks regressioonanalüüsile läbi ka korrelatsioonanalüüs, s.t leitakse paarikaupa kõikide tunnuste korrelatsioonikordajad ning võrreldakse sõltuvate tunnuste omavahelist korrelatsiooni ja korrelatsiooni sõltuva tunnusega.

Mida teha, kui multikollineaarsus esineb?

1. Jätta üks multikollineaarne tunnus mudelist välja.
2. Teisendada andmeid. Näiteks kahe kollineaarse tunnuse asemel kasutada nende suhet.
3. Suurendada valimi mahtu, mis võib vähendada multikollineaarsust.

9.16. Lineaarse mudeli vabaliige ja nullpunkti läbiv regressioonsirge

Alapeatükis 9.12 mainisime, et vabaliikme b statistilist olulisust mudelis

$$y = b + a_1x_1 + a_2x_2 + \ldots + a_kx_k + \varepsilon$$

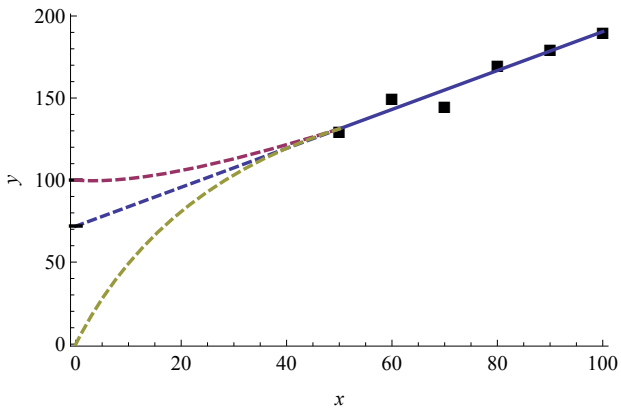
tavaliselt ei kontrollita. Seda sellepärast, et vabaliikmeta lineaarset mudelit enamasti ei kasutata. Ka siis, kui vabaliige on statistiliselt mitteoluline (t-testi tulemuseks on H_0), jäetakse see mudelisse.

Vabaliikme olemasolu on vajalik vähimruutude meetodi kasutamise seisukohalt. Vabaliige garanteerib, et regressioonijääkide summa $\sum u_i = 0$. Seda omadust kasutatakse mitmete regressioonanalüüsi käigus leitavate suuruste valemite tuletamisel. Sellisel juhul on regressioonmudeli diagnostikaks kasutatavad suurused, nagu determinatsioonikordaja R^2 ja F-statistik, usaldusväärsed. Kui me paneme aga peale kitsenduse, et vabaliige $b = 0$, siis $\sum u_i \neq 0$ ja me ei saa enam korrektselt läbi viia regressiooni dispersioonanalüüsi ega kasutada mudeli iseloomustamiseks determinatsioonikordajat.

Kas vabaliiget tuleks tõlgendada või mitte?

Vabaliikme tõlgendamisest

1. Matemaatiliselt näitab mitmese lineaarse mudeli vabaliige tunnuse Y väärtust juhul, kui kõik argumenttunnused on korraga nullid. Reaalsuses enamasti sellist situatsiooni ei esine. Küll on aga ühe argumenttunnusega mudeli $y = ax + b$ korral seoseid, kus vabaliikmel on kindel sisu. Mõned näited:
   - kulufunktsioonis on vabaliige püsikulud;
   - nõudlusfunktsioonis vastab nõutava koguse puudumisele piirhind;
   - tarbimismudelites näitab vabaliige tuludest sõltumatut autonoomset tarbimist.
2. Kui olemasolevad andmed on nullpunktist kaugel, pole meil tõestust selle kohta, et lineaarsus säilib ka nullpunkti lähedal. Seepärast peab vabaliikme väärtuse tõlgendamisega olema ettevaatlik (vt joonis 9.31).



Joonis 9.31. Piirkonnas (0, 50) andmed puuduvad ja me ei tea, kas mudel on seal ka lineaarne ning sirget võib pikendada punktini $\hat{y}(0) = 70$ või esineb mittelineaarsus ja $\hat{y}(0) = 100$ või hoopis $\hat{y}(0) = 0$

Näide 9.24. Kulutused transpordile: vabaliikme tõlgendus

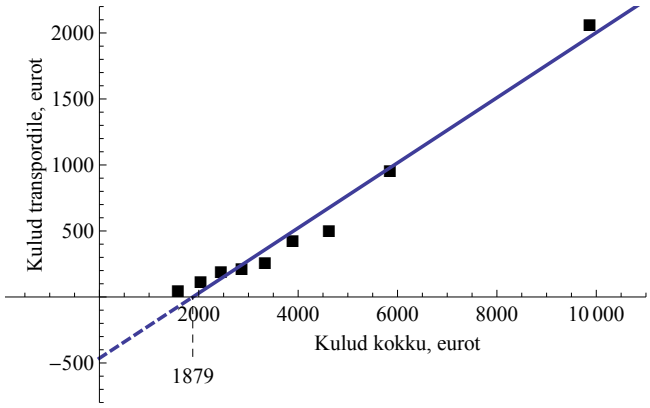
Näites 9.3 lk 431 analüüsisime, kuidas pereliikme kulutused erinevatele kaupadele ja teenustele sõltuvad kogukuludest. Transpordikulude jaoks saime

$$\hat{y}_{TR} = -464{,}1 + 0{,}2467x, \tag{9.75}$$

kus y_{TR} on kulud transpordile pereliikme kohta aastas (eurot) ja x kulutused kokku pereliikme kohta aastas (eurot). Kuidas tõlgendada negatiivset vabaliiget -464? Matemaatikast tuntud vabaliikme tõlgenduse järgi saame mudelist (9.75), et kui kogukulud $x = 0$, siis kulud transpordile on $\hat{y}_{TR}(0) = -464{,}1$ eurot. See pole võimalik, sest kulud ei saa negatiivsed olla.
Me saame aga leida, millal tekivad kulud transpordile. Selleks leiame, millise kogukulude väärtuse korral kulud transpordile võrduvad nulliga:

$$\begin{aligned} \hat{y}_{TR} &= 0 \\ -464{,}1 + 0{,}2467x &= 0 \\ x &= \frac{464{,}1}{0{,}2467} \\ x &\approx 1881. \end{aligned}$$

Järelikult kulutused transpordile tekivad siis, kui kogukulud pereliikme kohta on ligikaudu 1881 eurot aastas (*ca* 157 eurot kuus). See on seletatav sellega, et suur osa vaesemast elanikkonnast (pensionärid, paljulapselised pered) transpordile kulutama ei pea. Keskmine vanaduspension oli aastal 2012 näiteks 312 eurot kuus.



Mõnikord tuleb siiski hinnata lineaarset mudelit, kus teatud kaalutlustest lähtudes peab vabaliige puuduma. Seda nimetatakse **regressiooniks läbi nullpunkti** (*Regression through the Origin, RTO*) ja sellise mudeli üldkuju on

Regressioon läbi nullpunkti

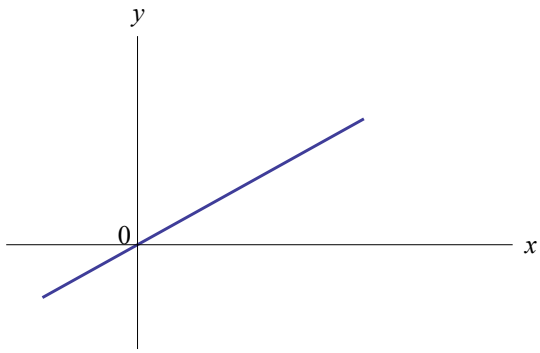
$$y = a_1x_1 + a_2x_2 + \ldots + a_kx_k + \varepsilon. \tag{9.76}$$

Ühe argumenttunnuse korral

$$y = ax + \varepsilon. \tag{9.77}$$

Vähimruutude meetodi kasutamisel saame mudeli (9.77) parameetri hinnanguks

$$a = \frac{\sum y_i x_i}{\sum x_i^2}. \tag{9.78}$$



Joonis 9.32. Võrdelise seose $y = ax$ graafik, sirge läbib nullpunkti

Kui kirjutame mudeli (9.77) välja sirgel asuvate mudelväärtuste $\hat{y}$ jaoks, on tegemist **võrdelise seosega**

Võrdeline seos

$$\hat{y} = ax, \tag{9.79}$$

kus a on võrdetegur (joonis 9.32).

Näide 9.25. Puukoore tootmine

Tarbepuidu tootmisel ülejäävat puukoort kasutatakse aianduses multšina ja hakkepuidu kõrval ka küttematerjalina. Seoses energiahindade tõusuga on seda hakatud kasutama küttematerjalina ka elektri tootmisel. Sellega seoses on nõudlus puukoore järele tõusnud. Ajakirjas Journal of Environmental Horticulture 2006. aastal ilmunud artiklis (Lu jt, 2006) analüüsiti, kuidas puukoore tootmine sõltub metsaraie mahust. Kasutati andmeid USA erinevatest piirkondadest aastatel 1986–2001.

On selge, et kui metsa ei raiuta, siis ka puukoort turule ei tule. Seepärast kasutati ilma vabaliikmeta lineaarset mudelit

$$y = ax + \varepsilon,$$

kus y on kuiva puukoore kogus (tuhat tonni) ja x raiemaht antud piirkonnas (miljonit kuupjalga). Regressioonanalüüs andis erinevate piirkondade jaoks erinevad võrdetegurid. Näiteks:

USA kirdeosa	lehtpuu	$\hat{y} = 0{,}89x$;
	okaspuu	$\hat{y} = 1{,}01x$;
USA kaguosa	lehtpuu	$\hat{y} = 1{,}38x$;
	okaspuu	$\hat{y} = 1{,}44x$.

Näeme, et USA kirdeosas annab üks miljon kuupjalga lehtpuu raiet keskmiselt 0,89 tuhat tonni puukoort, okaspuu korral aga 1,01 tuhat tonni. USA kaguosas saadakse puukoort rohkem: lehtpuu korral keskmiselt 1,38 tuhat tonni 1 miljoni kuupjala metsamaterjali kohta ja okaspuu korral 1,44 tuhat tonni.
Kasutades metsaraie prognoosi aastateks 2010 kuni 2050, leiti nende mudelite abil puukoore toodangu prognoos samaks perioodiks.

Näitena veel mõned majanduses tuntud võrdelised seosed, kus tuleb kasutada mudelit ilma vabaliikmeta:

- ettevõtte muutuvkulu on võrdeline tootmismahuga;
- Milton Friedmani püsiva sissetuleku hüpotees, mille järgi tarbija püsiv tarbimine on võrdeline püsiva sissetulekuga;
- monetarismi majandusteooria järgi on hindade muutumise määr (inflatsioon) võrdeline raha pakkumise muutusega;
- finantsvarade hindamise CAPM (*Capital Asset Pricing Model*) mudeli järgi on üksikult väärtpaberilt saadav riskipreemia (täiendav tulu võrreldes riskivaba tuluga) võrdeline võrdlusindeksi riskipreemiaga, kus võrdeteguriks on selle väärtpaberi beetakordaja.

Kõigi nende seoste korral tuleb võrdeteguri määramiseks hinnata regressioonmudelit kujul (9.79). Nullpunkti läbivat regressiooni tuleb kasutada ka siis, kui andmed on teisendatud standardiseeritud skaalasse (vt alapeatükk 9.20).

Tarkvarapakettides tuleb nullpunkti läbiva regressioonjoone saamiseks seda eraldi märkida. Näiteks programmis Excel tuleb vahendi *Regression* aknas teha valik *Constant is Zero*. Ka funktsiooni LINEST kasutamisel on selline valik olemas.

Näide 9.26. Koopiamasinate teenindamiseks kulunud aeg


N09Regressioon
N9.26

Koopiamasinaid hooldav ettevõte on huvitatud regressioonmudelist, mis aitaks planeerida tööjõu vajadust. Tööjõu vajaduse planeerimiseks on vaja teada, kui palju kulutatakse töötunde iga kliendi juures. Kulutatud töötundide arv sõltub loomulikult sellest, kui palju on kliendil hooldamist vajavaid koopiamasinaid. Mudeli koostamiseks koguti vastavad andmed ja regressioonanalüüsi läbiviimisel saadi järgmine tulemus (ruumi kokkuhoiu mõttes on ära jäetud ANOVA tabel):

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,935
R Square	0,874
Adjusted R Square	0,858
Standard Error	2,056
Observations	10

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0,101	1,267	0,079	0,939	−2,822	3,023
ARV	1,419	0,191	7,44	0,000074	0,979	1,859

Selle põhjal võib kirja panna mudeli:

$$\widehat{\text{AEG}} = 0{,}101 + 1{,}419 \cdot \text{ARV},$$

kus AEG on kliendi juures kulutatud tööaeg tundides ja ARV hooldatud koopiamasinate arv. Kuid loogiline on, et kui koopiamasinate arv on null, siis ei tohiks ka aega kuluda. Järelikult ei ole vabaliikme esinemine selles mudelis põhjendatud. Vabaliikme statistilise olulisuse testimisel tuleb vastu võtta nullhüpotees ($p = 0{,}939 > 0{,}05$), mis kinnitab, et vabaliige võib olla null. Seepärast viime läbi uue mudeli hindamise, kus regressioonjoon läbib nullpunkti (konstant on null).

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,982
R Square	0,964
Adjusted R Square	0,853
Standard Error	1,939
Observations	10

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A
ARV	1,432	0,0924	15,508	8,4E-08	1,223	1,641

Uueks mudeliks on

$$\widehat{\text{AEG}} = 1{,}432 \cdot \text{ARV}.$$

Selle mudeli järgi on keskmine koopiamasinale kulutatud aeg ligikaudu 1,4 tundi. Usalduspiirid 95%-lise usaldatavusega on 1,2–1,6 tundi.

Kui regressioonjoon ei läbi nullpunkti, siis jääkliikmete summa ei pruugi olla enam null, mis muidu on vähimruutude meetodi korral garanteeritud. Sellest tulenevalt ei pruugi kehtida ka tingimus $SST = SSR + SSE$ ning tekivad probleemid determinatsioonikordaja leidmisel. Determinatsioonikordajat on võimalik defineerida erineval moel (Greene, 2002, lk 36) ning erinevad tarkvarapaketid arvutavad seda erinevalt (Eisenhauer, 2003). Ühel meetodil arvutatud determinatsioonikordaja võib tulla mõnikord negatiivne, teisel meetodil arvutatult võib see tulla suurem kui 1. Interpretatsioon „kui suur osa koguhajuvusest on mudeli abil ära seletatud“ ei kehti. Seepärast sellisel juhul ei ole soovitatav determinatsioonikordajat mudeli juures esitada, kui just ei teata, millisel meetodil determinatsioonikordaja arvutatakse. Samuti ei saa omavahel võrrelda vabaliikmeta ja vabaliikmega mudeli determinatsioonikordajaid.

Programmis Excel kasutatakse nullpunkti läbiva regressioonjoone korral koguhajuvuse SST ja regressioonhajuvuse SSR arvutamisel valemite (9.29) ja (9.31) asemel järgmisi valemeid:

$$SST = \sum y_i^2, \quad SSR = \sum \hat{y}_i^2 \tag{9.80}$$

ning determinatsioonikordaja leitakse valemist

$$R^2 = \frac{SSR}{SST} = \frac{\sum \hat{y}_i^2}{\sum y_i^2}. \tag{9.81}$$

Mida teha siis, kui loogilistest kaalutlustest lähtudes peaks regressioonjoon läbima nullpunkti, kuid vabaliikme statistilise olulisuse testimisel on nullhüpotees ümber lükatud? Siis tuleb arvestada järgmiste võimalustega.

1. Mudelist on võib olla välja jäänud mõni oluline tunnus. Selle tunnuse lisamisel mudelisse võib regressioonsirge oma asendit muuta nii, et läheb nullpunkti lähedalt ja vabaliikme jaoks kehtib H_0.

2. Võib-olla on lihtsalt „halb“ valim. See tähendab, et vabaliikme jaoks tegelikult nullhüpotees kehtib, kuid arvutused valimi põhjal annavad, et ei kehti (II liiki viga).
3. Nullpunkti lähedal võib esineda mittelineaarsus, kuid selle avastamiseks andmed seal piirkonnas puuduvad (vt joonis 9.31).

Kokku võttes rõhutame veel kord, et regressiooni läbi nullpunkti võib kasutada ainult põhjendatud kaalutlustel, kui seose loogika seda nõuab.

9.17. Lineariseerimine

Kui argumenttunnuste X_l valik on õnnestunud, võib siiski juhtuda, et saadud mudel ei rahulda meid. Mõnikord on see tingitud sellest, et seosed tunnuste Y ja X_l vahel on mittelineaarsed, mudel on meil valitud aga lineaarne. Järelikult tuleks kasutada mittelineaarset regressioonmudelit.

Üks võimalus on leida mittelineaarse regressioonmudeli parameetrid otse, kasutades vastava statistikapaketi võimalusi. Seda saab teha juhul, kui vastava kujuga mudeli hindamise võimalus on statistikapaketis olemas. Sellist võimalust vaatasime alapeatükis 9.7. Kuid tabelarvutuses ei ole sellisel juhul võimalik hinnata tunnuste standardvigu ega statistilist olulisust. Seepärast on eelistatud lineaarse mudeli kasutamine.

Päris paljude mittelineaarsete regressioonmudelite leidmist saab lihtsustada funktsiooni lineariseerimisega. Näiteks ruutfunktsiooni

$$y = ax^2 + bx + c \tag{9.82}$$

saab lineariseerida, kui võtame kasutusele uue tunnuse

$$z = x^2.$$

Siis võime mudeli (9.82) asemel hinnata kahe argumenttunnusega lineaarset mudelit

$$y = az + bx + c.$$

Näide 9.27. Kulud toidule ja mudeli lineariseerimine

N09Regressioon
N9.27

Näites 9.6 nägime, et toidule tehtavate kulutuste sõltuvust kogukuludest kirjeldab hästi ruutfunktsioon

$$\hat{y}_T = 111{,}2 + 0{,}3x - 2 \cdot 10^{-5}x^2, \tag{9.83}$$

kus x on kulud kokku. Hindame seda mudelit uuesti, kasutades ruutfunktsiooni lineariseerimist. Selleks arvutame iga objek-

ti jaoks x^2. Näitena on tabelis toodud arvutused esimese kahe vaatluse jaoks.

y_T	x	x^2
342,6	992,2	984460,84
554,1	1585,7	2514444,49
...	...	...

Nüüd kasutame programmi Excel analüüsivahendit *Regression* ja hindame mudelit

$$y_T = b + a_1 x + a_2 x^2 + \varepsilon.$$

Argumenttunnusteks võtame kaks tunnust: x ja x^2. Tulemus on järgmine:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,993
R Square	0,987
Adjusted R Square	0,983
Standard Error	46,42
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	1138943	569471,5	264,27	2,55E-07
Residual	7	15084,27	2154,895		
Total	9	1154027			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	111,2	50,6	2,20	0,06404	−8,49	230,9
x	0,301	0,0235	12,80	4,13E-06	0,245	0,356
x^2	−1,58E-05	2,11E-06	−7,52	0,00013	−2,08E-05	−1,1E-05

Näeme, et mudel on statistiliselt oluline, F-testi olulisuse tõenäosus $2{,}55 \cdot 10^{-7} < 0{,}05$. Nii kogukulud x kui ka nende ruut on statistiliselt olulised. Mudeli parameetrite hinnangud saame täpsemad, kui graafikule lisatud trendijoone korral. Mudel on

$$\hat{y}_T = 111{,}2 + 0{,}301x - 1{,}58 \cdot 10^{-5} x^2$$

ja langeb kokku näites 9.6 saadud mudeliga (9.83). Lisaks on nüüd leitud ka parameetrite usalduspiirid.

Tabelis (9.12) on esitatud näited erinevate mittelineaarsete funktsioonide lineariseerimisest. Vastava lineaarse mudeli parameetrite leidmiseks tuleb eelnevalt arvutada uued tunnused vastavalt veerule „Teisendused“. Eksponentsiaalse mudeli lineariseerimiseks võetakse mõlemast poolest naturaallogaritm:

$$y = ae^{bx}$$
$$\ln y = \ln\left(ae^{bx}\right) = \ln a + \ln e^{bx} = \ln a + bx.$$

Teisendamisel on siin kasutatud korrutise logaritmi omadust.

Tabel 9.12. Mittelineaarsete funktsioonide lineariseerimine

Mittelineaarne funktsioon	Teisendused	Lineaarne funktsioon
$y = ax^2 + bx + c$	$z = x^2$	$y = az + bx + c$
$y = ax^3 + bx^2 + cx + d$	$z = x^3$, $v = x^2$	$y = az + bv + cx + d$
$y = a\sqrt{x} + b$	$z = \sqrt{x}$	$y = az + b$
$y = \dfrac{a}{x} + b$	$z = \dfrac{1}{x}$	$y = az + b$
$y = a \ln x + b$	$z = \ln x$	$y = az + b$
$y = ae^{bx}$	$z = \ln y$	$z = \ln a + bx$

Üht näidet mittelineaarse funktsiooni lineariseerimise kohta vaatame alapeatükis 9.22.

9.18. Kvalitatiivsed seletavad tunnused

Seni vaadeldud regressioonmudelites olid kõik seletavad tunnused intervallskaalas. On aga palju suurusi, mida ei saa mõõta intervallskaalas, kuid nende mõju erinevatele funktsioontunnustele on olemas. Näiteks isikuküsitluste analüüsimisel võivad uuritavat suurust mõjutada inimese sugu (nimiskaalas, kaheväärtuseline), ametikoht (nimiskaalas), haridustase (järjestusskaalas), elukoht (linnas või maal, kaheväärtuseline). Ettevõtete analüüsimise korral on tihti üheks seletavaks tunnuseks tegevusala, võib-olla ka omandivorm, mis on mõlemad nimiskaalas. Nimiskaalas ja järjestusskaalas mõõdetud tunnuseid nimetame kvalitatiivseteks tunnusteks.

Oletame, et ehitusettevõttes töötab kolm müügijuhti: Kask, Lepp ja Tamm. Müügijuht on isik, kes koostab hinnapakkumised ning sõlmib lepinguid. Ettevõte soovib konstrueerida mudelit, mille järgi prognoosida ehitusobjekti kasumimarginaali (puhaskasumi ja käibe suhe) selle järgi, kes kolmest müügijuhist pidas kliendiga läbirääkimisi ning sõlmis ehitustöödeks lepingu. Mudel peaks siis olema järgmine:

$$\hat{y} = b + \gamma_1 D_1 + \gamma_2 D_2, \tag{9.84}$$

kus y on kasumimarginaal, b, γ_1 ja γ_2 mudeli parameetrid ning

$$D_1 = \begin{cases} 1, & \text{kui lepingu sõlmis Lepp;} \\ 0, & \text{kui lepingut ei sõlminud Lepp;} \end{cases}$$

$$D_2 = \begin{cases} 1, & \text{kui lepingu sõlmis Tamm;} \\ 0, & \text{kui lepingut ei sõlminud Tamm.} \end{cases}$$

Tunnuste D_1 ja D_2 väärtused kõigi kolme müügijuhi jaoks on toodud tabelis 9.13. Tabeli viimases veerus on leitud kasumimarginaali mudelväärtus. Kirjutame tulemused eraldi välja:

$$\text{Kask} \quad \hat{y}_K = b, \tag{9.85}$$

$$\text{Lepp} \quad \hat{y}_L = b + \gamma_1, \tag{9.86}$$

$$\text{Tamm} \quad \hat{y}_T = b + \gamma_2. \tag{9.87}$$

Tabel 9.13. Valemist (9.84) leitud kolme müügijuhi kasumimarginaal

Müügijuht	D_1	D_2	Kasumimarginaal
Kask	0	0	$\hat{y}_K = b + \gamma_1 D_1 + \gamma_2 D_2 = b + \gamma_1 \cdot 0 + \gamma_2 \cdot 0 = b$
Lepp	1	0	$\hat{y}_L = b + \gamma_1 D_1 + \gamma_2 D_2 = b + \gamma_1 \cdot 1 + \gamma_2 \cdot 0 = b + \gamma_1$
Tamm	0	1	$\hat{y}_T = b + \gamma_1 D_1 + \gamma_2 D_2 = b + \gamma_1 \cdot 0 + \gamma_2 \cdot 1 = b + \gamma_2$

Analüüsime, mida näitavad fiktiivsete tunnuste D_1 ja D_2 kordajad γ_1 ja γ_2. Kasumimarginaali avaldistest (9.86) ja (9.85) avaldame γ_1:

$$\gamma_1 = \hat{y}_L - b = \hat{y}_L - \hat{y}_K.$$

Näeme, et fiktiivse tunnuse D_1 kordaja γ_1 mudelis (9.84) näitab, kui palju erineb kasumimarginaal Lepa ja Kase sõlmitud lepingute korral.

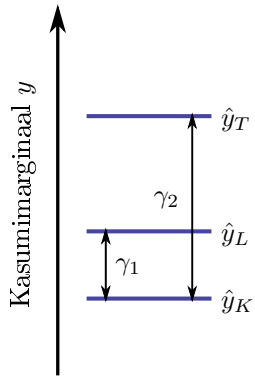
Müügijuhi Tamme kasumimarginaali valemist (9.87) avaldame kordaja γ_2:

$$\gamma_2 = \hat{y}_T - b = \hat{y}_T - \hat{y}_K.$$

Kordaja γ_2 näitab, kui palju erineb kasumimarginaal Tamme ja Kase sõlmitud lepingute korral.

Kasumimarginaali erinevad väärtused $\hat{y}_K$, $\hat{y}_L$ ja $\hat{y}_T$ on kujutatud joonisel 9.33. Nende erinevused on määratud fiktiivsete tunnuste kordajatega γ_1 ja γ_2. Müügijuht Kask, kelle korral mõlemad fiktiivsed tunnused on nullid, on **baasväärtus** ning kasumimarginaal ülejäänud kahe müügijuhi korral leitakse selle suhtes. Kokkuvõtteks võime öelda, et fiktiivse tunnuse kordaja näitab kvalitatiivse tunnuse väärtusele vastava funktsioontunnuse väärtuse erinevust baasväärtusega määratud väärtusest.

Kasumimarginaal sõltub ühest kvalitatiivsest tunnusest „müügijuht“, mis on mõõdetud nimiskaalas ja võib omada kolme erinevat



Joonis 9.33. Kolme erineva müügijuhi kasumimarginaal

taset. Selle tunnuse mõju modelleerimiseks on vaja kaht kaheväärtuselist tunnust D_1 ja D_2, mida nimetatakse fiktiivseteks tunnusteks (*dummy variable*) ehk indikaatortunnusteks (*indicator variable*). Kahe kaheväärtuselise tunnusega saab täielikult ära määrata kolm kvalitatiivse tunnuse taset (vt tabel 9.13).

Fiktiivne tunnus

Fiktiivne tunnus on kaheväärtuseline tunnus, millel võib olla väärtus 0 või 1 ning mis vastab kvalitatiivse tunnuse kindlale tasemele.
Fiktiivse tunnuse kordaja näitab kvalitatiivse tunnuse väärtusele vastava funktsioontunnuse väärtuse erinevust baasväärtusega määratud väärtusest.

Kui kvalitatiivsel tunnusel on neli taset, siis on vaja kolme fiktiivset tunnust (tabel 9.14), viie taseme puhul nelja tunnust jne. Fiktiivseid tunnuseid nimetatakse seepärast fiktiivseteks, et need ei vasta mitte üksikutele tunnustele, vaid ühe kvalitatiivse tunnuse erinevatele tasemetele.

Tabel 9.14. Nelja tasemega kvalitatiivne tunnus vajab kolme fiktiivset tunnust: D_1, D_2 ja D_3

Kvalitatiivse tunnuse tasemed	D_1	D_2	D_3
A	0	0	0
B	1	0	0
C	0	1	0
D	0	0	1

Fiktiivsete tunnuste arv

Fiktiivsete tunnuste **arv** on ühe võrra väiksem kvalitatiivse tunnuse tasemete arvust. Väärtus, mille fiktiivset tunnust mudelis pole, on **baasväärtus**.

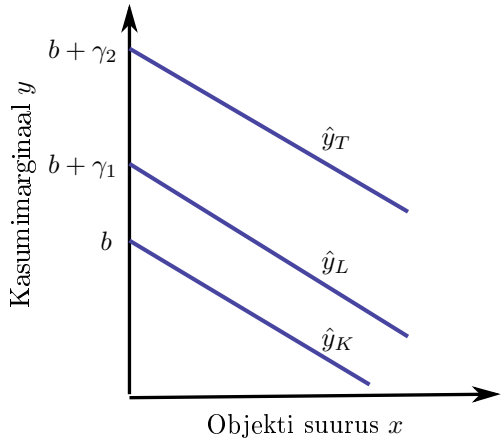
Sõltugu nüüd ehitusobjekti kasumimarginaal lisaks lepingu sõlminud müügijuhi isikust ka ehitusobjekti maksumusest x:

$$\hat{y} = ax + b + \gamma_1 D_1 + \gamma_2 D_2. \tag{9.88}$$

Mudelist (9.88) kasumimarginaali mudelväärtus sõltuvalt müügijuhist:

$$\begin{aligned}
\text{Kask} \quad & \hat{y}_K = ax + b + \gamma_1 \cdot 0 + \gamma_2 \cdot 0 = ax + b, \\
\text{Lepp} \quad & \hat{y}_L = ax + b + \gamma_1 \cdot 1 + \gamma_2 \cdot 0 = ax + b + \gamma_1, \\
\text{Tamm} \quad & \hat{y}_T = ax + b + \gamma_1 \cdot 0 + \gamma_2 \cdot 1 = ax + b + \gamma_2.
\end{aligned}$$

Näeme, et fiktiivsete tunnuste kordajad liituvad vabaliikmele b, s.t muudavad vabaliiget. Need kolm mudelit vastavad kolmele sirgele, millel on ühesugune tõus a ja erinev vabaliige. Suuremate objektide korral on konkurents suurem ning võimalik kasumimarginaal seetõttu väiksem. Seepärast on loogiline oletada, et kordaja a on negatiivne ja joonisel 9.34 on kujutatud langevad sirged.



Joonis 9.34. Kasumimarginaali sõltuvus objekti suurusest ja müügijuhist

Näide 9.28. iPodide hinnad eBay oksjonitel

N09regressioon
N9.28

Ajakirjas Journal of Applied Econometrics ilmunud artiklis analüüsiti, millest sõltub eBay[a] oksjoni lõpphind. Kasutati Apple'i iPod mini oksjonite 27. juuni kuni 18. juuli 2008 andmeid, kokku

1225 oksjonit. Oksjonite kohta olid kogutud järgmised andmed (Rezende, 2008):

HIND — lõpphind dollarites;
ALGH — algind dollarites;
ARV — pakkumiste arv;
SKORD — müügil oleva iPodi seisukord:
−1 — halb (mõrane klaas, katkine kõrvaklappide ühendus, aku mahutavus väike);
0 — keskmine (kriimustused);
1 — hea.

IPodi seisukord on kvalitatiivne tunnus ja selle mudelisse panekuks tuleb luua vastavad fiktiivsed tunnused. Kuna seisukorral on kolm taset, läheb vaja kaht fiktiivset tunnust. Baasväärtuseks võtame seisukorra „hea“ (1) ning fiktiivsed tunnused loome seisukorra „keskmine“ (0) ja „halb“ (−1) jaoks:

$$D_1 = \begin{cases} 1, & \text{kui seisukord on keskmine (0);} \\ 0, & \text{kui seisukord on muu;} \end{cases}$$
$$D_2 = \begin{cases} 1, & \text{kui seisukord on halb (−1);} \\ 0, & \text{kui seisukord on muu.} \end{cases}$$

Andmed mõnede erinevas seisukorras olevate iPodide kohta pärast fiktiivsete tunnuste loomist:

HIND	ALGH	ARV	SKORD	D1	D2
109,5	0,01	9	1	0	0
127,5	25	7	1	0	0
125	125	1	0	1	0
103,5	1	9	0	1	0
56	25	2	−1	0	1
31,01	26	2	−1	0	1
...	...	...	...	...	...

Regressioonmudel, mille parameetreid tuleb hinnata:

$$\text{HIND} = b + a_1\text{ALGH} + a_2\text{ARV} + \gamma_1 D_1 + \gamma_2 D_2 + \varepsilon.$$

Excelis saame järgmise regressioonanalüüsi aruande:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,684
R Square	0,468
Adjusted R Square	0,467
Standard Error	31,23
Observations	1225

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	1048770,045	262192,5	268,8	1,13E-165
Residual	1220	1190166,318	975,5462		
Total	1224	2238936,363			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	59,76	3,35	17,8	1,69E-63	53,19	66,33
ALGH	0,554	0,0250	22,2	5,87E-92	0,51	0,60
ARV	6,569	0,341	19,2	2,59E-72	5,90	7,24
D1	−16,11	2,07	−7,8	1,44E-14	−20,17	−12,05
D2	−40,47	3,24	−12,5	8,67E-34	−46,837	−34,12

Mudel on statistiliselt oluline: F-testi olulisuse tõenäosus $p = 1{,}13 \cdot 10^{-165} < 0{,}05$. Kõik parameetrid on statistiliselt olulised. Mudel, mis kirjeldab iPodi lõpphinda, on järgmine:

$$\widehat{\text{HIND}} = 59{,}76 + 0{,}554\,\text{ALGH} + 6{,}569\,\text{ARV} - 16{,}11D_1 - 40{,}47D_2,$$
$$R^2 = 0{,}468.$$

Tõlgendame mudeli parameetreid:

- kui alghind on 1 dollari võrra suurem, siis lõpphind on 0,554 dollarit suurem;
- pakkumiste arvu suurenemine ühe võrra suurendab lõpphinda 6,569 dollari võrra;
- keskmises seisukorras oleva iPodi hind on 16,11 dollari võrra väiksem heas seisukorras oleva iPodi hinnast;
- halvas seisukorras oleva iPodi hind on 40,47 dollari võrra väiksem heas seisukorras oleva iPodi hinnast.

Mudel kirjeldab ära 46,8% iPodide hinna varieerumisest eBay oksjonitel.

Kui soovime teada, kui palju erineb halvas seisukorras oleva iPodi hind keskmises seisukorras oleva seadme hinnast, leiame vastavate fiktiivsete tunnuste kordajate vahe:

$$-40{,}47 - (-16{,}11) = -24{,}36.$$

Halvas seisukorras oleva iPodi hind on 24,36 dollarit väiksem keskmises seisukorras oleva iPodi hinnast.

[a] Interneti kauplemiskeskkond `http://www.ebay.com/`

Fiktiivsete tunnuste statistilist olulisust kontrollitakse samamoodi t-testiga nagu kvantitatiivsete seletavate tunnuste korral. Kuid mitteolulisi fiktiivseid tunnuseid ei tohi mudelist üksikult eemaldada. Fiktiivsete tunnuste komplekt vastab ühele kvalitatiivsele tunnusele ning eemaldada võib vaid kõik sellele tunnusele vastavad fiktiivsed tunnused korraga. Seda tehakse tavaliselt siis, kui kõikidele kvalitatiivse tunnuse väärtustele vastavad fiktiivsed tunnused on mitteolulised. Sellisel juhul ei ole tõestatud, et vastav kvalitatiivne tunnus mõjutab sõltuvat tunnust.

Näide 9.29. Eesti noorte säästmisharjumused

N09Regressioon
N9.29

2015. aasta kevadel kaitsti TTÜ-s magistritöö, milles uuriti Eesti noorte säästmisharjumusi (Tiitso, 2015). Magistritöö jaoks viidi läbi küsitlus 18–35-aastaste noorte seas, sellele vastas 1006 isikut. Üheks küsimuseks oli „Kui suur on Teie hädareserv ootamatute kulude katteks (eurodes)?“. Analüüsime, kuidas hädareservi suurus sõltub järgmistest tunnustest:

sissetulek ST — viimase 6 kuu keskmine sissetulek (eurot);
säästuprotsent SP — mitu protsenti on keskmiselt säästetud viimase 6 kuu sissetulekust;
laste arv L;
haridustase:
0 — alg- või põhiharidus,
1 — keskharidus,
2 — keskeriharidus,
3 — kõrgharidus.

Kuna haridustase on kvalitatiivne tunnus, millel on neli erinevat väärtust, tuleb mudeli koostamiseks luua kolm fiktiivset tunnust:

$$D_1 = \begin{cases} 1, & \text{kui haridustase on 1;} \\ 0, & \text{kui haridustase ei ole 1;} \end{cases}$$
$$D_2 = \begin{cases} 1, & \text{kui haridustase on 2;} \\ 0, & \text{kui haridustase ei ole 2;} \end{cases}$$
$$D_3 = \begin{cases} 1, & \text{kui haridustase on 3;} \\ 0, & \text{kui haridustase ei ole 3.} \end{cases}$$

Hinnatav mudel on

$$y = b + a_1\mathrm{ST} + a_2\mathrm{SP} + a_3\mathrm{L} + \gamma_1 D_1 + \gamma_2 D_2 + \gamma_3 D_3 + \varepsilon.$$

Tabelarvutuses tuleb mudeli hindamiseks luua andmete tabelisse uued veerud fiktiivsete tunnuste jaoks:

Haridus	D1	D2	D3
0	0	0	0
1	1	0	0
2	0	1	0
3	0	0	1

Excelis tehtud regressioonanalüüsi tulemusena saame järgmise aruande (välja on jäetud ANOVA tabel ja usalduspiirid):

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,646
R Square	0,418
Adjusted R Square	0,414
Standard Error	1089,5
Observations	1006

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	−383	137	−2,80	0,0052
Sissetulek ST	0,535	0,056	9,58	7,6E-21
Säästuprotsent SP	59,66	3,46	17,22	2,1E-58
Laste arv L	125,6	55,9	2,25	0,025
D1	245	144	1,70	0,089
D2	224	161	1,40	0,162
D3	632	145	4,37	1,3E-05

Sissetulek ja säästuprotsent on statistiliselt olulised nivool 0,01, laste arv nivool 0,05. Haridustasemele vastavate fiktiivsete tunnuste statistilise olulisuse kontrollimine:

D_1: $p = 0{,}089 < 0{,}1$, on oluline nivool 0,1;

D_2: $p = 0{,}162 > 0{,}1$, ei ole statistiliselt oluline;

D_3: $p = 1{,}3 \cdot 10^{-5} < 0{,}01$, on oluline nivool 0,01.

Kuigi keskeriharidusele vastav fiktiivne tunnus D_2 ei ole statistiliselt oluline, tuleb see mudelisse jätta, sest haridustase kui kvalitatiivne tunnus on statistiliselt oluline.

Kõrgharidusega noortel (D_3) on hädareserv keskmiselt 632 eurot suurem kui alg- või põhiharidusega noortel. Keskharidusega noortel (D_1) on hädareserv keskmiselt 245 eurot suurem kui alg- või põhiharidusega noortel. See, kas keskeriharidusega noortel (D_2) on hädareservi suurus erinev alg- või põhiharidusega noortest, tõestatud ei ole.

Vaatame, mis juhtub, kui me näites 9.29 eemaldame mudelist keskeriharidusele vastava mitteolulise fiktiivse tunnuse D_2. Sellisel juhul jäävad mudelisse fiktiivsed tunnused D_1 ja D_3 ning haridustasemete kodeerimine on esitatud tabelis 9.15.

Tabel 9.15. Haridustasemele vastavad fiktiivsed tunnused ilma keskeriharidusele vastava tunnuseta D_2

Haridustase	Haridustaseme kood	D_1	D_3
Alg- ja põhiharidus	0	0	0
Keskharidus	1	1	0
Keskeriharidus	2	0	0
Kõrgharidus	3	0	1

Nagu näeme, on alg- ja põhihariduse ning keskerihariduse korral fiktiivsete tunnuste väärtused ühesugused ning mudelis nendel vahet ei tehta. See tähendab, et oleme alg- ja põhihariduse ning keskerihariduse kokku pannud. Niimoodi grupeerida ei ole loogiline.

Kvalitatiivse tunnuse väärtused, mille jaoks mudelis fiktiivne tunnus puudub, on üheskoos grupeeritud baasväärtuseks.

Kui mõnele kvalitatiivse tunnuse tasemele vastav fiktiivne tunnus ei ole statistiliselt oluline, võib kvalitatiivse tunnuse tasemeid vähendada, grupeerides mõned tasemed kokku. Näiteks hariduse korral võib kokku võtta keskhariduse ning keskerihariduse. Sellisel juhul tuleb luua uus fiktiivne tunnus

$$D_{12} = \begin{cases} 1, & \text{kui haridustase on 1 või 2} \\ 0, & \text{kui haridustase ei ole 1 või 2.} \end{cases}$$

Sellisel juhul oleme haridusel eristanud kolme taset ning vastavad fiktiivsed tunnused on esitatud tabelis 9.16.

Tabel 9.16. Haridustasemete ümbergrupeerimine

Haridustase	Haridustaseme kood	D_{12}	D_3
Alg- ja põhiharidus	0	0	0
Keskharidus või keskeriharidus	1 või 2	1	0
Kõrgharidus	3	0	1

N09Regressioon
N9.30

Näide 9.30. Eesti noorte säästmisharjumused II

Kuna näites 9.29 oli keskeriharidusele vastav fiktiivne tunnus D_2 statistiliselt mitteoluline, eristame hädareservi mudelis kolme haridustaset nagu tabelis 9.16. Hindame mudelit

$$y = b + a_1\mathrm{ST} + a_2\mathrm{SP} + a_3\mathrm{L} + \gamma_1 D_{12} + \gamma_3 D_3 + \varepsilon,$$

kus y on hädareservi suurus, ST sissetulek, SP säästmisprotsent ning L laste arv. Fiktiivne tunnus D_{12} vastab kesk- ja keskeriharidusele ning D_3 kõrgharidusele, baasväärtuseks on alg- ja põhiharidus. Excelis tehtud regressioonanalüüsi tulemuseks on järgmine aruanne (välja on jäetud ANOVA tabel ja usalduspiirid):

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,646
R Square	0,418
Adjusted R Square	0,415
Standard Error	1088,9
Observations	1006

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	−383	137	−2,800	0,0052
Sissetulek ST	0,5343	0,0558	9,582	7,2E-21
Säästuprotsent SP	59,64	3,46	17,232	1,8E-58
Laste arv L	125,3	55,9	2,243	0,025
D12	239	141	1,702	0,089
D3	633	144	4,378	1,3E-05

Korrigeeritud determinatsioonikordaja $R_a^2 = 0{,}415$, mis on veidi suurem kui näites 9.29.
Tunnuse D_{12} kordaja näitab, et kesk- või keskeriharidusega noortel on hädareserv kesmiselt 239 euro võrra suurem kui alg- või põhiharidusega noortel.

Kvalitatiivseid tunnuseid võib mudelis olla ka rohkem kui üks, siis luuakse iga kvalitatiivse tunnuse jaoks oma fiktiivsete tunnuste komplekt. Igas komplektis on fiktiivsete tunnuste arv ühe võrra väiksem kui vastava kvalitatiivse tunnuse tasemete arv.

Näide 9.31. Kas ametiühingusse kuulumine mõjutab töötasu?


N09Regressioon
N9.31

F. Vella ja M. Verbeek analüüsisid oma 1988. aastal ilmunud artiklis, kas ametiühingusse kuulumine mõjutab noorte meeste töötasu (Vella ja Verbeek, 1998). Valimisse kuulus 545 täistööajaga töötavat meesterahvast, kes lõpetasid kooli 1980. aastal ja olid vaatluse all aastatel 1980–1987. Andmed võeti USA riiklikust pikaajalisest uuringust (*National Longitudinal Survey*). Kuna iga isiku jaoks olid kaheksa aasta andmed, siis valimi maht on $545 \cdot 8 = 4360$. Ametiühingusse kuulumise või mittekuulumise määras ära vastus küsimusele, kas palk oli fikseeritud kollektiivse palgalepinguga või mitte.

Palkade jaotus on tavaliselt asümmeetriline ja seepärast modelleeritakse enamasti palga naturaallogaritmi, sest siis väheneb üksikute suurte palkade mõju. Seletavad tunnused, mis mudelisse võeti:

KOOL — haridustee pikkus aastates;
STAAZ — tööstaaž aastates;
STAAZ^2 — tööstaaž ruudus;
AFR — kui afroameeriklane, siis 1, vastasel juhul 0 (fiktiivne tunnus);
AB — kui abielus, siis 1, vastasel juhul 0 (fiktiivne tunnus);
AMÜ — kui ametiühingu liige, siis 1, vastasel juhul 0 (fiktiivne tunnus).

Seletavate tunnuste hulgas on haridustee pikkus ning tööstaaž ja selle ruut intervallskaalas, ülejäänud kolm on kvalitatiivsed tunnused. Kuna kvalitatiivsed tunnused on kõik kaheväärtuselised, siis lisaks fiktiivsed tunnuseid luua pole vaja.

Mudel, mida hindame, on

$$\ln \text{PALK} = b + a_1\,\text{KOOL} + a_2\,\text{STAAZ} + a_3\,\text{STAAZ}^2 + \gamma_1\,\text{AFR} + \\ + \gamma_2\,\text{AB} + \gamma_3\,\text{AMÜ} + \varepsilon.$$

Excelis tehtud regressioonanalüüsi tulemus on järgmine (ilma ANOVA tabeli ja usalduspiirideta):

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,432
R Square	0,186
Adjusted R Square	0,185
Standard Error	0,481
Observations	4360

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	−0,0240	0,0630	−0,38144	0,703
KOOL	0,09876	0,00460	21,45849	3,51E-97
STAAZ	0,0889	0,0101	8,80158	1,91E-18
STAAZ2	−0,00283	0,000707	−4,00335	6,35E-05
AFR	−0,1469	0,0232	−6,32664	2,76E-10
AB	0,1075	0,0157	6,850532	8,39E-12
AMÜ	0,1807	0,0171	10,56947	8,46E-26

Kõik mudelisse võetud tunnused on statistiliselt olulised nivool 0,01. Näeme, et afroameeriklastel on palk väiksem (kordaja on negatiivne), abielus olemine suurendab palka ja ametiühingus olemine suurendab samuti (kordajad on positiivsed). Tööstaaži kasvades palk suureneb, kuid kasv aeglustub, sest ruutliikme kordaja on negatiivne (vt joonis 9.16 (b)).
Kordajate arvväärtuste tõlgendamine on mõnevõrra komplitseeritud, sest sõltuvaks tunnuseks on palga naturaallogaritm. Võime näiteks öelda, et ametiühingusse kuulumine suurendab palga naturaallogaritmi 0,1807 võrra. Seda, kuidas leida mõju palgale, on näidatud lisas A.13. Arvutame valemi (A.81) järgi:

$$e^{0,1807} - 1 \approx 0,198.$$

Ametiühingusse kuulumine suurendab palka keskmiselt 19,8%. Sama valemi (A.81) järgi arvutades saame, et afroameeriklastel on palk keskmiselt 13,7% väiksem ning abielus olijatel 11,4% suurem.

9.19. Ühikute teisendamine

Regressioonmudeli hindamiseks kasutatavad andmed on alati esitatud konkreetsetes ühikutes. Rahaliste suuruste esitamiseks kasutatakse erinevaid valuutasid (eurod, dollarid), meetrite ja liitrite asemel kasutatakse USA-s näiteks jalgu või galloneid. Lisaks võib kasutada ühikute kordseid: tuhanded eurod, miljonid eurod.

Olgu tunnus X Eesti SKP miljonites eurodes. 2014. aasta IV kvartalis oli $x = 5129$ miljonit eurot. Kui $\tilde{X}$ on Eesti SKP miljardites eurodes, siis

$$\tilde{x} = 0,001x = 0,001 \cdot 5129 = 5,129 \text{ mld eurot.} \qquad (9.89)$$

Olgu Y keskmine brutokuupalk kroonides, mis 2010. aastal oli 13 184 krooni. Kui $\tilde{Y}$ on keskmine brutokuupalk eurodes, siis teisendamiseks

peame kasutama Eesti krooni ja euro vahelist kurssi, $1\,€ \approx 15{,}65\,\mathrm{kr}$.

$$\tilde{y} = \frac{1\,€}{15{,}65\,\mathrm{kr}} \cdot y = \frac{1\,€}{15{,}65\,\mathrm{kr}} \cdot 13\,184\,\mathrm{kr} \approx 842\,€. \tag{9.90}$$

Kordajaid 0,001 ja 1/15,65 valemites (9.89) ja (9.90) nimetatakse mastaabikordajateks (*scale factor*). Mastaabikordaja võib olla ühikuta (valemis (9.89)) või ühikuga suurus (valemis (9.90)).

Mastaabikordaja

Mastaabikordaja on kordaja, millega tuleb ühtedes ühikutes esitatud suurust korrutada, et saada sama suuruse väärtus teistes ühikutes.

Vaatame nüüd, kuidas tuleb ühikute teisendamisel teisendada regressioonmudeli parameetreid. Olgu meil USA-s müüdud majade põhjal leitud regressioonmudel, mis kirjeldab maja hinna Y (\$) sõltuvust maja pindalast X (ruutjalgades):

$$\hat{y} = 42000 + 38x. \tag{9.91}$$

Mudeli vabaliikme ühik on sama, mis sõltuval tunnusel Y, seega, vabaliige on 42 000 \$. Korrutise $38x$ ühik peab samuti olema dollarid, sest liita saab ainult samades ühikutes olevaid suurusi. Kuna X on mõõdetud ruutjalgades (ft^2), siis lineaarliikme kordaja on $38\,\$/\mathrm{ft}^2$. See garanteerib, et Y tuleb dollarites. Näiteks, kui maja pindala on $1500\,\mathrm{ft}^2$, siis hind

$$\hat{y} = 42\,000\,\$ + 38\frac{\$}{\mathrm{ft}^2} \cdot 1500\,\mathrm{ft}^2 = 99\,000\,\$.$$

Soovime nüüd maja hinda esitada eurodes kursiga, mis oli mudeli koostamise ajal $1\,\$ = 0{,}7\,€$. Mastaabikordaja on siis $0{,}7\,€/\$$:

$$\tilde{y} = 0{,}7\frac{€}{\$}y. \tag{9.92}$$

Uus vabaliige on

$$0{,}7\frac{€}{\$} \cdot 42\,000\,\$ = 29\,400\,€. \tag{9.93}$$

Uus lineaarliikme kordaja

$$38\frac{\$}{\mathrm{ft}^2} \cdot 0{,}7\frac{€}{\$} = 38 \cdot 0{,}7\frac{€}{\mathrm{ft}^2} = 26{,}6\frac{€}{\mathrm{ft}^2} \tag{9.94}$$

ja maja hinna mudel, kus hind $\tilde{Y}$ on eurodes,

$$\hat{\tilde{y}} = 29400 + 26{,}6x. \tag{9.95}$$

Valemitest (9.93) ja (9.94) näeme, et kui muudame sõltuva tunnuse ühikuid, siis vabaliiget ja lineaarliikme kordajat tuleb korrutada vastava mastaabikordajaga.

Nüüd soovime maja pindala teisendada ruutmeetriteks, arvestades, et $1\,\mathrm{ft}^2 = 0{,}093\,\mathrm{m}^2$:

$$\tilde{x} = 0{,}093\frac{\mathrm{m}^2}{\mathrm{ft}^2}x. \tag{9.96}$$

Vabaliikmele see teisendus ei mõju. Uus lineaarliikme kordaja

$$26{,}6\frac{€}{\mathrm{ft}^2}\cdot\frac{1\,\mathrm{ft}^2}{0{,}93\,\mathrm{m}^2} = \frac{26{,}6}{0{,}93}\frac{€^2}{\mathrm{m}^2} \approx 286\frac{€^2}{\mathrm{m}^2}. \tag{9.97}$$

Valemitest (9.96) ja (9.97) näeme, et seletava tunnuse ühikute teisendamisel tuleb lineaarliikme kordajat jagada vastava mastaabikordajaga. Maja hinna mudel, kus hind $\tilde{Y}$ on eurodes ja maja pindala $\tilde{X}$ ruutmeetrites:

$$\hat{\tilde{y}} = 29400 + 286\tilde{x}. \tag{9.98}$$

Parameetrite teisendamine

Olgu meil hinnatud regressioonmudel

$$y = b + a_1x_1 + a_2x_2 + \ldots + a_kx_k + \varepsilon. \tag{9.99}$$

Kui soovime teisendada ühikuid, nii et

$$\tilde{y} = vy, \quad \tilde{x}_l = w_lx_l, \tag{9.100}$$

kus v ja w_l on mastaabikordajad, siis uue mudeli parameetrid avalduvad mudeli (9.99) parameetrite kaudu järgmiselt:

$$\tilde{b} = vb, \quad \tilde{a}_l = \frac{v}{w_l}a_l. \tag{9.101}$$

Kuna parameetrite standardvead on samades ühikutes, mis parameetrid ise, tuleb uutes ühikutes esitatud standardvigade jaoks kasutada samu teisendusi (9.101)

$$\tilde{se}(b) = v\,se(b), \quad \tilde{se}(a_l) = \frac{v}{w_l}se(a_l). \tag{9.102}$$

Hüpoteeside kontrollimiseks vajalikud t-statistikud aga ei muutu, need on **mastaabi-invariantsed**:

$$t_b = \frac{vb}{v\,se(b)} = \frac{b}{se(b)}, \quad t_{a_l} = \frac{\frac{v}{w_l}a_l}{\frac{v}{w_l}se(a_l)} = \frac{a_l}{se(a_l)}. \tag{9.103}$$

Järelikult jäävad kõik otsused parameetrite statistilise olulisuse kohta samaks. Kuna determinatsioonikordaja R^2 on ühikuta suurus, siis ka see on mastaabi-invariantne, mudeli kirjeldusvõime jääb ühikute teisendamisel samaks.

Näide 9.32. Kasutatud autode hind

N09Regressioon
N9.32

Millest sõltub kasutatud auto hind? Kindlasti sõltub see auto vanusest, läbisõidust, mudelist, tõenäoliselt ka mootori mahust. Hinna mudeli saamiseks on kasutatud 804 General Motorsis toodetud autot. Mudelite valik oli Buick (80), Cadillac (80), Chevrolet (320), Pontiac (150), SAAB (114) ja Saturn (60). Kõik autod olid väga heas seisukorras ning vähem kui aasta vanad. (Kuiper ja College, 2008) Mudelis olevad tunnused:

Y — soovituslik jaehind, \$;
X_1 — läbisõit miilides;
X_2 — mootori maht liitrites;
X_3 — uste arv.

Mudeliks saadi

$$y = 13684 - 0{,}1622x_1 + 4899x_2 - 1135x_3 + \varepsilon, \qquad (9.104)$$
$$R^2 = 0{,}339.$$

Teisendame selle mudeli kujule, kus hind oleks eurodes, läbisõit kilomeetrites ja mootori maht kuupsentimeetrites. Selleks leiame vastavad mastaabikordajad. Valuuta muutmiseks kasutame 2005. aasta keskmist kurssi: $1\,\$ = 0{,}8\,€$, $1\,\text{miil} = 1{,}609\,\text{km}$ ja $1\,\text{l} = 1000\,\text{cm}^3$. Vastavad mastaabikordajad on siis

$$v = 0{,}8\frac{€}{\$}, \quad w_1 = 1{,}609\frac{\text{km}}{\text{miil}}, \quad w_2 = 1000\frac{\text{cm}^3}{\text{l}}.$$

Uue mudeli parameetrid nüüd valemitest (9.101):

$$\begin{aligned}
\tilde{b} &= 13684 \cdot 0{,}8 \approx 10947,\\
\tilde{a}_1 &= -0{,}1622 \cdot \frac{0{,}8}{1{,}609} \approx -0{,}0806,\\
\tilde{a}_2 &= 4899 \cdot \frac{0{,}8}{1000} \approx 3{,}919,\\
\tilde{a}_3 &= -1135 \cdot 0{,}8 \approx -907{,}9
\end{aligned}$$

ja uus mudel

$$\tilde{y} = 10947 - 0{,}0806\tilde{x}_1 + 3{,}919\tilde{x}_2 - 907{,}9\tilde{x}_3 + \varepsilon, \quad R^2 = 0{,}339,$$

kus $\tilde{y}$ on hind eurodes, $\tilde{x}_1$ läbisõit kilomeetrites, $\tilde{x}_2$ mootori maht kuupsentimeetrites ning $\tilde{x}_3$ uste arv.

Näeme, et auto, mille läbisõit on ühe km võrra suurem, maksab 0,08 eurot vähem ehk läbisõidu suurenemine 1000 km võrra vähendab hinda 80 euro võrra. Auto, mille mootori maht on 1 cm^3 võrra suurem, maksab ligikaudu 3,9 eurot rohkem. Uste arv võis olla kas 2 või 4 ning huvitav on märkida, et odavamad on need autod, millel uste arv on väiksem. Nelja uksega auto on kahe uksega autost ligikaudu 1800 eurot odavam.

9.20. Standardiseeritud kordajad

Regressioonmudeli argumenttunnused on enamasti erinevates ühikutes ning seetõttu me ei saa võrrelda, milline tunnus mõjutab funktsioontunnust rohkem, milline vähem. Näites 9.12 tööjõu pakkumise kohta saime mudeli

$$\widehat{\text{TUNNID}} = 2445 - 47{,}6\ \text{TTASU} + 0{,}0264\ \text{VARAD} - 8{,}66\ \text{VANUS}, \tag{9.105}$$

kus TUNNID on töötundide arv aastas, TTASU tunnitasu dollarites, VARAD majapidamise varade suurus dollarites ning VANUS isiku vanus aastates. Kuigi me võime öelda, et tunnitasu suurenemine 1 dollari võrra mõjutab tööjõu pakkumist rohkem kui varade suurenemine 1 dollari võrra, pole selline võrdlus korrektne. Sest keskmine tunnitasu on 2,77 dollarit, aga keskmine varade suurus 6265 dollarit ning järelikult ühe dollariline muutus on tunnitasu korral suur muutus, aga varade korral väike. Vanuse kordajat $-8{,}66$ ei saa me aga teiste kordajatega üldse võrrelda. Selgituseks paneme kordajad kirja koos ühikutega:

$$\begin{aligned}
\text{TTASU}: &\quad -47{,}6\,\frac{\text{tund}}{\$},\\
\text{VARAD}: &\quad 0{,}0264\,\frac{\text{tund}}{\$},\\
\text{VANUS}: &\quad -8{,}66\,\frac{\text{tund}}{\text{aasta}}.
\end{aligned}$$

Me ei saa võrrelda, kumma kordaja absoluutväärtus on suurem, kas 47,6 tund/\$ või 8,66 tund/aasta, sest dollarid ja aastad ei ole võrreldavad.

Et mudelis olevate seletavate tunnuste kordajaid omavahel võrrelda, peavad need olema kas samades ühikutes või siis ühikuta suurused. Ühikuta kordajate saamiseks teisendatakse kõik tunnused standardiseeritud skaalasse (vaata alapeatükk 3.6). Selleks tuleb iga objekti jaoks leida kõigi tunnuste z-skoorid:

$$z_{Yi} = \frac{y_i - \bar{y}}{s(Y)},\ z_{li} = \frac{x_{li} - \bar{x}_l}{s(X_l)}, \tag{9.106}$$

kus $\bar{y}$ ja $\bar{x}_l$ on vastava tunnuse aritmeetilised keskmised ning $s(Y)$ ja $s(X_l)$ valimi standardhälbed. Seejärel hinnatakse regressioonmudelit, kus kasutatakse nii sõltuva kui seletavate tunnuste z-skoore. Standardiseeritud skaalas esitatud tunnuste regressioonmudeli kordajaid nimetatakse **standardiseeritud kordajateks**. Kuna tunnuse X standardhälve $s(X)$ on samades ühikutes, mis tunnus ise, on standardiseeritud skaalas esitatud tunnused ühikuta ning samuti on ühikuta vastava regressioonmudeli kordajad.

Standardiseeritud kordajad

Lineaarse regressioonmudeli **standardiseeritud kordaja** näitab, mitme standardhälbe võrra muutub funktsioontunnus, kui argumenttunnus suureneb ühe standardhälbe võrra.

Kuna standardiseeritud skaalas esitatud suuruse aritmeetiline keskmine on null, siis valemist (9.16) järeldub, et standardiseeritud tunnustega lineaarse mudeli vabaliige on 0, s.t regressioonjoon läbib nullpunkti.

Näide 9.33. Tööjõu pakkumine ja standardiseeritud kordajad

N09Regressioon
N9.33,34

Viime näites 9.12 toodud andmete põhjal uuesti läbi lineaarse regressioonmudeli hindamise, kasutades seekord standardiseeritud skaalas esitatud väärtusi. Selleks tuleb eelnevalt leida kõigi tunnuste jaoks valimi keskmine ning standardhälve ja siis kasutada tunnuste teisendamiseks valemit (9.106).
Näiteks töötundide korral $\bar{y} = 2137{,}4$ ja $s(Y) = 64{,}03$. Objekti 1 korral on töötundide arv 2157. Vastav z-skoor

$$z_{\text{TUNNID}_1} = \frac{2157 - 2137{,}4}{64{,}0} \approx 0{,}3064.$$

Tunnitasu korral $\bar{x} = 2{,}772$ ja $s(X) = 0{,}4565$. Objekti 1 korral on tunnitasu 2,905. Vastav z-skoor

$$z_{\text{TTASU}_1} = \frac{2{,}905 - 2{,}772}{0{,}4565} \approx 0{,}2909.$$

Niimoodi arvutame kõikide tunnuste z-skoorid iga objekti jaoks. Järgnevas tabelis on esitatud arvutustulemused kolme esimese elemendi jaoks. Kahel viimasel real on tunnuste aritmeetilised keskmised $\bar{x}$ ja standardhälbed s.

	Algandmed			
	TUNNID	TTASU	VARAD	VANUS
	2157	2,905	7250	38,5
	2174	2,97	7744	39,3
	2062	2,35	3068	40,1
	...	...	...	
$\bar{x}$	2137,4	2,772	6265	39,35
s	64,03	0,4565	2913	4,222

	Standardiseeritud skaalas		
TUNNID	TTASU	VARAD	VANUS
0,3064	0,2909	0,3380	−0,2016
0,5719	0,4332	0,5076	−0,0121
−1,1774	−0,9250	−1,0975	0,1773
...	...	...	...

Regressioonanalüüsis kasutame nüüd standardiseeritud väärtusi ja arvestame sellega, et regressioonjoon peab läbima nullpuntki (vabaliige puudub). Mudeliks saame

$$\hat{z}_{\text{TUNNID}} = -0{,}339 z_{\text{TTASU}} + 1{,}202 z_{\text{VARAD}} - 0{,}571 z_{\text{VANUS}}. \tag{9.107}$$

Kuna selle mudeli kordajad on ühikuta suurused, siis võime järeldada, et kõige rohkem mõjutab töötundide arvu varade suurus: varade suurenemisel ühe standardhälbe võrra suureneb töötundide arv 1,202 standardhälbe võrra. Vanuse suurenemisel ühe standardhälbe võrra töötundide arv väheneb 0,571 standardhälbe võrra ning tunnitasu suurenemisel ühe standardhälbe võrra töötundide arv väheneb 0,339 standardhälvet.

Seos standardiseeritud ja standardiseerimata kordajate vahel

Standardiseeritud kordajate leidmiseks ei pea tingimata läbi viima regressioonanalüüsi standardiseeritud tunnustega. Kehtib seos

$$\tilde{a}_l = a_l \frac{s(X_l)}{s(Y)}, \tag{9.108}$$

kus $\tilde{a}_l$ on tunnuse X_l standardiseeritud kordaja, a_l vastava tunnuse kordaja standardiseerimata regressioonmudelis ning $s(X_l)$ ja $s(Y)$ vastavalt tunnuse X_l ning funktsioontunnuse Y valimite standardhälbed. Kui regressioonanalüüsiks on kasutatud standardiseerimata tunnuseid, siis tuleb lisaks leida kõikide tunnuste standardhälbed ning standardiseeritud kordajad saab arvutada valemist (9.108).

Näide 9.34. Tööjõu pakkumine ja standardiseeritud kordajate arvutus

N09Regressioon
N9.33,34

Leiame tööjõu pakkumise mudeli (9.107) kordajad, kasutades standardiseerimata tunnuste korral leitud mudeli (9.105) kordajaid ning näites 9.33 toodud standardhälbeid:

$$\text{TTASU:} \qquad \tilde{a}_1 = -46{,}7 \cdot \frac{0{,}4565}{64{,}03} \approx -0{,}339,$$

$$\text{VARAD:} \qquad \tilde{a}_2 = 0{,}0264 \cdot \frac{2913}{64{,}03} \approx 1{,}202,$$

$$\text{VANUS:} \qquad \tilde{a}_3 = -8{,}66 \cdot \frac{4{,}222}{64{,}03} \approx -0{,}571.$$

Nagu näha, tulevad kordajad samad, mis regressioonanalüüsil standardiseeritud skaalasse teisendatud tunnustega näites 9.33.

Majanduses kasutatakse sageli mudeleid, kus nii sõltuv tunnus kui ka seletavad tunnused on logaritmitud:

$$\ln y = b + a_1 \ln x_1 + a_2 \ln x_2 + \ldots + a_k \ln x_k + \varepsilon. \tag{9.109}$$

Sellise mudeli korral näitab kordaja a_l, mitu protsenti suureneb y, kui x_l suureneb 1%. Kuna selle mudeli kordajad on ühikuta, siis nende väärtused on võrreldavad. Üht sellist mudelit vaatame alapeatükis 9.22.

9.21. Regressioonanalüüsi etapid ja mudeli korrektne esitamine

Regressioonanalüüsi põhieesmärgiks on vaatlusandmeid kasutades saada võimalikult palju informatsiooni uuritava nähtuse kohta. Loetleme kokkuvõtlikult mudeli saamiseks läbitavad **etapid**.

1. Probleemi püstitamine.
2. Tunnuste valik: mis on funktsioontunnus ja millised on potentsiaalsed seletavad tunnused (regressorid).
3. Mudeli üldkuju valik: lineaarne või mittelineaarne. Kui mittelineaarne, siis millise kujuga.
4. Andmete kogumine.
5. Mudeli parameetrite hindamine regressioonanalüüsi abil.
6. Mudeli diagnostika: mudeli statistilise olulisuse testimine, parameetrite statistilise olulisuse testimine, jääkide analüüs.
7. Vajadusel mudeli korrigeerimine ning uue mudeli hindamine ja diagnostika.
8. Lõpliku mudeli tõlgendamine.

9. Soovi korral arvutused mudeli põhjal (prognoos).

Kui on vaja leida sobiv seletavate tunnuste komplekt ning tuleb hinnata mitmeid erineva seletavate tunnuste komplektiga mudeleid, siis tabelarvutuses on see üpris tülikas. Tabelarvutuses peavad kõik mudelisse võetavad seletavad tunnused asuma kõrvuti veergudes ning erinevate tunnuste valikuks tuleb veergusid ümber tõsta. Sellisel juhul on soovitav kasutada vabavara Gretl[3]. Lühike õpetus programmi Gretl kasutamiseks on õpiku autori kodulehel[4].

Mudeli esitamine

Regressioonanalüüsi tulemus esitatakse vastava mudelina, mille juurde märgitakse:

- valimi maht n;
- determinatsioonikordaja R^2;
- parameetrite standardvead se (sulgudes parameetrite all);
- kasutatud tähistuste seletused koos ühikutega.

$$y = \underset{(se(b))}{b} + \underset{(se(a_1))}{a_1} x_1 + \ldots + \underset{(se(a_k))}{a_k} x_k + \varepsilon, \quad R^2 = \ldots$$

$$n = \ldots$$

Ümardamine

Mudeli parameetrite ja standardvigade väärtuste esitamisel ei tohiks anda liigseid numbrikohti, väärtused tuleb sobivalt ümardada. Standardviga ja determinatsioonikordaja esitatakse tavaliselt kolme tüvenumbriga[5]. Mudeli parameetrid ümardatakse selle kohani, kus asub standardvea kolmas tüvenumber. Kolm tüvenumbrit on näiteks arvudes 0,125, 0,100 ja 120. Mõned näited on esitatud tabelis 9.17. Kui tulemustega on vaja teha arvutusi, võib säilitada mõned lisakohad vältimaks ümardamisega seotud ebatäpsusi.

Tabel 9.17. Näited parameetrite ja standardvigade ümardamisest

Ümardamata		Ümardatud	
Parameetri hinnang	Standardviga	Parameetri hinnang	Standardviga
10,17926616	1,130859143	10,18	1,13
0,001727347	0,000085672	$172{,}73 \cdot 10^{-5}$	$8{,}57 \cdot 10^{-5}$
12530,03845	5586,811232	$12{,}53 \cdot 10^{3}$	$5{,}59 \cdot 10^{3}$

Näitena esitame tööjõu pakkumise mudeli (näited 9.12 ja 9.16):

$$\text{TUNNID} = \underset{(93,6)}{2444{,}8} - \underset{(23,0)}{47{,}6}\ \text{TTASU} + \underset{(0,00393)}{0{,}02641}\ \text{VARAD} - \underset{(1,71)}{8{,}66}\ \text{VANUS} + \varepsilon, \quad n = 39, \quad R^2 = 0{,}715,$$

[3] `http://gretl.sourceforge.net/`

[4] `http://www.sauga.pri.ee/gretl/`

[5] Mõõtmisteoorias soovitatakse viga esitada kahe tüvenumbriga ja mõõtmistulemus ümardada vea teise tüvenumbrini.

kus TUNNID on töötatud tundide arv aastas, TTASU tunnitasu ($), VARAD majapidamise varade suurus ($) ning VANUS isiku vanus aastates.

Kui hinnatakse mitut erineva seletatavate tunnuste komplektiga mudelit ja soovitakse neid võrrelda, siis ei ole mudeli matemaatilise kuju esitamine otstarbekas, sest sellisel kujul on lugejal raske erinevaid tunnuste komplekte võrrelda. Mudelite kordajad ning standardvead esitatakse sellisel juhul tabelis. Tärnidega märgitakse, millised kordajad on statistiliselt olulised: *** tähistab olulisust nivool 0,01, ** olulisust nivool 0,05 ning * olulisust nivool 0,1. Valimi maht, determinatsioonikordaja ja korrigeeritud determinatsioonikordaja esitatakse tabeli viimastel ridadel. Tabelis 9.18 on esitatud käibe mudeli hindamise tulemused näidetest 9.13 ja 9.20. E on majanduskasvu indeks, M turunduskulud (tuhat eurot), C ühiku kulu indeks ning P toote hind eurodes.

Tabel 9.18. Regressioonmudelite hindamise tulemused tabeli kujul

	Mudel 1	Mudel 2	Mudel 3	Mudel 4
Vabaliige	−3989	−3527	−2642	−2639
	(774)	(514)	(679)	(693)
E	50,13***	42,88***	40,10***	39,46***
	(7,45)	(5,08)	(4,98)	(5,21)
M		332,9***	469,7***	478,5***
		(66,4)	(96,6)	(99,9)
C			−6,29*	−4,35
			(3,40)	(4,85)
P				68
				(119)
R^2	0,716	0,885	0,906	0,908
R_a^2	0,700	0,872	0,888	0,883
n	20	20	20	20

Mõnikord tuuakse sulgudes standardvigade asemel kas t-statistikud või neile vastavad olulisuse tõenäosused. Seepärast tuleb alati märkida, mis suurused on sulgudes.

9.22. Näide: autotööstuse Cobbi-Douglase tootmisfunktsioon

Tootmisfunktsioon kirjeldab maksimaalset võimalikku kogutoodangut antud tootmissisendite korral. Tüüpilisteks tootmissisenditeks on ka-

pital ja tööjõud. Sagedasti kasutatakse Cobbi-Douglase tootmisfunktsiooni

$$q(K,L) = AK^{\alpha}L^{\beta}, \tag{9.110}$$

Cobbi-Douglase tootmisfunktsioon

kus q on tootmismaht, K kapital ja L tööjõud. Omades andmeid antud tegevusalal tegutsevate ettevõtete tootmismahu, varade ja kasutatava tööjõu kohta, on võimalik leida tootmisfunktsiooni mudel selle tegevusala ettevõtete jaoks, leides parameetrite A, α ja β väärtused. Selleks lineariseerime funktsiooni (9.110), võttes mõlemalt poolt naturaallogaritmi[6]:

$$\ln q = \ln(AK^{\alpha}L^{\beta}) = \ln A + \ln K^{\alpha} + \ln L^{\beta} = \ln A + \alpha \ln K + \beta \ln L.$$

Mudelit

$$\ln q = \ln A + \alpha \ln K + \beta \ln L, \tag{9.111}$$

mis kirjeldab seost tunnuste logaritmide vahel, nimetatakse **log-log mudeliks**[7]. Võttes kasutusele tähistused

$$y = \ln q, \tag{9.112}$$
$$x_1 = \ln K, \tag{9.113}$$
$$x_2 = \ln L, \tag{9.114}$$
$$b = \ln A, \tag{9.115}$$

saame järgmise lineaarse mudeli:

$$y = b + \alpha x_1 + \beta x_2. \tag{9.116}$$

Lineaarse mudeli (9.116) parameetrite hindamiseks on vaja eelnevalt leida tunnuste q, K ja L naturaallogaritmid (valemid (9.112)–(9.114)). Kapitali K astmenäitaja α ning tööjõu astmenäitaja β on lineaarse mudeli kordajad ning nende hinnangud saame kohe peale lineaarse mudeli hindamist. Mittelineaarse mudeli (9.110) parameeter A saadakse lineaarse mudeli vabaliikmest b, kui tehakse teisenduse (9.115) pöördteisendus:

$$A = e^{b}. \tag{9.117}$$

Cobbi-Douglase tootmisfunktsioonis (9.110) esinevate astmenäitajate tõlgendus on järgmine:

Parameetrite tõlgendus

- kapitali K astmenäitaja α näitab, mitu protsenti suureneb kogutoodang q, kui kapitali suurendada 1%;
- tööjõu L astmenäitaja β näitab, mitu protsenti suureneb kogutoodang q, kui tööjõudu suurendada 1%.

Suurust, mis seob sõltumatu ja sõltuva tunnuse protsentuaalseid

Elastsus

[6] Võib kasutada ka kümnendlogaritmi, kuid regressioonanalüüsis eelistatakse logaritmi alusel e, s.t naturaallogaritmi.

[7] Tuletame meelde, et ingliskeelses kirjanduses on naturaallogaritmi tähistuseks log.

muutusi, nimetatakse majandusteaduses funktsiooni $y(x)$ **elastsuseks** ehk **elastsuskordajaks** (lähemalt vt näiteks (Aasma ja Levin, 2013, lk 129)). Järelikult on astmenäitaja α kapitali elastsus ning β tööjõu elastsus.

Mudelis (9.110) esinevat parameetrit A interpreteeritakse kui tehnoloogiast sõltuvat konstanti: ühesuguse kapitali ja tööjõuga ettevõtetel on kogutoodangu erinevus põhjustatud erinevast tehnoloogiast.

1999. aasta 19. aprilli Postimehes ilmusid andmed 20 maailma suurima autotootja kohta: varad, tööjõud ja käive. Leiame nende andmete põhjal autotööstuse tootmisfunktsiooni kujul (9.110), kus q on käive (miljard dollarit), K varad (miljard dollarit) ja L töötajate arv. Varade all on mõeldud omakapitali.

Mudeli lineariseerimiseks leiame suuruste q, K ja L naturaallogaritmid. Toome ära fragmendi andmetabelist, kuhu on lisatud kolm veergu: $\ln K$, $\ln L$ ja $\ln q$.

N09Regressioon
P9.22

Ettevõte	Varad K, mld \$	Töötajate arv L	Käive q, mld \$	$\ln K$	$\ln L$	$\ln q$
General Motors	228888	608000	178174	12,341	13,318	12,091
Ford	279097	363900	153627	12,539	12,805	11,942
Toyota	103894	159000	95137	11,551	11,977	11,463
Daimler Benz	76191	300000	71561	11,241	12,612	11,178
Daewoo	44861	265000	71526	10,711	12,487	11,178
...	...	...	...	...	...	...

Kasutades kolmes viimases veerus leitud naturaallogaritme, hindame mudelit

$$\ln q = b + \alpha \ln K + \beta \ln L + \varepsilon. \tag{9.118}$$

Esitame Excelis tehtud regressioonanalüüsi tulemused.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,976
R Square	0,952
Adjusted R Square	0,946
Standard Error	0,1742
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	10,2600	5,1300	169,0	6,06E-12
Residual	17	0,5162	0,0304		
Total	19	10,7761			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2,002	0,508	3,944	0,00105	0,9311	3,0729
ln K	0,674	0,0685	9,833	1,98E-08	0,5292	0,8183
ln L	0,130	0,0687	1,896	0,0751	−0,0147	0,2751

Mudel tervikuna on statistiliselt oluline, F-testi olulisuse tõenäosus $6{,}06 \cdot 10^{-12} < 0{,}05$. Kapitali logaritm on statistiliselt oluline nivool 0,01 ($p = 1{,}98 \cdot 10^{-8} < 0{,}01$). Tööjõu L logaritm on statistiliselt oluline nivool 0,1: $p = 0{,}0751 < 0{,}1$. Kuna majandusteooriast on teada, et kogutoodang sõltub tööjõust, siis seda tunnust me mudelist ei eemalda ja kasutame olulisuse nivood 0,1.

Varade K naturaallogaritmi kordaja $\alpha = 0{,}674$ ja töötajate arvu L naturaallogaritmi kordaja $\beta = 0{,}130$. Vabaliige $b = \ln A = 2{,}002$, millest

$$A = e^{2{,}002} \approx 7{,}4.$$

Autotööstuse tootmisfunktsioon on

$$\hat{q} = 7{,}4K^{0{,}67}L^{0{,}13}, \quad R^2 = 0{,}952, \tag{9.119}$$

kus q on käive ja K varad miljardites dollarites ning L töötajate arv. Kapitali 1% võrra suurendades suureneb käive 0,67%. Tööjõu suurenemisel 1% võrra on käibe suurenemine 0,13%. Kui mõlemat tootmissisendit suurendada 1%, suureneb käive $0{,}67\% + 0{,}13\% = 0{,}8\%$.

Leiame mudeli põhjal käibe mudelväärtused. Näiteks General Motorsi jaoks

$$\hat{q} = 7{,}4 \cdot 228888^{0{,}67} \cdot 608000^{0{,}13} = 171192 \quad \text{mld \$}. \tag{9.120}$$

Tegelik väärtus 178 174 mld \$ on mudelväärtusest 4% suurem. Niimoodi leiame käibe mudelväärtused kõigi ettevõtete jaoks ning lisaks ka tegelike väärtuste suhtelised erinevused mudelväärtustest.

Ettevõte	Varad K, mld \$	Töötajate arv L	Käive q, mld \$	Käibe mudelväärtus $\hat{q}$, mld \$	Suhteline erinevus, %
General Motors	228888	608000	178174	171192	4%
Ford	279097	363900	153627	183018	−16%
Toyota	103894	159000	95137	84437	13%
Daimler Benz	76191	300000	71561	74419	−4%
Daewoo	44861	265000	71526	51249	40%
...	...	...	...	...	...

Näeme, et kõige suurem erinevus mudelväärtusest on Daewool. Selle käibe tegelik väärtus 71 526 mld \$ on mudelväärtusest 40% suurem.

Leiame jäägid (9.52) ja standardiseeritud jäägid (9.53). Excelis tuleb selleks regressioonanalüüs uuesti teha ning valida *Residuals* ja *Standardized Residuals*.

	Predicted ln q	*Residuals*	*Standard Residuals*
General Motors	12,051	0,0400	0,24
Ford	12,117	−0,1751	−1,06
Toyota	11,344	0,1193	0,72
Daimler Benz	11,217	−0,0392	−0,24
Daewoo	10,844	0,3334	2,02
...	...	...	...

Paneme tähele, et käibe mudelväärtuse võib leida ka lineaarse mudeli (9.118) mudelväärtusest $\ln \hat{q}$, mis on tabeli veerus *Predicted ln q*. General Motorsi korral

$$\ln \hat{q} = 12{,}051,$$
$$\hat{q} = e^{12{,}051} = 171192.$$

Tulemus on sama, nagu saime arvutusest (9.120).

Jääkide analüüs näitab, et Daewoo standardiseeritud jääk on $2{,}02 > 2$, s.t see ettevõte on ebatüüpiline. Ülejäänud ettevõtete korral on standardiseeritud jäägi absoluutväärtus väiksem kui 2.

Kas Daewoo ebatüüpilisus on hea või halb? Esialgu paistab, et hea, sest tegelik käive on oluliselt suurem kui mudelväärtus. Kuid neli kuud pärast nende andmete avaldamist Postimehes (17. august 1999), kirjutas Äripäev:

> „Daewoo lammutatakse.
> Finantsraskustes Lõuna-Korea tööstusgrupi Daewoo kreeditorid otsustasid eile lammutada riigi suuruselt teise konglomeraadi ja jätta alles vaid kuus autotootmisüksust. ... Tänavune restruktureerimisplaan näeb ette grupi võlgade ja omakapitali suhte kahandamise 1998. a 527 protsendilt 196 protsendile.“

Daewoo tegelik käive oli peaaegu sama suur kui Daimler Benzil, töötajate arv ligikaudu sama suur, kuid varad (omakapital) oluliselt väiksem. Leiame, kui palju kapitali oleks pidanud Daewool olema, et saavutada sellist käivet nagu tal oli. Selleks paneme tootmisfunktsiooni avaldisse (9.119) Daewoo tegeliku käibe $q = 71526$ ja töötajate arvu

$L = 265000$ ning avaldame varad K (miljardites dollarites):

$$\begin{aligned} 71526 &= 7{,}4K^{0{,}67} \cdot 265000^{0{,}13} \\ K^{0{,}67} &= \frac{71526}{7{,}4 \cdot 265000^{0{,}13}} \\ K &= \left(\frac{71526}{7{,}4 \cdot 265000^{0{,}13}}\right)^{1/0{,}67} \approx 73580. \end{aligned}$$

Erinevus andmetes avaldatud varadest:

$$73580 - 44861 = 28719,$$

mis peaks siis olema võlgade suurus miljardites dollarites. Suhteline erinevus

$$\frac{28719}{44861} \approx 0{,}64.$$

Mudeli põhjal hinnatud võlgade ja omakapitali suhe on 64%.

Nagu nägime, võib teatud tegevusala ettevõtete andmeid kasutades leida vastava tegevusala tootmisfunktsiooni ning siis standardiseeritud jääkide abil teha kindlaks ebatüüpilised ettevõtted.

Log-log mudelit

$$\ln y = b + a_1 \ln x_1 + \ldots + \ln a_k x_k + \varepsilon$$

kasutatakse ka mujal kui tootmisfunktsiooni leidmisel. Kuna selle mudeli kordajad $a_1, \ldots, a_k$ on elastsuskordajad, siis on sellisel kujul mudeli hindamine õigustatud juhul, kui elastsuskordajad on konstantsed: funktsioontunnuse suhtelise muudu sõltuvus argumenttunnuste suhtelistest muutustest on erinevate väärtuste korral ühesugune.

9.23. Ülesanded

9.1. Iga tunnustepaari korral määrata, kumb on sõltuv tunnus Y ja kumb seletav tunnus X:

a) maja hind ja maja pindala;
b) telesaadete vaadatavus ja reklaamiminuti hind;
c) SKP elaniku kohta ja autode arv 1000 elaniku kohta;
d) SKP ja tööealise elanikkonna suurus.

VASTUS lk 681.

Lineaarse mudeli parameetrite tõlgendamine, arvutused

9.2. Leida mudeli (9.4) põhjal, kui suur on 5-aastaste poisslaste keskmine pikkus. Kuidas muutub keskmine pikkus vanuse suurenedes?

VASTUS lk 681.

9.3. Avaldatava reklaami hind on erinevatel ajakirjandusväljaannetel erinev. See sõltub mitmest tegurist. USA-s läbiviidud uuring hõlmas

48 ajakirja. Analüüsi all oli ühekordse leheküljesuuruse neljavärvilise reklaami hind y dollarites. Argumenttunnusteks võeti tellijaskonna suurus x_1 (tuhandetes), naiste osakaal lugejate hulgas x_2 (protsentides) ja lugejate mediaansissetulek dollarites x_3. Läbiviidud regressioonanalüüs andis järgmise mudeli:

$$\hat{y} = -8643 + 5{,}28x_1 - 11{,}00x_2 + 1{,}22x_3, \quad R^2 = 0{,}694.$$

Tõlgendada mudeli parameetreid. VASTUS lk 681.

9.4. J. Kitt ja P. Sträter analüüsisid reklaami mõju USA maiustusteturul (Kitt ja Sträter, 2008). Nad kasutasid 100 erineva brändi müügiandmeid aastatest 1996–2000, mis moodustas 78% kogu maiustusteturust. Analüüs näitas, et reklaamikulud mõjutasid oluliselt müügitulu. Köhakommide jaoks said nad järgmise mudeli:

$$\hat{y} = 6{,}7x_1 + 2{,}7x_2 - 0{,}01x_3, \quad R^2 = 0{,}97,$$

kus

y — müügitulu aastas (tuhat dollarit);
x_1 — müügimaht aastas (tuhat tonni);
x_2 — kulud reklaamile aastas (tuhat dollarit);
x_3 — kulud teiste magusatoodete reklaamile aastas (tuhat dollarit).

Anda tõlgendus mudeli parameetritele. Kui suur on köhakommide müügitulu, kui müügimaht aastas on 25 tuhat tonni, kulud köhakommide reklaamile 20 tuhat dollarit aastas ja kulud teiste maiustuste reklaamile 300 tuhat dollarit aastas? VASTUS lk 681.

9.5. 1990. aasta suvel viis ajakiri Journal of Personal Selling & Sales Management läbi uuringu tööstustoodangu turustajate hulgas. Küsitleti 244 meest ja 153 naist 16-st USA kaguosa tööstusettevõttest. Muu hulgas uuriti, millest sõltub tööga rahulolu. Kasutati järgmist mudelit:

$$y = b + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + \varepsilon,$$

kus
y — tööga rahulolu (mõõdetud 40-pallises skaalas);
x_1 — vanus (aastates);
x_2 — haridus (aastates);
x_3 — töökogemus (kuudes);
x_4 — müügikogemus (kuudes).
Tõlgendada mudeli parameetreid a_1, a_2, a_3 ja a_4. VASTUS lk 681.

9.6. Turunduses on oluline määrata, kes on toote potentsiaalsed ostjad, neid nimetatakse sihtrühmaks. Sihtrühmi võib eristada inimeste vanuse, soo, sissetuleku ja muude näitajate põhjal. Ajakirjas Journal

Mittelineaarne mudel

of Advertising Research 1980. aastal ilmunud artiklis uuriti erinevate mittealkohoolsete jookide tarbimist 21-aastaste ja vanemate inimeste hulgas (Wheatley, Chiu ja Stevens, 1980). Selleks viidi Seattle'i piirkonnas (USA) läbi telefoniküsitlus, millele vastas 252 isikut. Kogutud andmete põhjal koostati mudel, mis kirjeldas kohvi tarbimise sõltuvust inimese vanusest:

$$\hat{C} = -13{,}195 + 1{,}288A - 0{,}011A^2, \quad R^2 = 0{,}094,$$

kus C on kohvi tarbimine ööpäevas untsides (1 unts $\approx$ 28,3 g) ja A vanus aastates. Kohvi tarbimine tähendas valmis tehtud joogi tarbimist, mitte kohviubade või jahvatatud kohvi hulka.

1. Mitu grammi kohvi ööpäevas joob selle mudeli järgi 30-aastane inimene?
2. Kas 30-aastase isiku kohvi tarbimine vanuse kasvades suureneb või väheneb?
3. Kas 65-aastase isiku kohvi tarbimine vanuse kasvades suureneb või väheneb?
4. Kui vanad inimesed joovad ööpäevas kohvi rohkem kui 400 grammi?
5. Millises vanuses isikud joovad kohvi kõige rohkem?

VASTUS lk 681.

9.7. Gini kordaja on ühiskonna tulude jaotuse ebavõrdsuse näitaja, mille võttis kasutusele Itaalia statistik Corrado Gini (1884–1965). Gini kordaja jääb vahemikku 0–1 ja mida suurem see on, seda suurem on ebavõrdsus tulude jaotuses. Gini indeks on sama näitaja protsentides (korrutatud sajaga).

2016. aastal TTÜ-s kaitstud bakalaureusetöös uuriti tulude ebavõrdsuse mõju majanduskasvule (Kondjukova, 2016). Töö autor kasutas 28 Euroopa Liidu riigi andmeid aastatest 2004–2014. Gini indeks varieerus vahemikus 22,7–38,9. Mudelis võeti sõltuvaks tunnuseks SKP kasvumäär (protsentides) ning seletavateks tunnusteks Gini indeks (GIN), töötuse määr (UNE, protsentides), ekspordi osatähtsus SKP-s (EKS, protsentides), sündivus (SYN) ja oodatav eluiga aastates (ELU). Mudeliks saadi

$$\begin{aligned}\mathrm{SKP} = {} & 47{,}6 + 2{,}45\,\mathrm{GIN} - 0{,}0386\,\mathrm{GIN}^2 - 0{,}325\,\mathrm{UNE} + \\ & + 0{,}0939\,\mathrm{EKS} - 7{,}32\,\mathrm{SYN} - 0{,}947\,\mathrm{ELU} + \sum_{i=2}^{11} D_i dt_i + \varepsilon.\end{aligned}$$

Fiktiivsed tunnused dt_i vastasid aastatele ning nende kordajate D_i väärtusi pole siin toodud.

1. Kuidas mõjutab majanduskasvu töötuse määr?
2. Kuidas mõjutab majanduskasvu ekspordi osatähtsus SKP-s?

3. Kas suur ebavõrdsus mõjutab majanduskasvu negatiivselt või positiivselt?
4. Kui väike peab Gini indeks olema, et selle suurenemine mõjutaks majanduskasvu positiivselt?

VASTUS lk 681.

Korrigeeritud determinatsioonikordaja

9.8. Valimisse kuulus 20 objekti ja iga objekti jaoks registreeriti neli tunnust: Y, X_1, X_2 ja X_3. Eesmärgiks oli leida lineaarne regressioonmudel funktsioontunnuse Y modelleerimiseks. Kasutatud mudelid olid järgmised

$$\begin{aligned} &y = b + a_1x_1 + \varepsilon, && R^2 = 0{,}45,\\ &y = b + a_1x_1 + a_2x_2 + \varepsilon, && R^2 = 0{,}53,\\ &y = b + a_1x_1 + a_2x_2 + a_3x_3 + \varepsilon, && R^2 = 0{,}55. \end{aligned}$$

Milline neist kolmest mudelist on parim? VASTUS lk 681.

9.9. Lineaarse regressioonmudeli hindamisel saadi determinatsioonikordaja väärtuseks 0,775 ja korrigeeritud determinatsioonikordaja tuli 0,739. Mitu argumenttunnust oli mudelis, kui valimi maht oli 30? VASTUS lk 681.

9.10. Tihti esitatakse korrigeeritud determinatsioonikordaja valem kujul

$$R_a^2 = R^2 - (1 - R^2)\frac{k}{n-k-1},$$

kus R^2 on determinatsioonikordaja, n valimi maht ja k argumenttunnuste arv. Näidata, et see on ekvivalentne valemiga (9.58).

Parameetrite statistiline olulisus

9.11. On teada, et veini hind sõltub selle vanusest: mida vanema aastakäiguga vein, seda kallim. Ashenfelter jt analüüsisid oma 1995. aastal ilmunud artiklis „Bordeaux Wine Vintage Quality and the Weather“ (Ashenfelter, Ashmore ja Lalonde, 1995), kas veini hind sõltub ka kliimatingimustest, mis valitsesid vastava viinamarjasaagi valmimise eel ja ajal. Analüüsiks kasutasid nad andmeid Bordeaux' veinide kohta aastatest 1952–1980, välja arvatud aastakäigud 1954 ja 1956, mida müüakse väga harva. Hinnad võeti Londoni veiniturult aastatel 1990–1991. Regressioonanalüüs viidi läbi kolme erineva mudeli korral, kus sõltuvaks tunnuseks võeti hinna y naturaallogaritm: sellisel juhul näitab lineaarse mudeli kordaja hinna muutust protsentides.

$$\begin{aligned} &\text{Mudel 1:} && \ln y = b + a_1x_1 + \varepsilon;\\ &\text{Mudel 2:} && \ln y = b + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + \varepsilon;\\ &\text{Mudel 3:} && \ln y = b + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5 + \varepsilon, \end{aligned}$$

kus
x_1 — veini vanus aastates;

x_2 — keskmine õhutemperatuur (°C) viinamarjasaagi kasvuperioodil (aprill – september);
x_3 — sademete hulk (mm) augustis ja septembris;
x_4 — sademete hulk (mm) enne viinamarja kasvuperioodi (oktoober – märts);
x_5 — keskmine õhutemperatuur (°C) septembris.

Tabelis 9.19 on esitatud parameetrite hinnangud ja sulgudes nende standardvead, välja ei ole toodud vabaliiget (Ashenfelter, Ashmore ja Lalonde, 1995, tabel 2).

Tabel 9.19. Ülesande 9.11 juurde

Mudelis olev tunnus	Mudel 1	Mudel 2	Mudel 3
Veini vanus (x_1)	0,0354	0,0238	0,0240
	(0,0137)	(0,00717)	(0,00747)
Temperatuur apr–sept (x_2)		0,616	0,608
		(0,0952)	(0,116)
Sademed aug ja sept (x_3)		−0,00386	−0,00380
		(0,00081)	(0,000950)
Sademed okt–märts (x_4)		0,001173	0,00115
		(0,000482)	(0,000505)
Temperatuur sept (x_5)			0,00765
			(0,0565)
R^2	0,212	0,828	0,828
Valimi maht n	27	27	27

1. Kontrollida parameetrite statistilist olulisust kõikide mudelite korral, kasutades olulisuse nivood 0,05.
2. Kas on tõestatud, et kliimatingimused enne viinamarjade kasvuperioodi ja kasvuperioodil mõjutavad veini hinda?
3. Kas on tõestatud, et septembrikuu temperatuur eraldi mõjutab veini hinda?
4. Artikli autorid kasutasid üht mudelit veinide hindade prognoosimiseks. Milline mudel sobib prognoosimiseks kõige paremini?

VASTUS lk 681.

Ühikute teisendamine

9.12. Eesti piimandussektori modelleerimisel aastatel 2004–2008 saadi järgmine piima kokkuostuhinna mudel (Põldaru, Roots ja Viira, 2009):

$$\hat{y} = 0{,}221 + 0{,}111x_1 - 0{,}146x_2 + 1{,}655x_3,$$

kus on kasutusel järgmised tunnused:
y — kokkuostetava piima hind Eesti siseturul (kr/kg),
x_1 — juustu hind EL-i siseturul (kr/kg),
x_2 — või hind EL-i siseturul (kr/kg),

x_3 — odra hind Eesti siseturul (kr/kg).

1. Teisendada hinnad eurodesse kursiga 1 € = 15,65 kr ja panna kirja teisendatud mudel.
2. Tõlgendada teisendatud mudeli parameetreid.

VASTUS lk 682.

ÜL09Regressioon

Järgmiste ülesannete andmed on failis ÜL09Regressioon

Lineaarne regressioon

A.9.1. Tabelis on esitatud SKP elaniku kohta ja keskmine brutokuupalk Eesti maakondades 2012. aastal[8]. Mõlemad suurused on eurodes. Leida lineaarne mudel, mis kirjeldab keskmise brutokuupalga sõltuvust elaniku kohta tulevast SKP-st. Kui suur osa keskmise brutokuupalga varieerumisest erinevates maakondades on tingitud sellest, et regionaalne SKP elaniku kohta on erinev? VASTUS lk 682.

A.9.2. Tarbimismudelis on sõltuvaks tunnuseks kulud mingile hüvisele ning seletavaks tunnuseks tarbimiskulud kokku. Tabelis on toodud 2012. aasta Eesti leibkonnauuringus osalenud inimeste hulgast juhuslikult valitud 500 isiku andmed: kulud riietele ja jalanõudele, kulud transpordile, kulud vabale ajale ning tarbimiskulud kokku (*Leibkonna eelarve uuring* 2012). Kõik kulud on eurodes leibkonnaliikme kohta aastas.

1. Leida riiete, transpordi ja vaba aja tarbimismudelid.
2. Tõlgendada saadud mudeleid. Millise hüvise tarbimine kasvab kogukulude kasvades kõige kiiremini ja millise hüvise tarbimine kõige aeglasemalt?
3. Millise mudeli kirjeldusvõime on kõige kõrgem?
4. Mudelite võrdlemiseks konstrueerida diagramm, kus on mudelite põhjal leitud kolm sirget.
5. Kulud mingile hüvisele saavad olla vaid positiivsed. Milliste kogukulude väärtuste korral tekivad kulud riietele, transpordile ja vabale ajale?

VASTUS lk 682.

A.9.3. Kuidas kasutatud sõiduauto hind sõltub selle vanusest? Tabelis on portaalis Auto24 2014. aasta aprillis müügil olnud 12 kasutatud sõiduauto vanus ja hind[9]. Kõikide autode mark on Mazda, käigukast manuaal ja kütus bensiin.

1. Leida lineaarne mudel, mis kirjeldab auto hinna sõltuvust auto vanusest.

[8] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel RAA0050: sisemajanduse koguprodukt (ESA 2010), tabel PA5321: keskmine bruto- ja netokuupalk.

[9] Auto24 `http://www.auto24.ee`

2. Kuidas muutub kasutatud Mazda hind vanuse kasvades?
3. Kui palju maksaks leitud mudeli järgi uus Mazda?
4. Kui palju maksaks leitud mudeli järgi Mazda, mille vanus on kuus aastat?
5. Võrrelda müügil olnud autode tegelikku hinda mudeli järgi arvutatud hinnaga.

VASTUS lk 683.

A.9.4. Kasutades autotootja Audi AG 2014. aasta aruandest (Audi, 2014) võetud andmeid, leida

a) kulufunktsioon (kulude sõltuvus toodetud autode arvust);
b) püsikulu usaldusvahemik usaldatavusega 95%;
c) piirkulu usaldusvahemik usaldatavusega 95%;
d) võrrelda Audi püsikulu ja piirkulu näites 9.4 leitud BMW vastavate kuludega.

VASTUS lk 683.

A.9.5. Modifitseeritud Phillipsi kõver on seos töötuse määra U_t ja inflatsioonimäära muutuse $\Delta\pi_t$ vahel (Kerem jt, 1988, lk 167):

$$\Delta\pi_t = -\alpha(U_t - U_{nom}), \tag{9.121}$$

kus inflatsioonimäära muutus $\Delta\pi_t = \pi_t - \pi_{t-1}$, U_t on tegelik ning U_{nom} nominaalne töötuse määr. Konstant α on positiivne, mis tähendab, et kui tegelik töötuse määr on suurem kui nominaalne, siis inflatsioon väheneb ($\Delta\pi_t < 0$).

Regressioonanalüüsi läbiviimiseks kirjutatakse mudel (9.121) kujul

$$\Delta\pi_t = b + aU_t, \tag{9.122}$$

kus

$$a = -\alpha, \quad b = \alpha U_{nom}. \tag{9.123}$$

Teades inflatsioonimäära muutust, mida väljendab tarbijahinnaindeksi muutus, ning töötuse määra, on võimalik hinnata mudeli (9.122) parameetreid ning seejärel valemitest (9.123) arvutada nominaalne töötuse määr.

Tabelis on tarbijahinnaineksi muutus (THIMUUT, %) ning töötuse määr (TÖÖTUS, %) Eestis aastatel 1999–2009[10].

1. Hinnata mudelit (9.122), kus $\Delta\pi_t$ on THIMUUT ning U_t on TÖÖTUS.
2. Kasutades leitud parameetrite hinnanguid, arvutada valemitest (9.123) nominaalne töötuse määr Eestis perioodil 1999–2009.

[10] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel TT35: töötuse määr soo ja vanuserühma järgi, tabel IA01: tarbijahinnaindeksi muutus võrreldes eelmise aastaga (1992–2009).

VASTUS lk 683.

Prognoosi usalduspiirid

A.9.6. 1994. aastal otsustas ühe USA suurlinna ajaleht alustada ilmumist lisaks argipäevadele ka pühapäeviti. Et hinnata, kui suur võiks olla pühapäevase numbri tiraaž, koguti andmeid 34 ajalehe argipäevaste ja pühapäevaste tiraažide kohta (Chatterjee ja Hadi, 2006). Tiraažid on tuhandetes.

1. Konstrueerida hajumisdiagramm, kus horisontaalteljel on tiraaž argipäeviti ning püstteljel pühapäevane tiraaž. Kas diagrammi põhjal võib soovitada lineaarse seose olemasolu?
2. Hinnata lineaarset regressioonmudelit, kus sõltuvaks tunnuseks on pühapäevane tiraaž. Lisada regressioonjoon ka hajumisdiagrammile.
3. Kui suured on vabaliikme ning lineaarliikme kordaja usalduspiirid usaldatavusega 95%?
4. Kui suur osa pühapäevaste tiraažide varieeruvusest on seletatud argipäevaste tiraažide varieerumisega?
5. Kui suur on keskmine pühapäevane tiraaž ajalehtedel, mille argipäevane tiraaž on 500 tuhat? Leida usalduspiirid usaldatavusega 95%.
6. Ajalehel, mis soovib hakata pühapäeviti ilmuma, on argipäevane tiraaž 500 tuhat. Prognoosida selle ajalehe pühapäevast tiraaži ning leida prognoosi usaldusvahemik usaldatavusega 95%. Võrrelda usaldusvahemiku poollaiust eelmises osas leitud usaldusvahemiku poollaiusega.
7. Üks teine ajaleht, mille argipäevane tiraaž on 2 miljonit, soovib samuti hakata pühapäeviti ilmuma. Prognoosida selle ajalehe pühapäevast tiraaži ning leida prognoosi usaldusvahemik usaldatavusega 95%. Võrrelda usaldusvahemiku poollaiust eelmises osas leitud usaldusvahemiku poollaiusega.

VASTUS lk 683.

Mittelineaarne regressioon

A.9.7. XX sajandi algul pandi tööstusettevõtetes tähele, et kehtib järgmine empiiriline seaduspärasus: toodete arvu kahekordistumisel väheneb ühe toote jaoks kulutatud aeg kindla protsendi võrra (Dutton ja Thomas, 1984; Wright, 1936). Kõverat, mis kirjeldab tooteühiku jaoks kulutatud aja sõltuvust toodete arvust, nimetatakse õppimiskõveraks (*learning curve*). Enamasti esitatakse see astmefunktsioonina

$$t_n = t_1 n^b, \tag{9.124}$$

kus t_1 on esimese toote valmistamiseks kulunud aeg, n toodete arv kokku ja t_n n-nda toote valmistamiseks kulunud aeg. Kui toodete arvu kahekordistumisel on ühiku tootmiseks vajamineva aja vähenemise määr r, siis

$$b = \log_2(1 - r), \tag{9.125}$$

(vt lisa A.14). Erinevate tehnoloogiate korral võib vähenemise määr olla vahemikus 5%–40% (Dutton ja Thomas, 1984). Kuna tööaeg on otseselt seotud tööjõukuluga, siis õppimiskõver väljendab ühiku tööjõukulu vähenemist tootmismahu suurenemisel.

Tabelis on esitatud andmed 384 Liberty klassi kaubalaeva kohta, mis ehitati laevatehases Bethlehem-Fairfield Shipyards ajavahemikul aprill 1941 kuni oktoober 1944 (P. Thompson, 2007). Iga dokist väljunud laeva jaoks on antud selle valmimise aeg ning ehitamiseks kulunud töötundide arv (mln). Laevade koguarv näitab, mitu laeva on valmistatud alates 1941. aasta detsembrist, mil valmis esimene laev.

1. Luua hajumisdiagramm, kus horisontaalteljel on laevade koguarv ning vertikaalteljel töötundide arv laeva kohta. Veenduda, et seos on mittelineaarne.
2. Hajumisdiagrammile lisada astmefunktsioonile (9.124) vastav regressioonjoon koos mudeli ja determinatsioonikordajaga.
3. Kasutades parameetri b hinnangut, leida valemist (9.125) töötundide arvu vähenemise määr r laevade arvu kahekordistumisel.

Vastus lk 683.

A.9.8. Transpordiettevõte teostab ühe kliendi toodangu laialivedu Soomes asuvale kauplustevõrgule. Reisid on alati erinevad ja sõltuvad konkreetsetest tellimustest, sihtkohtadeks võib olla kuni 18 erinevat kauplust üle kogu Soome. Tabelis on toodud erineva pikkusega reiside veohind. Andmed pärinevad aastast 2005, kroonid on teisendatud eurodeks.

1. Konstrueerida hajumisdiagramm, mis kirjeldab veohinna sõltuvust läbisõidust. Veenduda, et seos on mittelineaarne: algul suureneb veohind läbisõidu kasvades kiirenevalt, hiljem aeglustuvalt. Käänupunkt on 1500 km läbisõidu korral.
2. Konstrueerida diagramm, mis kirjeldab veohinna sõltuvust läbisõidust, kui läbisõit on kuni 1500 km. Lisada sobiva kujuga regressioonjoon. Milline kõver kirjeldab seda seost kõige paremini?
3. Konstrueerida diagramm, mis kirjeldab veohinna sõltuvust läbisõidust, kui läbisõit on üle 1500 km. Lisada sobiva kujuga regressioonjoon. Milline kõver kirjeldab seda seost kõige paremini?
4. Kui kiiresti kasvab veohind, kui läbisõit on
   a) 500 km,
   b) 1400 km,
   c) 2100 km?

Vastus lk 683.

A.9.9. Näites 9.7 analüüsisime, kuidas kalade söödakoefitsient sõltub veetemperatuurist. Lisaks söödakulule uurisid viidatud artikli autorid ka kalade kasvumäära sõltuvust veetemperatuurist. Tabelis on toodud

70–150-grammiste lõhede kasvumäär (%) nelja erineva veetemperatuuri juures (Handeland, Imsland ja Stefansson, 2008).

1. Konstrueerida hajumisdiagramm, kus on esitatud kasvumäära sõltuvus temperatuurist. Veenduda, et tegemist ei ole lineaarse seosega, vaid seose modelleerimiseks sobib ruutparabool.
2. Lisada hajumisdiagrammile ruutfunktsioonile vastav regressioonjoon koos võrrandi ja determinatsioonikordajaga.
3. Mudelit kasutades leida, millise temperatuuri juures on kasvumäär kõige kõrgem ning kui suur on maksimaalne kasvumäär.

VASTUS lk 683.

Mitmene regressioon

A.9.10. Seda, kuidas reklaam mõjutab käivet, analüüsiti ajakirjas Journal of Advertising Research 1981. aastal ilmunud artiklis (McDaniel, 1981). Tabelis on toodud ühe ettevõtte käibe suurenemine nädalas (tuhat dollarit), linnade arv, kus näidati selle ettevõtte reklaami telekanalis, ning reklaamide arv nädalas.

1. Kas käibe suurenemine sõltub sellest, mitmes linnas reklaami näidati ja kui tihti?
2. Kui suur on linnade arvu olulisuse tõenäosus?
3. Kui suur on reklaamide sageduse olulisuse tõenäosus?

VASTUS lk 686.

A.9.11. Ülesandes A.9.3 tuli leida kasutatud Mazda hinna sõltuvus auto vanusest. Nüüd on lisaks vanusele antud ka auto läbisõit kilomeetrites. Kontrollida, kas läbisõidu lisamine hinna mudelisse parandab mudelit. VASTUS lk 686.

A.9.12. Kellade kollektsionäär teab, et kella hind kasvab lineaarselt kella vanuse kasvades. Lisaks sellele püstitab ta hüpoteesi, et kella hind oksjonil sõltub lineaarselt ka oksjonil osalejate arvust. Vastavalt sellele peaks kella hinda saama modelleerida järgmise lineaarse mudeliga:

$$y = b + a_1 x_1 + a_2 x_2 + \varepsilon,$$

kus y on kella hind oksjonil (dollarit), x_1 kella vanus aastates, x_2 oksjonil osalejate arv ja ε juhuslik komponent. Oma hüpoteesi kontrollimiseks on kollektsionäär kogunud andmeid 32 oksjonil müüdud kella kohta: kella vanus aastates, oksjonil osalejate arv ja kella müügihind oksjonil.

Viia läbi vastav regressioonanalüüs ja kontrollida kollektsionääri hüpoteesi, et kella hind sõltub kella vanusest ja oksjonil osalejate arvust. VASTUS lk 686.

A.9.13. Ajakirjas The Review of Economics and Statistics analüüsiti, millest sõltus uute autode müük USA-s aastatel 1929–1956 (Suits, 1958). Kasutati järgmisi tunnuseid:

MÜÜK — müüdud uute autode arv aastas, miljonit;
HIND — autode jaemüügi hinnaindeks;
TULU — elanike kogusissetulek aastas, miljard dollarit;
ARV — registreeritud autode koguarv aasta algul, miljonit.

1. Hinnata mudelit, kus müüdud autode arv sõltub hinnaindeksist. Kuidas hinnatõus mõjutab müüdud autode arvu?
2. Lisada mudelisse ka elanike kogusissetulek ja registreeritud autode arv. Kuidas nüüd hinnatõus mõjutab müüdud autode arvu? Kuidas mõjutab müüdud autode arvu autode koguarv?

VASTUS lk 687.

Ebatüüpilised vaatlused ja erindid

A.9.14. 2013. aastal olid kohalike eelarvete tulud kokku 1477 tuhat eurot ning sellest 725,8 tuhat eurot moodustas füüsilise isiku tulumaks[11]. Järelikult tuleb ligikaudu 50% kohalike omavalitsuste tuludest füüsilise isiku tulumaksust. Tulumaksu suurus ühe elaniku kohta on erinevates omavalitsustes erinev ja sõltub mitmetest näitajatest.

Tabelis on toodud andmed 215 Eesti omavalituse kohta[12]:

TM — füüsilise isiku tulumaks elaniku kohta 2013. aastal (eurot);
RM — rahvaarvu muutus 2013–2014 (%);
RT — registreeritud töötus (%);
ÜM — ülalpeetavate määr (%).

Rahvaarvu muutus iseloomustab konkreetse territooriumi atraktiivsust ehk seda, kas inimesed soovivad selle omavalitsuse territooriumil elada. Ülalpeetavate määr näitab, milline on laste (0–14 aastat) ja pensionäride (65 ja vanemad) suhe tööealisse elanikkonda (15–64 aastat) ning iseloomustab tööjõu potentsiaali ja taastootmisvõimet. Registreeritud töötute osatähtsus iseloomustab omavalituse majanduslikku aktiivsust. (Servinski, Kivilaid ja Tischler, 2009)

1. Hinnata lineaarset mudelit, kus sõltuvaks tunnuseks on füüsilise isiku tulumaks elaniku kohta ja seletavateks tunnusteks ülejäänud kolm tunnust:

$$\mathrm{TM} = b + a_1\mathrm{RM} + a_2\mathrm{RT} + a_3\ddot{\mathrm{U}}\mathrm{M} + \varepsilon.$$

2. Millised tunnused on statistiliselt olulised nivool 0,01 ja millised nivool 0,05?
3. Kui suur on mudeli kirjeldusvõime?
4. Tõlgendada mudeli parameetreid.
5. Millised on ebatüüpilised omavalitsused ja millised on erindid?

VASTUS lk 687.

Tunnuste valik

A.9.15. Millised kinnisvara iseloomustavad tunnused on olulised kin-

[11] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel RR31: kohalike eelarvete tulud.

[12] Allikas: Eesti Statistikaameti piirkondliku statistika portaal `http://www.stat.ee/pp`

nisvara hinna kujunemisel? Tabelis on andmed 25 Pennsylvania osariigis müüdud eramu kohta (Narula ja Wellington, 1977):

HIND — maja müügihind, tuhat dollarit;
KRUNT — krundi pindala, tuhat ruutjalga ($1\,\mathrm{ft}^2 \approx 0{,}093\,\mathrm{m}^2$);
VANNIT — vannitubade arv;
GARAAZ — garaažikohtade arv;
KAMIN — kaminate arv;
VANUS — maja vanus aastates;
MAKS — kinnisvaramaks aastas, sada dollarit.

1. Leida mudel, mis võimaldab prognoosida maja müügihinda, kui on teada kinnisvara iseloomustavad tunnused. Nende hulka ei kuulu kinnisvaramaks. Sobiva mudeli leidmiseks kasutada edaspidist sammsammulist tunnuste lisamise meetodit, alustades ühe tunnusega mudelist. Tunnuste järjestamiseks kasutada korrelatsioonikordajaid sõltuva tunnusega ning mudelisse jätta tunnused, mille statistiline olulisus on tõestatud nivool 0,05.
2. Anda tõlgendus mudeli parameetritele.
3. Maakler väidab, et kuna kinnisvaramaksu määramisel võetakse arvesse kinnisvara iseloomustavad tunnused, piisab hinna prognoosimisel mudelist, kus sõltuvaks tunnuseks on ainult kinnisvaramaks. Hinnata seda mudelit ja võrrelda punktis 1 saadud mudeliga. Kumb mudel on parem?

Vastus lk 687.

A.9.16. Paljud autorid on analüüsinud, millised tegurid on seotud majanduskasvuga. W.H. Masanjala ja C. Papageorgiu uurisid 37 Aafrika riigi andmete põhjal, millest sõltub majanduskasv Aafrikas (Masanjala ja Papageorgiou, 2008). Nad kasutasid 24 potentsiaalset seletavat tunnust, tabelis on toodud nendest 9. Sõltuvaks tunnuseks on SKP suhteline kasv.

Growth — keskmine SKP suhteline kasv aastas perioodil 1960–1992;
GDP60 — SKP elaniku kohta aastal 1960, logaritmitud;
Mining — kaevandamise osatähtsus SKP-s;
PrimExp70 — primaarkaupade ekspordi suhe SKP-sse aastal 1970;
PrimSch60 — algkoolis osalemise määr aastal 1960;
LifExp60 — oodatav eluiga sünnihetkel aastal 1960;
RevCoup — keskmine revolutsioonide ja riigipöörete arv aastas (1960–1984);
Invest — kodumaiste investeeringute suhe SKP-sse;
YrsOpen — avatud majanduse aastate osakaal perioodil 1965–1990;
Muslim — moslemite osakaal elanikkonnas.

1. Sobiva mudeli saamiseks kasutada sammsammulist tunnuste eemaldamist (tagurpidine valik). Tunnuste statistilise olulisuse testimisel võtta olulisuse nivooks 0,1.
2. Millised tegurid mõjuvad majanduskasvule positiivselt, millised negatiivselt?

VASTUS lk 688.

A.9.17. Kas piima tarbimine sõltub reklaamist? Seda analüüsis H. Kinnucan oma artiklis, mis ilmus 1986. aastal ajakirjas Journal of Agricultural and Resource Economics (Kinnucan, 1986). Ta tugines New Yorgi andmetele aastatest 1971–1979:

piima tarbimine, gallonit ($\approx$ 3,8 liitrit) elaniku kohta aastas;
hind, senti kvardi ($\approx$ 1,14 liitrit) kohta;
reklaamikulud, senti elaniku kohta aastas;
sissetulek, dollarit elaniku kohta aastas.

Leida parim mudel, mis kirjeldab piima tarbimist, ja anda tõlgendus mudeli parameetritele. Kui suure osa piima tarbimise varieerumisest mudel ära seletab? VASTUS lk 688.

Multikollineaarsus

A.9.18. Transporditeenuse pakkujate jaoks on oluline kaubavedude mahu prognoosimine. Üks võimalus prognoosimiseks on sobivate seletavate tunnustega regressioonmudeli kasutamine. Kui on olemas vastavate seletavate tunnuste prognoosid, siis saab regressioonmudeli põhjal prognoosida ka sõltuvat tunnust.

Ajakirjas Neurocomputing 2015. aastal ilmunud artiklis analüüsiti kaubavedude mahu muutust Šanghais aastatel 2000–2010 ning selle seost Šanghai piirkonna SKP ja sinna tehtud põhivara investeeringutega (Y. Yang, 2015). Olgu

y kaubavedude maht (miljonit tonni);
x_1 SKP (miljardit jüaani);
x_2 põhivara investeeringud (miljardit jüaani).

Kasutades artikli autori poolt toodud andmeid, leida sobiv regressioonmudel kaubavedude mahu prognoosimiseks. Valik teha järgmiste mudelite hulgast:

$$\hat{y} = ax_1 + b, \tag{9.126}$$

$$\hat{y} = ax_2 + b, \tag{9.127}$$

$$\hat{y} = ax_1 + ax_2 + b. \tag{9.128}$$

Viimase mudeli korral kontrollida multikollineaarsuse esinemise võimalust. VASTUS lk 688.

A.9.19. Tänapäeval on hoonete projekteerimisel väga oluline võimalikult väike energiakulu. Tabelis on andmed 19 nelja ja viie tärni hotelli kohta Hiina Hainani provintsis (Xin jt, 2012):

KWH — energiakulu (1000 kWh);
PINDALA — hotelli pindala (m^2);
TOAD — tubade arv;
TÄITUVUS — keskmine täituvus (%);
VANUS — hotelli vanus aastates.

1. Leida mudel, mis kirjeldab hotelli energiakulu sõltuvust hotelli iseloomustavatest teguritest. Sobiva mudeli leidmiseks kasutada edaspidist sammsammulist tunnuste lisamise meetodit, alustades ühe tunnusega mudelist. Tunnuste järjestamiseks võtta aluseks korrelatsioonikordajad sõltuva tunnusega ning mudelisse jätta tunnused, mis on olulised nivool 0,1.
2. Põhjendada, miks tunnused PINDALA ja TOAD ei sobi mõlemad korraga mudelisse.
3. Anda tõlgendus mudeli parameetritele.
4. Kui suure osa hotellide energiakulu varieerumisest mudel ära seletab?

VASTUS lk 688.

Regressioon läbi nullpunkti

A.9.20. Ajakirjas Journal of Advertising Research 1981. aastal ilmunud artiklis analüüsiti toidukaupade müügikäivet USA erinevates linnades (McDaniel, 1981). Andmed pärinevad aastast 1975. Rahvaarv on miljonites ning toidukaupade käive miljardit dollarit aastas.

1. Leida lineaarne regressioonmudel, mis kirjeldab toidukaupade müügikäibe sõltuvust linna rahvaarvust. Veenduda, et vabaliige on statistiliselt mitteoluline.
2. Loogiline on eeldada, et linnas, milles elanikke pole, ei ole ka vajadust toidukaupade järele. Seepärast viia läbi nullpunkti läbiva regressioonjoone hindamine. Kuidas sõltub toidukaupade käive rahvaarvust?

VASTUS lk 689.

A.9.21. Toidus sisalduva energia allikaks on rasvad, süsivesikud ja valgud. R. Johnson (1995) kasutas 13 toiduaine andmeid analüüsimaks, kuidas nende komponentide sisaldus mõjutab toidu energiasisaldust. Ta sai mudeliks

$$\widehat{\mathrm{CAL}} = 8{,}888\,\mathrm{RASV} + 4{,}266\,\mathrm{VALK} + 3{,}978\,\mathrm{SV}, \tag{9.129}$$

kus CAL on energiasisaldus (kcal), RASV rasvade sisaldus (g), VALK valkude sisaldus (g) ning SV süsivesikute sisaldus (g) 100 grammis toidus. Kui toitaines puuduvad rasvad, süsivesikud ja valgud, siis ei saa toidust ka energiat. Seepärast kasutas Johnson regressiooni läbi nullpunkti.

Tabelis on 35 toiduaine energiasisaldus (kcal) ning rasvade, valkude ja süsivesikute sisaldus 100 grammis toidus. Kasutatud on Tervise

Arengu Instituudi toidu koostise andmebaasi Nutridata (Pitsi, Kambek ja Jõeleht, 2014).

1. Viia läbi vabaliiget sisaldava lineaarse mudeli hindamine, kus sõltuvaks tunnuseks on energiasisaldus ning seletavateks tunnusteks rasvade, valkude ja süsivesikute sisaldus. Veenduda, et vabaliige on statistiliselt mitteoluline.
2. Viia läbi ilma vabaliikmeta mudeli hindamine.
3. Kas Johnsoni mudeli (9.129) kordajad langevad eelmises punktis hinnatud mudeli kordajate 95%-lise usaldatavusega leitud usalduspiiridesse?
4. Milline komponent annab kõige rohkem energiat ja milline kõige vähem?
5. Nutridata andmebaasi järgi on 100 grammis McDonaldsi hamburgeris 8,5 grammi rasva, 30,2 g süsivesikuid ning 12,3 g valke. Kui palju energiat annab 100 g hamburgeri söömine?

VASTUS lk 689.

Mittelineaarse mudeli lineariseerimine

A.9.22. Ülesandes A.9.2 tuli lineaarsete mudelite abil hinnata, kuidas kulutused riietele, jalanõudele, transpordile ning vabale ajale sõltuvad summaarsetest tarbimiskuludest. Kuid on tähele pandud, et summaarsete tarbimiskulude suurenemisel kulutused mitmetele hüvistele ei muutu konstantselt, vaid mudelis esineb mittelineaarsus. Kasutades ülesandes A.9.2 toodud andmeid, kontrollida, kas mittelineaarsus esineb ka nende hüviste kulude suurenemisel. Selleks lisada lineaarsetesse mudelitesse „Kulud kokku“ ruutliige, s.t hinnata mudeleid kujul

$$y_{\text{hüvis}} = a_1 x + a_2 x^2 + b + \varepsilon,$$

kus x on kulud kokku ning $y_{\text{hüvis}}$ kulud riietele, transpordile või vabale ajale. Kas ruutliige on kõigi kolme hüvise korral statistiliselt oluline ja kas selle lisamine muudab mudeli paremaks? Kas kogukulude kasvades kulud konkreetsele hüvisele kasvad kiirenevalt või aeglustuvalt? Märkus: ruutliikme lisamiseks tuleb tabelarvutuses eelnevalt leida uus tunnus „Kulud kokku“ ruudus ning seejärel viia läbi regressioonanalüüs nagu kahe tunnusega lineaarse mudeli korral. VASTUS lk 689.

A.9.23. Korruptsiooni tajumise indeks (CPI, *Corruption Perceptions Index*) on Transparency Internationali igal aastal leitav indeks, mille koostamisel lähtutakse rahvusvaheliste reitinguagentuuride ning uurimisasutuste kogutud andmetest korruptsiooni tajumise kohta avalikus sektoris. Indeksi arvutamisel kasutatakse Maailmapanga, Maailma Majandusfoorumi, Freedom House'i jt institutsioonide tehtud uuringuid. Andmed standardiseeritakse ning agregeeritakse 100 palli skaalale, nii et 100 palli saanud riik loetakse täiesti korruptsioonivabaks[13].

[13] Allikas: Korruptsioonivaba Eesti `http://www.transparency.ee`

Shao jt analüüsisid korruptsiooni tajumise indeksi ja riigi elatustaseme vahelist seost ning näitasid, et rikkamates riikides on korruptsiooni tajumise indeks suurem (korruptsiooni tajutakse vähem) (Shao jt, 2007). Riigi elatustaseme näitajaks oli neil sisemajanduse koguprodukt elaniku kohta (*GDP per capita*), mõõdetuna USA dollarites. Seose modelleerimiseks kasutasid nad astmefunktsiooni

$$CPI = aGDP^{\mu}, \tag{9.130}$$

kus CPI on korruptsiooni tajumise indeks 100-pallisel skaalal, GDP sisemajanduse koguprodukt elaniku kohta USA dollarites ning a ja μ positiivsed konstandid. Artikli autorid kasutasid andmeid aastatest 2000–2005.

Tabelis on toodud korruptsiooni tajumise indeks[14] ja SKP elaniku kohta[15] 169 riigis. Riigid on jagatud järgmistesse piirkondadesse:

AF — Lõuna-Aafrika;
AM — Ameerika;
AP — Aasia ja Vaikne ookean;
EE — Ida-Euroopa ja Kesk-Aasia;
EU — Euroopa Liit ja Lääne-Euroopa;
ME — Kesk-Ida ja Põhja-Aafrika.

Parameetrite a ja μ määramiseks tuleb seos (9.130) viia lineaarsele kujule, võttes mõlemalt poolt naturaallogaritmid:

$$\ln CPI = \ln a + \mu \ln GDP. \tag{9.131}$$

1. Arvutada kõikide riikide jaoks $\ln CPI$ ja $\ln GDP$.
2. Hinnata mudelit (9.131) ning leida ka standardiseeritud jäägid.
3. Kas mudel on statistiliselt oluline?
4. Kirjutada välja parameetri μ hinnang koos usalduspiiridega ning võrrelda seda Shao jt 2005. aasta andmete põhjal saadud tulemusega: $\mu \approx 0{,}26 \pm 0{,}02$.
5. Leida erinditena esilekerkivad riigid, kus korruptsiooni tajumise indeks on oluliselt madalam (korruptsiooni tajumine oluliselt suurem), kui elatustaseme põhjal võiks seda prognoosida.
6. Hinnata mudelit (9.131) ainult Euroopa Liidu ja Lääne-Euroopa riikide põhjal (EU).
7. Kas mudel on statistiliselt oluline?
8. Kirjutada välja parameetri μ hinnang koos usalduspiiridega.
9. Kasutades viimasena hinnatud mudelit, leida Eesti jaoks korruptsiooni tajumise indeksi mudelväärtus ning võrrelda seda empiirilise väärtusega.

[14] Allikas: Transparency International `http://www.transparency.org/`
[15] Allikas: International Monetary Fund `http://www.imf.org/`

VASTUS lk 689.

A.9.24. Daniel Belmont (1958) analüüsis, kuidas kahe linna vaheline lennureisijate arv sõltub lennureisijate koguarvust kummaski linnas ning linnadevahelisest kaugusest. Ta sai järgmise mudeli:

$$T_{ij} = \frac{k(T_i T_j)^p}{D^q}, \tag{9.132}$$

kus T_i ja T_j on siselendude lennureisijate koguarv linnades i ja j kindla ajavahemiku jooksul, T_{ij} lennureisijate arv nende kahe linna vahel ning D linnadevaheline kaugus. Mudeli kontrollimiseks kasutas ta 23 USA linna vahelise 41 lennuliini statistikat aastast 1955 ning leidis parameetrite k, p ja q hinnangud.

1. Lineariseerida mudel (9.132), selleks võtta mõlemalt poolt naturaallogaritmid.
2. Hinnata vastavat lineaarset mudelit.
3. Aasta varem ilmunud artiklis oli sama autor kasutanud mudelit, kus puudus linnadevaheline kaugus (Belmont, 1957). Kas kauguse mõju linnadevahelisele reisijatevoole on tõestatud?
4. Kasutades hinnatud mudeli parameetreid, kirjutada mudel kujul (9.132).

VASTUS lk 689.

A.9.25. 1866. aastal näitas Edward Jarvis, et mingist piirkonnast pärit patsientide arv vaimuhaiglas on pöördvõrdelises sõltuvuses selle piirkonna kaugusega haiglast (Jarvis, 1866). Selle seaduspäraga, mida nimetatakse ka Jarvise seaduseks (*Jarvis' law*), tuleb arvestada vaimuhaiglate planeerimisel. Neil on tavaliselt väiksem teeninduspiirkond kui muudel haiglatel. Jarvise seaduse kehtivust on hiljem kontrollinud mitmed autorid.

J. M. Hunter ja G. W. Shannon (1984) kasutasid USA Massachusettsi osariigis asuva Worcesteri vaimuhaigla andmeid aastatest 1832–1849. Tabelis on erinevatest maakondadest pärit patsientide arv (PTSARV), vastava maakonna rahvaarv (RHV) ning kaugus haiglast D miilides.

1. Hinnata mudelit
$$y = b + \frac{a}{D}, \tag{9.133}$$
kus sõltuv tunnus y on patsientide arv 1000 elaniku kohta.
2. Kas konstant b on statistiliselt oluline?
3. Loogiline on eeldada, et kui piirkonna kaugus D läheneb lõpmatusele ($1/D \to 0$), läheneb sellest piirkonnast pärit patsientide arv nullile. Seega ei peaks mudelis (9.133) konstantset liiget b olema. Hinnata mudelit ilma konstantse liikmeta.
4. Kas Jarvise seadus leidis kinnitust?

Vastus lk 689.

Kvalitatiivsed seletavad tunnused

A.9.26. Millest sõltub käimisel kulutatud energia? Ajakirjas The Journal of Physiology ilmunud artiklis mõõdeti 50 isiku poolt kulutatud kalorite hulk, kui nad jalutasid kiirusega 3 miili (1,6 km) tunnis (Mahadeva, Passmore ja Woolf, 1953). Valitud isikud olid erineva elustiiliga: üliõpilased, laboritehnikud, kooliõpilased ja pensionärid vanuses 13 kuni 79 aastat ja nende hulgas oli 35 meest ning 15 naist. Registreeriti katsealuste sugu, vanus, kehakaal (kg) ja jalutamisega kulutatud kilokalorite hulk kümnes minutis.

1. Viia läbi lineaarse regressioonmudeli hindamine, kus sõltuvaks tunnuseks on kulutatud energia ning seletavaks tunnuseks isiku kehakaal. Kas kehakaal mõjutab energiakulu?
2. Lisada mudelisse isiku vanus. Kas vanus mõjutab energiakulu?
3. Hinnata mudelit, kus seletavateks tunnusteks on kaal ja sugu (0 on naine ja 1 mees). Kas sugu mõjutab energiakulu?

Vastus lk 690.

A.9.27. Tabelis on 2002 kodulaenu andmed, mis väljastati USA-s 1990. aastal Bostonis (Munnell jt, 1996):

LAEN — laenu suurus (tuhat dollarit);
HIND — laenu eest muretsetava kinnisvara hind (tuhat dollarit);
HSUHE — hinna suhe laenutaotleja sissetulekusse;
OMAND — 0 isiklik omand, 1 kaasomand;
PEREK — 0 laenutaotleja ei ole abielus, 1 on abielus.

1. Leida keskmine laenu ja hinna suhe usaldatavusega 0,95.
2. Leida keskmine hinna ja sissetuleku suhe usaldatavusega 0,95.
3. Leida usaldatavusega 0,95, kui suur osa laenudest taotletakse kaasomandis oleva kinnisvara soetamiseks.
4. Hinnata lineaarset mudelit, kus sõltuvaks tunnuseks on laenu suurus ning seletavateks tunnusteks kõik ülejäänud tunnused.
5. Kas kõik tunnused on statistiliselt olulised nivool 0,01?
6. Kuidas laenu suurus sõltub
   a) kinnisvara hinnast;
   b) kinnisvara omandivormist;
   c) laenutaotleja perekonnaseisust?

Vastus lk 690.

A.9.28. Tuginedes 209 USA ettevõtte andmetele, mis on avaldatud ajakirjas Businessweek 6. mail 1991. aastal (Wooldridge, 2002), analüüsida, kuidas töötasu (TTASU, tuhat dollarit aastas) sõltub ettevõtte käibest (KÄIVE, miljon dollarit aastas), omakapitali rentaablusest ROE (*return on equity*) ja ettevõtte tegevusalast (ALA). Ettevõtted jagunevad nelja tegevusala vahel:

1 — tööstus,

2 — finantsvahendus,
3 — tarbekaupade tootmine,
4 — transport.

1. Hinnata järgmist mudelit:

$$\ln \text{TTASU} = b + a_1 \ln \text{KÄIVE} + a_2 \text{ROE} + \gamma_1 D_1 + \gamma_2 D_2 + \gamma_3 D_3 + \varepsilon,$$

kus ettevõtte tegevusala jaoks on loodud kolm fiktiivset tunnust:

$$D_1 = \begin{cases} 1, & \text{kui on finantsvahendus} \\ 0, & \text{kui ei ole finantsvahendus} \end{cases}$$
$$D_2 = \begin{cases} 1, & \text{kui on tarbekaubad} \\ 0, & \text{kui ei ole tarbekaubad} \end{cases}$$
$$D_3 = \begin{cases} 1, & \text{kui on transport} \\ 0, & \text{kui ei ole transport.} \end{cases}$$

2. Millised fiktiivsed tunnused on olulised nivool 0,01, nivool 0,05 ja nivool 0,1?
3. Kas töötasu sõltub ettevõtte tegevusalast?
4. Kui suur on keskmine töötasu ettevõttes, mille käive aastas on 10 miljardit dollarit, omakapitali rentaablus 20 ja mille tegevusala on tööstus?
5. Kui suur on keskmine töötasu ettevõttes, mille käive aastas on 10 miljardit dollarit, omakapitali rentaablus 20 ja mille tegevusala on finantsvahendus?
6. Millisel tegevusalal on töötasu kõige suurem ja millisel kõige väiksem, kui muud tunnused on samad?
7. Kui palju on finantsvahendusega tegelevas ettevõttes keskmise töötasu naturaallogaritm suurem kui tööstusettevõttes, kui muud tunnused on samad?
8. Kui palju on finantsvahendusega tegelevas ettevõttes keskmine töötasu suurem kui tööstusettevõttes, kui muud tunnused on samad? Näpunäide: vt lisa A.13.
9. Kui palju on transpordiga tegelevas ettevõttes keskmine töötasu väiksem kui tööstusettevõttes, kui muud tunnused on samad?

VASTUS lk 690.

A.9.29. Ülesandes A.9.15 saadi maja hinna prognoosimiseks mudel, kus seletavateks tunnusteks oli vannitubade arv, krundi pindala, garaažikohtade arv ning maja vanus. On teada ka müüdud majade konstruktsiooni tüüp (KONSTR):

1) telliskivi;
2) puitsõrestik ja telliskivi;
3) alumiiniumsõrestik;
4) puitsõrestik.

Kas maja hind sõltub ka konstruktsiooni tüübist? VASTUS lk 690.

Standardiseeritud kordajad

A.9.30. Ülesandes A.9.13 tuli leida mudel, mis kirjeldaks aastas müüdud uute autode arvu sõltuvust autode jaemüügi hinnaindeksist, elanike kogusissetulekust aastas ning registreeritud autode koguarvust.

1. Leida vastavate seletavate tunnuste standardiseeritud kordajad.
2. Milline tunnus mõjutab müüdud autode arvu kõige rohkem ja milline kõige vähem?

VASTUS lk 690.

Cobbi-Douglase tootmisfunktsioon

A.9.31. Analüüsida, kuidas Euroopa farmaatsiaettevõtete käive y (tuhat eurot) sõltub töötajate arvust L, varadest K (tuhat eurot) ning teadus- ja arenduskuludest T (tuhat eurot). Kasutada on 92 Euroopa farmaatsiaettevõtte andmed aastatest 2012–2014. Andmed on võetud Euroopa ettevõtete andmebaasist Amadeus[16] ja pärinevad erinevatest aastatest, sest iga ettevõtte korral on kasutatud viimaseid andmebaasis olevaid andmeid. Valimisse on võetud ainult need ettevõtted, millel teadus- ja arenduskulud on nullist erinevad ning töötajate arv väiksem kui 100 tuhat.

Eeldada, et käibe modelleerimiseks saab kasutada kolme sisendiga Cobbi-Douglase tootmisfunktsiooni kujul

$$y = AK^{\alpha}L^{\beta}T^{\gamma}. \tag{9.134}$$

Parameetrite hinnangute leidmiseks kasutada vastavat lineariseeritud mudelit:

$$\ln y = c + \alpha \ln K + \beta \ln L + \gamma \ln T + \varepsilon.$$

Peale parameetrite hindamist kirjutada välja mudel kujul (9.134) ning interpreteerida mudeli parameetreid α, β ja γ. VASTUS lk 690.

A.9.32. Cobbi-Douglase tootmisfunktsioon on saanud oma nime USA majandusteadlaste Charles Cobbi (1875–1949) ja Paul Douglase (1892–1976) järgi, kes 1928. aastal avaldasid töö USA tootmisfunktsiooni kohta (Cobb ja Douglas, 1928). Nad kasutasid andmeid aastatest 1899–1922 ning USA kogutoodang Q, kapital K ja tööjõud L olid alusindeksid, nii et 1899. aasta väärtuseks oli kõigil tunnustel 100.

1. Lähtudes Cobbi-Douglase tootmisfunktsioonist kujul

$$Q = AK^{\alpha}L^{\beta}, \tag{9.135}$$

   hinnata vastavat lineariseeritud mudelit.
2. Leida astmenäitajate α ja β summa.

[16] `https://amadeus.bvdinfo.com/`

Kui tootmisfunktsioonis (9.135) astmenäitajate summa $\alpha + \beta = 1$, siis tuleb hinnata vaid üht astmenäitajat ning Cobbi-Douglase tootmisfunktsioon on kujul (Kerem jt, 1988, lk 35)

$$Q = AK^{\alpha}L^{1-\alpha}. \tag{9.136}$$

Sellisel juhul tuleb kogutoodangu tõstmiseks 1% võrra nii kapitali kui ka tööjõudu suurendada 1% võrra. Astmenäitaja α hindamiseks regressioonanalüüsi abil teisendatakse mudelit (9.136) järgmiselt:

$$\begin{aligned} \ln Q &= \ln A + \alpha \ln K + (1-\alpha)\ln L, \\ \ln Q - \ln L &= \ln A + \alpha(\ln K - \ln L), \\ \ln \frac{Q}{L} &= \ln A + \alpha \ln \frac{K}{L}. \end{aligned} \tag{9.137}$$

3. Kasutades lineaarset mudelit (9.137), kus sõltuvaks tunnuseks on $\ln(Q/L)$ ning seletavaks tunnuseks $\ln(K/L)$, leida selle mudeli parameetrite hinnangud.
4. Lähtudes lineaarse mudeli (9.137) konstandi $\ln A$ hinnangust, arvutada konstandi A väärtus ja kirjutada mudel kujul (9.136).
5. Cobb ja Douglas said oma töös USA tootmisfunktsiooni kujul

$$Q = 1{,}01K^{1/4}L^{3/4}.$$

   Kas hinnatud mudel ühtib C. Cobbi ja P. Douglase leitud mudeliga?

VASTUS lk 691.

Kulufunktsioon ja prognoosimine

A.9.33. Regressioonanalüüs on üks meetod, mida kasutatakse kulude analüüsimisel ja prognoosimisel. Erinevat liiki kulude prognoosimiseks on vaja teada, kuidas vastavad kulud sõltuvad kulujuhtidest, s.t on vaja teada kulufunktsiooni. Kulujuhid (*cost drivers*) on tegurid, mis määravad ära vastavate kulude suuruse. Kulufunktsioonis on funktsioontunnuseks kulud ja seletavateks tunnusteks kulujuhid.

Continental Airlines on maailmas suuruselt kaheksas lennufirma. 2008. aasta, mil algas ülemaailmne majanduskriis, mõjus ka sellele ettevõttele ning põhjustas tulude vähenemise. Ajakirjas Issues in Accounting Education 2011. aastal ilmunud artiklis on toodud Continentali kogutulu, tegevuskulud komponentide kaupa ja tegevusega seotud kulujuhid ajavahemikul 2000 I kuni 2008 IV kvartal (Román, 2011). Lennufirma tegevuskulud jagunevad kümneks kululiigiks, andmed on toodud lehel A.9.33 (a). Kogutulu ja kõik kulud on miljonites dollarites. Kulude kirjeldused koos potentsiaalsete kulujuhtidega on järgmised:

1) Kulud kütusele sõltuvad kütuse hinnast ja kütusekulust gallonites.

2) Tööjõukulu sisaldab lendurite, stjuuardesside ja maapealse personali tööjõukulusid. Kulujuhiks võiks olla kohtmiilid. Kohtmiil on lennuliini pikkuse (miilides) ja lennuki mahutavuse (kohtade arv) korrutis.
3) Kulud siseliinidele on tingitud lendude ostmisest siseliine teenindavatelt lennufirmadelt (Expressjet, Chautauca, CommutAir ja Cogan). Kulujuhiks on kohtmiilid siseliinidel.
4) Lennukite liisimiskulud sõltuvad liisitud lennukite arvust.
5) Lennujaamamaksudel on kulujuhiks reisijate arv.
6) Turustuskulud sisaldavad mitmesuguseid allahindlusi ning komisjonitasusid broneeringuid vahendavatele firmadele ja on määratud reisija poolt makstava tasuga. Kulujuhiks võiks olla kogutulu.
7) Lennukite hoolduskulud sisaldavad perioodilist hooldust, varuosi, lennukikere ja mootori remonti. Kulujuhiks võiks olla lennukite koguarv või kohtmiilid.
8) Amortisatsioonikulu sisaldab lennukite ja maapealse vara amortisatsiooni ning on põhiliselt määratud varade suuruse ja raamatupidamisreeglitega. Planeerimisel võib võtta aluseks eelnevate kvartalite aritmeetilise keskmise.
9) Reisijate teenindus sisaldab reisile registreerimist, toitlustust, pagasi käitlemist, lennukisalongi koristamist ning kulujuhiks on reisijate arv.
10) Muud kulud sisaldavad mitmesuguseid kõrvaltegevusega seotud kulusid nagu turvateenused, reklaamikulud, muud varustuskulud jms. Kulujuhid pole täpselt määratletud. Kuna suur osa muudest kuludest on reklaamikulud, mis on määratud osana müügitulust, siis peamiseks kulujuhiks võiks olla kogutulu.

Lehel A.9.33 (b) on Continental Airlinesi kogutulu ja kulujuhtide prognoos 2009. aasta neljaks kvartaliks. Kütuse hind on määratud pikaajaliste lepingutega, millega on fikseeritud ka ostetatava kütuse kogus.

1. 2000 I kuni 2008 IV kvartali andmete põhjal leida kõik 10 kulufunktsiooni. Regressioonanalüüsi käigus kontrollida ka soovitatud kulujuhi statistilist olulisust nivool 0,1. Kui kulujuht ei ole statistiliselt oluline, siis sõltuvus sellest pole tõestatud ning prognoosimisel võib aluseks võtta vastavate kulude keskmise väärtuse.
2. Kasutades 2009. aasta kulujuhtide ja kogutulu prognoose ning leitud kulufunktsioone, leida 2009. aasta kvartalite jaoks
   a) kõikide kululiikide prognoosid;
   b) kogukulude prognoos;
   c) kasumi prognoos;
   d) kasumi prognoos aastaks 2009.

Vastus lk 691.

Peatükk 10

Aegread

Seni vaatlesime peamiselt kogumeid, kus olid mingite tunnuste väärtused erinevate objektide jaoks ühel ja samal ajaperioodil (või -momendil). Selliseid andmeid nimetatakse ristandmeteks ehk läbilõikeandmeteks (*cross-sectional*). Erinevate nähtuste ja protsesside analüüsimisel on aga väga oluline ka ajas toimuva muutumise ehk dünaamika jälgimine. Kuna majandusprotsessidel on teatud inertsust, siis teades mingi tunnuse käitumist minevikus, on teatud täpsusega võimalik prognoosida selle käitumist tulevikus.

10.1. Aegrea mõiste

Aegreaks ehk kronoloogiliseks reaks nimetatakse arvandmete rida, mis kirjeldab suuruse ajalist muutumist. Aegrida saadakse korduvvaatluse läbiviimisel. Harilikult esitatakse aegrida ajamomentide (kuupäev, kellaaeg) või ajaperioodide (kuu, kvartal, aasta) ja neile vastavate suuruste väärtuste kogumina, nii et iga kogumi elemendi jaoks on fikseeritud ajamoment või -periood, millele see väärtus vastab. Aegrea väärtused on alati järjestatud. Kui ristandmete analüüsimise korral pole objektide järjestus oluline, siis aegrea korral on tunnuse väärtuste järjestus oluline ja selle muutmine ei ole lubatud.

Tunnuste liigitus

Tunnuste liigitus aegridade analüüsil:

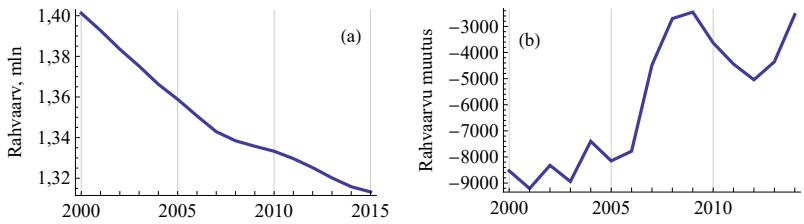
- **varusuurused** on sellised, mille väärtuse saab leida mingi ajamomendi jaoks. Näiteks hind, klientide arv, tööjõu, kapitali või materjali hulk;
- **voosuurused** on sellised, millel esineb väärtus mingi ajaperioodi jaoks. Näiteks kulud, tulu, kasum;
- **intensiivsussuurused** on erinevate varu- ja voosuuruste suhted. Näiteks tööviljakus, mitmesugused rahandussuhtarvud.

Moment- ja perioodread

Aegread jagunevad kahte klassi:

- **momentread** — iga element on seotud teatud ajamomendiga;

- **perioodread** — iga element on seotud mingi ajavahemikuga, perioodiga;
  - võrdperioodsed, kus ajamomentide vahed on võrdsed;
  - mittevõrdperioodsed, kus ajamomentide vahed ei ole võrdsed.



Joonis 10.1. (a) Momentrida: Eesti rahvaarv (mln) seisuga 1. jaanuar. (b) Perioodrida: Eesti rahvaarvu muutus aastas

Momentreast võime saada perioodrea. Näiteks elektrienergia arvesti näidu muutumine ajas on momentrida. Selle alusel saame elektrienergia kulu tunnis (päevas, kuus), mis on perioodrida. Varude suurus on momentrida, varude muutus perioodrida. Rahvaarv on momentrida (joonis 10.1 (a)), rahvaarvu muutus perioodrida (joonis 10.1 (b))[1].

Ajaliselt **sõre** suurus — tunnuse väärtusel on mõte vaid esitatud ajamomentidel, vahepealsetel ajamomentidel tunnuse väärtus puudub, mõtet pole. Näiteks aktsia sulgemishind: hind eksisteerib vaid kauplemispäevadel, s.o tööpäevadel. Puhkepäevadel väärtus puudub. Pideva suuruse korral on väärtus olemas ka vahepealsetel ajamomentidel, puuduvad vaid vaatlusandmed nende kohta.

Aegridade analüüsi eesmärgid

Aegridade analüüsimisel on järgmised **eesmärgid**:

1) aegrea omaduste kirjeldamine ja selle aluseks oleva protsessi parameetrite hindamine;
2) aegrea mudeli konstrueerimine, mis kirjeldaks aegreana esitatud tunnuse käitumist;
3) tunnuse väärtuste prognoosimine;
4) aegrida genereeriva protsessi mõjutamine.

Üks tähtsamaid aegridade analüüsi eesmärke on prognoosimine. Äriplaneerimisel kasutatakse aegridadel põhinevat prognoosimist näiteks nõudluse prognoosimisel, varude juhtimisel, transpordi planeerimisel, tootmise planeerimisel. Börsil kauplejad prognoosivad aktsiate tulumäärade muutumist. Makromajanduse tasemel prognoositakse sisemajanduse koguprodukti, keskmist palka, töötuse määra, intressi-

[1] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel RV021: rahvastik, 1. jaanuar.

määrasid jms, mida valitsused kasutavad strateegiate väljatöötamisel ja riigieelarve koostamisel.

Aegridade analüüsis võib eristada kahte taset:

- **elementaaralanalüüsil** püütakse aegrida kirjeldada võimalikult lihtsalt, uuritakse keskmist taset, kasutatakse erinevaid silumismeetodeid, iseloomustatakse muutlikkust ja leitakse arengutrende;
- **kompleksanalüüsi** korral eraldatakse aegreast trend ja perioodiline komponent ning uuritakse neid eraldi.

10.2. Aegridade keskmised tasemed

Keskmise taseme määramine sõltub sellest, kas tegemist on periood- või momentreaga. Perioodridade korral kasutatakse keskmiste tasemete leidmiseks aritmeetilist keskmist. Võrdperioodsete ridade korral kasutatakse lihtsat aritmeetilist keskmist

$$\bar{x} = \frac{\sum x_i}{T}, \tag{10.1}$$

kus x_i on suuruse väärtus i-ndal perioodil ja T on perioodide arv. Näiteks riikliku statistilise aastaaruande „Majandusnäitajad“ täitmisel tuleb leida aasta keskmine töötajate arv. Selleks summeeritakse 12 kuu keskmised töötajate arvud ja jagatakse 12-ga.

Mittevõrdperioodsete ridade korral kasutatakse kaalutud aritmeetilist keskmist

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}. \tag{10.2}$$

Kaalude f_i leidmiseks tuleb perioodid ühtlustada. Selleks valitakse mingi lühem ajavahemik, mille pikkuse kordseteks on kõik vaatlusperioodide pikkused. Tihti on see lühima vaatlusperioodi pikkus. Kui on vaja, teisendatakse suuruse väärtused ühtlustatud perioodidele vastavateks ja seejärel kasutatakse kaalutud keskmise valemit, kus kaal näitab, mitmel ühtlustatud perioodil vastav väärtus x_i esines.

Näide 10.1. Keskmine käive kvartalis

Ettevõtte 2013. aasta I kvartali käive oli 0,5 miljonit eurot, II kvartali käive oli 0,55 miljonit eurot ja 2013. aasta teise poolaasta käive 1,3 miljonit eurot. 2014. aastal oli keskmine kvartalikäive 0,75 miljonit eurot. Leiame keskmise kvartalikäibe perioodil 2013–2014.

Ühtlustatud perioodiks on sobiv valida kvartal.

Periood	Käive perioodil, mln €	Keskmine kvartalikäive antud perioodil x_i, mln €	Kaal f_i	$f_i x_i$
2013. a I kv	0,5	0,5	1	0,5
2013. a II kv	0,55	0,55	1	0,55
2013. a II poolaasta	1,3	$1{,}3/2 = 0{,}65$	2	1,3
2014. a		0,75	4	3
KOKKU			8	5,35

Leiame kaalutud aritmeetilise keskmise:

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{5{,}35}{8} \approx 0{,}67.$$

Vastus: keskmine kvartalikäive antud ajavahemikul oli 0,67 miljonit eurot.

Pideva suuruse momentrea korral on keskmise leidmine komplitseeritum, sest me ei tea suuruse väärtusi vahepealsetel ajamomentidel. Momentridade korral kasutatakse **kronoloogilist keskmist**. Selle leidmiseks teisendatakse momentrida kõigepealt perioodreaks ja leitakse perioodide keskmised. Perioodi keskmine on perioodi alg- ja lõppmomendi väärtuste aritmeetiline keskmine. Seejärel leitakse perioodide keskmiste aritmeetiline keskmine.

Kronoloogiline keskmine

Kronoloogiline keskmine on momentrea andmete põhjal leitud perioodrea perioodide keskmiste aritmeetiline keskmine.

Näide 10.2. Keskmine kassaseis ja kronoloogiline keskmine

Olgu meil teada ettevõtte kassaseis iga kuu algul. Paneme tähele, et tegemist on momentreaga.

Kuupäev	1.01	1.02	1.03	1.04	1.05	1.06	1.07
Kassaseis, tuh €	40	190	30	660	580	420	620

Leiame poolaasta keskmise kassaseisu. Algul leitakse kuude keskmised, siis kvartalite keskmised ja nende põhjal poolaasta keskmine.

	I kvartal			II kvartal		
	jaan	veebr	märts	apr	mai	juuni
Kuu keskmine kassaseis, tuh €	$\frac{40+190}{2}=115$	$\frac{190+30}{2}=110$	345	620	500	520
Kvartali keskmine kassaseis, tuh €	$\frac{115+110+345}{3}=190$			546,7		
Poolaasta keskmine kassaseis, tuh €	$\frac{190+546{,}7}{2}\approx 368{,}3$					

Näites 10.2 tehtud järkjärgulised arvutused võib poolaasta keskmise kassaseisu leidmiseks välja kirjutada ühe arvutusena, kus kasutame vahetult esialgse momentrea väärtusi:

$$\frac{40+2\cdot 190+2\cdot 30+2\cdot 660+2\cdot 580+2\cdot 420+620}{12}\approx 368{,}3.$$

Tegemist on kaalutud aritmeetilise keskmisega, kus momentrea keskmiste väärtuste kaal on kaks korda suurem äärmiste väärtuste kaalust.

Kronoloogiline keskmine momentreast, millel on T väärtust:

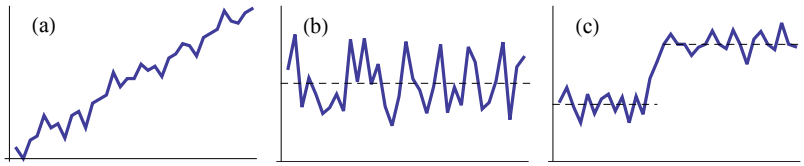
$$\bar{x}_{kron}=\frac{\frac{x_1}{2}+x_2+\ldots+x_{T-1}+\frac{x_T}{2}}{T-1}. \qquad (10.3)$$

Loomulikult tuleb enne aegrea keskmise taseme leidmist otsustada, kas keskmine tase on üldse antud aegrea korral esinduslik karakteristik. Mõnel aegreal keskmine tase puudub ja mõne aegrea korral võib keskmine tase muutuda (vt joonis 10.2).

10.3. Juurdekasvud ja kasvutempod

Muutuste iseloomustamisel kasutatakse enamasti kolme tüüpi suurusi:

- absoluutne ahel- ja alusjuurdekasv;
- suhteline ahel- ja alusjuurdekasv ehk juurdekasvutempo;
- suhteline ahel- ja aluskasvutempo ehk indeks.



Joonis 10.2. Aegrea (a) korral keskmine tase ilmselt puudub. Aegrea (b) korral on keskmine tase olemas ja aegrea (c) korral võib erinevatel perioodidel näha kaht erinevat keskmist taset

Nähtust iseloomustava suuruse absoluutset muutumist iseloomustab aegrea kahe väärtuse vahe, mida nimetatakse **absoluutseks juurdekasvuks**.

Aheljuurdekasv ja alus-juurdekasv

Aheljuurdekasv on absoluutne juurdekasv aegrea **eelmise** väärtusega võrreldes:

$$\Delta y^a = y_t - y_{t-1}, \tag{10.4}$$

kus y_t on aegrea elemendi väärtus vaadeldaval ajamomendil või perioodil, y_{t-1} elemendi väärtus eelmisel ajamomendil (perioodil).

Alusjuurdekasv on absoluutne juurdekasv mingi varasema **baasiks** võetava väärtusega võrreldes:

$$\Delta y^b = y_t - y_0, \tag{10.5}$$

kus y_0 on aegrea elemendi väärtus baasiks võetud ajamomendil (perioodil).

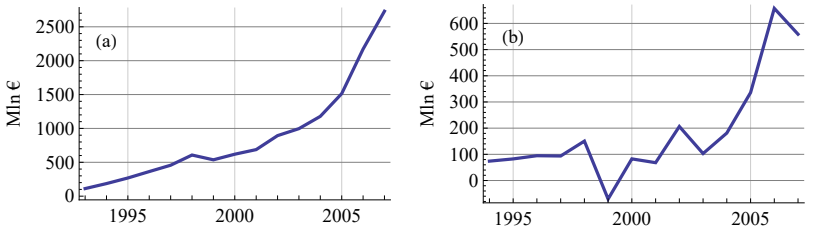
Absoluutne juurdekasv võib olla nii positiivne kui ka negatiivne:
- $\Delta y > 0$ korral on tegemist kasvamisega;
- $\Delta y < 0$ korral esineb kahanemine.

Näide 10.3. Eestis tehtud ehitustööde maht aastatel 1993–2007

N10 Aegread
N10.3

Joonisel 10.3 (a) on omal jõul Eestis tehtud ehitustööd jooksevhindades aastatel 1993–2007, miljonit eurot[a]. Diagrammil (b) on ehitustööde absoluutne aheljuurdekasv. Sellelt jooniselt on näha, et aastatel 1994–1997 oli juurdekasv ligikaudu ühesugune. 1998. aastal absoluutne juurdekasv suurenes, aga 1999. aastal oli juurdekasv negatiivne: ehitustööde maht vähenes. Kõigil teistel aastatel oli absoluutne juurdekasv positiivne ja suurenes jõud-

sasti aastatel 2004–2006. Aastal 2007 oli juurdekasv väiksem kui aastal 2006. Seda on vasakpoolselt diagrammilt (a) raske näha.



Joonis 10.3. (a) Eestis tehtud ehitustööd jooksevhindades, mln €. (b) Ehitustööd Eestis, absoluutne aheljuurdekasv, mln €

[a] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel EH02: omal jõul Eestis tehtud ehitustööd jooksevhindades.

Keskmine absoluutne juurdekasv leitakse aritmeetilise keskmise abil:

Keskmine absoluutne juurdekasv

$$\overline{\Delta y} = \frac{1}{n}\sum_{t=1}^{n} \Delta y_t. \tag{10.6}$$

Seda valemit kasutatakse siis, kui on antud absoluutsed juurdekasvud ning n on siin juurdekasvude arv. Kui aga soovime keskmist absoluutset juurdekasvu leida aegrea alg- ja lõppväärtuse abil, siis

$$\begin{aligned}\overline{\Delta y} &= \frac{\Delta y_1 + \Delta y_2 + \ldots + \Delta y_n}{n} = \\ &= \frac{y_2 - y_1 + y_3 - y_2 + \ldots + y_T - y_{T-1}}{T-1} = \frac{y_T - y_1}{T-1}.\end{aligned} \tag{10.7}$$

Paneme tähele, et juurdekasvude arv on ühe võrra väiksem y_t väärtuste arvust ehk aegrea pikkusest T: $n = T - 1$.

Jooniselt 10.3 (b) on näha, et keskmine absoluutne juurdekasv ei sobi terve ajavahemiku 1994–2007 juurdekasvude iseloomustamiseks, sest pärast 2004. aastat on toimunud oluline juurdekasvu suurenemine. Eestis tehtud ehitustööde keskmise absoluutse juurdekasvu võib leida aastateks 1994–2004 ja see on 96,9 miljonit eurot.

Juurdekasvutempo ehk **suhteline juurdekasv** on absoluutse juurdekasvu ja suuruse väärtuse jagatis. Sõltuvalt sellest, kas jagatakse eelmise väärtusega või vaadeldava perioodi väärtusega, tuntakse aheljuurdekasvutempot ja tagasivaatavat aheljuurdekasvutempot.

Juurdekasvutempo

Aheljuurdekasvutempo iseloomustab seda, kuidas suuruse väärtus on muutunud perioodi algusega võrreldes :

$$j^a = \frac{\Delta y^a}{y_{t-1}} = \frac{y_t - y_{t-1}}{y_{t-1}}. \tag{10.8}$$

Tagasivaatav aheljuurdekasvutempo iseloomustab, kuidas muutus suurus perioodi jooksul, võrreldes väärtusega perioodi lõpus:

$$j^{tva} = \frac{\Delta y^a}{y_t} = \frac{y_t - y_{t-1}}{y_t}. \tag{10.9}$$

Juurdekasvutempo võib olla nii positiivne kui ka negatiivne. Sellest sõltuvalt esineb

- kasvamine, kui $\frac{\Delta y^a}{y_{t-1}} > 0$;
- kahanemine, kui $\frac{\Delta y^a}{y_{t-1}} < 0$.

Kasvutempo ehk indeks on nähtust iseloomustava arvväärtuse suhe mingi eelmisel ajamomendil (või perioodil) olnud arvväärtusesse.

Ahelindeks ja alusindeks

Ahelindeks on kasvutempo aegrea eelmise elemendiga võrreldes

$$i^a = \frac{y_t}{y_{t-1}}, \tag{10.10}$$

kus y_t on aegrea elemendi väärtus vaadeldaval ajamomendil või perioodil ja y_{t-1} aegrea elemendi väärtus eelmisel ajamomendil (perioodil).

Alusindeks on kasvutempo mingi muu aluseks võetud ajamomendi või perioodi (baasi) suhtes

$$i^b = \frac{y_t}{y_0}, \tag{10.11}$$

kus y_0 on aegrea elemendi väärtus baasiks võetud ajamomendil (perioodil).

Kasvamise või kahanemise määramiseks tuleb kasvutempot i võrrelda arvuga 1. Analüüsitav suurus

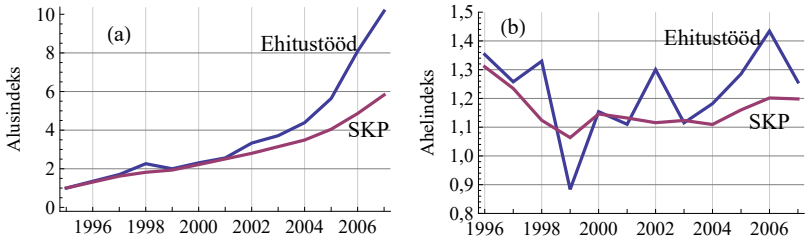
- kasvab, kui $i > 1$;
- jääb samaks, kui $i = 1$;
- kahaneb, kui $i < 1$.

Ahelindeks näitab kasvamist (kahanemist) eelmise elemendiga y_{t-1} võrreldes, alusindeks võrdleb aga baasperioodi y_0 väärtusega.

Näide 10.4. Eestis tehtud ehitustööde mahu ja Eesti SKP indeksid


N10Aegread
N10.4

Kui me soovime Eestis tehtud ehitustööde mahtude aegrida võrrelda Eesti SKP aegreaga[a], siis nende suuruste absoluutarve ei saa me esitada ühel ja samal diagrammil. Põhjuseks on nende arvude väga suur erinevus: ehitustööde mahu maksimumväärtus on ligikaudu 2,7 miljardit eurot (aastal 2007), aga SKP väärtus oli samal aastal ligikaudu 16 miljardit eurot. Nüüd tuleb appi alusindeks.



Joonis 10.4. Eesti SKP ja Eestis tehtud ehitustööd jooksevhindades. (a) Alusindeksid, 1995. aastal on indeksi väärtus 1. (b) Ahelindeksid

Joonisel 10.4 (a) on mõlema suuruse alusindeksid. Baasaastaks on 1995, s.t aastal 1995 on mõlema suuruse alusindeksi väärtus üks. Näeme, et kuni aastani 2001 kasvasid mõlemad suurused ligikaudu ühesuguse kiirusega, alusindeksid on ligikaudu samad. Alates aastast 2001 kasvas aga ehitustööde maht kiiremini kui SKP ja kasvukiiruse erinevus suurenes.
Joonisel 10.4 (b) on mõlema suuruse ahelindeksid. Ahelindeksit kasutame siis, kui tahame analüüsida muutust eelmise perioodiga võrreldes. Näeme, et aastatel 1996–1999 SKP küll kogu aeg kasvas (ahelindeks on suurem kui üks), kuid kasv aeglustus. Ehitustööde mahu ahelindeksi väärtus on aastal 1999 väiksem kui üks. Järelikult 1999. aastal ehitustööde maht 1998. aastaga võrreldes vähenes. Kõige rohkem kasvas ehitustööde maht aastal 2006, siis oli kasv eelneva aastaga võrreldes 43% (ahelindeks 1,43).

[a] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel RAA0012: sisemajanduse koguprodukt ja kogurahvatulu.

Oletame, et ettevõtte käive Y oli 2015. a detsembris 2000 eurot. 2016. a jaanuaris oli käibe kasvutempo 1,05, veebruaris 1,1 ja märtsis 0,9. Et leida märtsikuu käivet, tuleb detsembrikuu käive korrutada

järjest jaanuari, veebruari ja märtsi kasvutempoga:

$$Y_{03.16} = 2000 \cdot 1{,}05 \cdot 1{,}1 \cdot 0{,}9 = 2079.$$

Kui suur oli aga kuu keskmine kasvutempo 2016. aasta I kvartalis? Alapeatükis 2.6 nägime, et sellisel juhul tuleb kasutada geomeetrilist keskmist:

$$\bar{i^a} = \sqrt[3]{1{,}05 \cdot 1{,}1 \cdot 0{,}9} \approx 1{,}013.$$

Kontrollimiseks kasutame märtsikuu käibe leidmiseks keskmist kasvutempot:

$$Y_{03.16} = 2000 \cdot 1{,}013 \cdot 1{,}013 \cdot 1{,}013 \approx 2079.$$

Keskmine kasvutempo

Keskmine kasvutempo leitakse üksikute kasvutempode geomeetrilise keskmisena:

$$\bar{i^a} = \sqrt[n]{\prod_{k=1}^{n} i_k^a}, \tag{10.12}$$

kus n on kasvutempode arv.

Kui me soovime aga keskmist kasvutempot leida aegrea alg- ja lõppväärtuse abil, siis valemist (10.12)

$$\bar{i^a} = \sqrt[n]{i_1^a \cdot i_2^a \cdot \ldots \cdot i_n^a} = \sqrt[T-1]{\frac{y_2}{y_1} \cdot \frac{y_3}{y_2} \cdot \ldots \cdot \frac{y_T}{y_{T-1}}} = \sqrt[T-1]{\frac{y_T}{y_1}}, \tag{10.13}$$

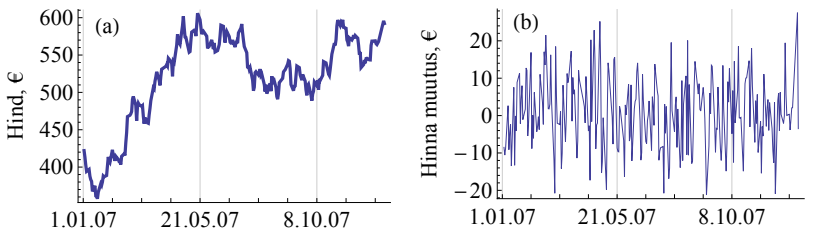
kus T on aegrea liikmete arv.

Lisaks juurdekasvude keskmiste leidmistele on tihti otstarbekas uurida juurdekasvude jaotust (jaotushistogrammi) ning selle jaotuse sümmeetriat ja püstakust.

N10 Aegread
N10.5

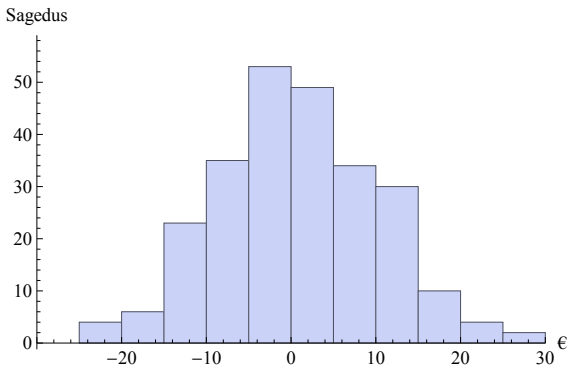
Näide 10.5. Bensiini maailmaturu hinna dünaamika

Joonisel 10.5 (a) on toodud ühe tonni bensiini turuhind ajavahemikul 2.01.–28.12.2007 (Terep, 2008). Hind on teisendatud eurodesse. On näha, et keskmist hinda ei ole sobiv kasutada, sest esimese viie kuu jooksul oli kasvutrend. Seepärast analüüsime juurdekasvusid.
Joonisel 10.5 (b) on esitatud hinna absoluutne aheljuurdekasv.



Joonis 10.5. (a) Ühe tonni bensiini turuhind 2.01.–28.12.2007. (b) Hinna absoluutne aheljuurdekasv

Joonisel 10.6 on toodud hinna absoluutsete aheljuurdekasvude jaotust iseloomustav histogramm. On näha, et kõige sagedamini esines kuni viieeuroseid hinnalangusi (aheljuurdekasv 0 kuni -5 €). Vaatlusandmete põhjal arvutatud keskmine aheljuurdekasv on aga positiivne, 0,682 eurot, aritmeetiline keskmine jääb moodist paremale. Asümmeetriakordaja on 0,12, s.t jaotus on väikese parempoolse asümmeetriaga. See on tingitud sellest, et maksimaalsed hinnatõusud on suuremad (25–30 eurot) kui maksimaalsed hinnalangused (juurdekasv -20 kuni -25 eurot). Püstakuse kordaja on $-0{,}32$ ja nagu näha ka jooniselt, on tegemist suhteliselt lameda jaotusega.



Joonis 10.6. Bensiini hinna aheljuurdekasvude jaotus 2.01.–28.12.2007

10.4. Aegridade silumine

Aegridade silumise ehk tasandamise eesmärgiks on mitmesuguste perioodiliste ja juhuslike muutuste kõrvaldamine ning arengutendentside väljaselgitamine. See võimaldab saada rohkem informatsiooni aegrea

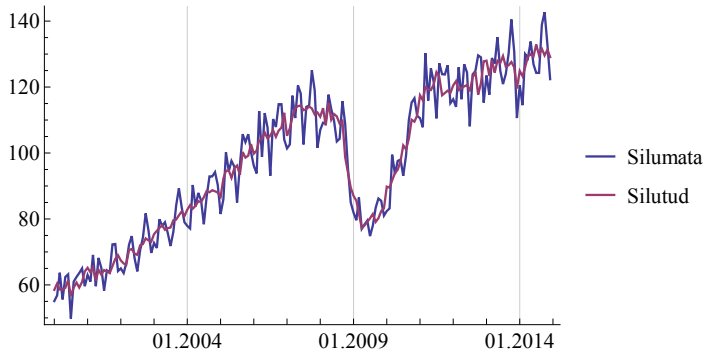
kirjeldatud suuruse muutumise kohta, selgitada välja trende ning teha prognoose.

Tuntumad **silumismeetodid** on:

Silumismeetodid

- libisev keskmine (*moving average*, MA):
  - lihtne libisev keskmine (*simple moving average*, SMA);
  - kaalutud libisev keskmine (*weighted moving average*, WMA);
- eksponentsilumine (*exponential smoothing*, ES):
  - lihtne eksponentsilumine;
  - trendiga eksponentsilumine;
  - trendi ja sesoonsusega eksponentsilumine ehk Holti-Wintersi mudel;
- silumine regressioonjoonega:
  - lineaarne regressioon;
  - mittelineaarne regressioon (eksponentsiaalne, logaritmiline, polünoom).

Joonisel 10.7 on toodud tööstustoodangu mahuindeks jaanuar 2007 – detsember 2014, silumata ja silutud väärtus[2]. Silutud väärtust on vaja selleks, et kõrvaldada sesoonsed kõikumised ja muuta võrreldavate perioodide tingimused sarnasemaks. Lisaks korrigeeritakse seda indeksit tööpäevade arvuga kuus, et indeks ei oleks mõjutatud tööpäevade arvust (erinevatel kuudel on erinev arv tööpäevi). Lähemalt võib sesoonsest korrigeerimisest lugeda Eesti Statistikaameti peametoodiku Mihkel Tähe artiklist „Aegridade sesoonne korrigeerimine“ (Täht, 2007).



Joonis 10.7. Tööstustoodangu mahuindeks, 2010. aastal on indeksi väärtus 100. Silumata väärtus ning sesoonselt ja tööpäevade arvuga korrigeeritud väärtus

Konkreetse aegrea korral tuleb alati leida sobiv silumismeetod. Sobivuse hindamiseks kasutatakse silumisjääkide ehk vigade hindamist.

[2] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabelid TO0053 ja TO0054.

Näiteks minimeeritakse keskmist ruutviga MSE (*Mean Square Error*) (vt alapeatükk 10.13).

Aegridade analüüsimiseks kasutatava tarkvara valik on väga suur. Võib kasutada tabelarvutust, mõnd statistikapaketti (SPSS, EViews, Gretl, R) või spetsiaalset tarkvara.

10.5. Libisev keskmine

Libisevaks keskmiseks nimetatakse pikemat ajavahemikku hõlmava aegrea teatavast arvust järjestikustest elementidest leitavat lühema perioodi keskmist.

Lihtne libisev keskmine

Lihtsa libiseva keskmise arvutamiseks kasutatakse lihtsat aritmeetilist keskmist aegrea viimasest q väärtusest:

$$MA_t = \frac{1}{q}\sum_{i=1}^{q} y_{t-i+1}. \tag{10.14}$$

Iga järgmise arvväärtuse leidmisel nihkutakse edasi, nii et keskmise arvutamisel hõlmatakse üks uus element ja jäetakse välja kõige varasem element. Aegrea liikmete arvu q, mida kasutatakse iga libiseva keskmise väärtuse arvutamisel, nimetatakse **libiseva keskmise sammuks**. Näiteks libisev keskmine sammuga kolm:

$$MA_t = \frac{1}{3}(y_t + y_{t-1} + y_{t-2}).$$

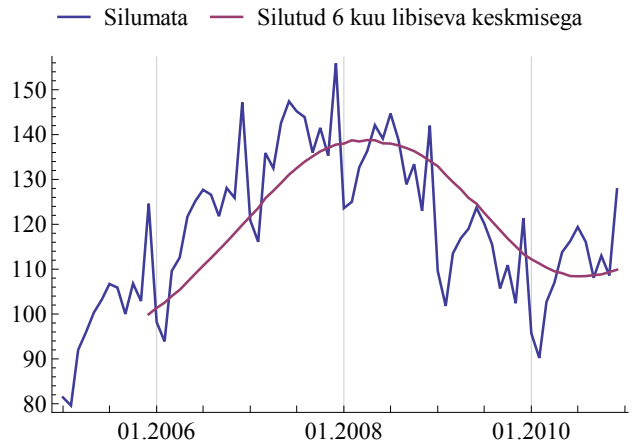
Tabelis 10.1 on näitena toodud kolme kuu libiseva keskmise arvutus.

Tabel 10.1. Kolme kuu libiseva keskmise MA arvutus

Aeg, kuudes	1	2	3	4	5	6
y	10	17	13	22	15	17
MA			$\frac{10+17+13}{3} \approx 13{,}3$	$\frac{17+13+22}{3} \approx 17{,}3$	16,7	18

Joonisel 10.8 on toodud jaemüügi mahuindeks jaanuar 2005 – detsember 2010, mida on silutud kuue kuu libiseva keskmisega[3].

[3] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel KM0022: jaemüügi mahuindeks.



Joonis 10.8. Jaemüügi mahuindeks (2005. a indeks = 100) jaanuar 2005 – detsember 2010, silumata ja silutud 6 kuu libiseva keskmisega

Libiseva keskmise väärtuse perioodi t jaoks saab avaldada ka eelmise perioodi libiseva keskmise kaudu:

$$MA_t = MA_{t-1} + \frac{y_t - y_{t-q}}{q}. \tag{10.15}$$

Tõestuseks kirjutame välja libisevad keskmised ajaperioodide t ja $t-1$ jaoks:

$$MA_t = \frac{1}{q}\left(y_t + y_{t-1} + y_{t-2} + \ldots + y_{t-q+1}\right);$$
$$MA_{t-1} = \frac{1}{q}\left(y_{t-1} + y_{t-2} + \ldots + y_{t-q+1} + y_{t-q}\right).$$

Nüüd

$$MA_t = \frac{1}{q}y_t + \frac{1}{q}\left(y_{t-1} + y_{t-2} + ... + y_{t-q+1} + y_{t-q}\right) - \frac{1}{q}y_{t-q} =$$
$$= \frac{1}{q}y_t + MA_{t-1} - \frac{1}{q}y_{t-q} = MA_{t-1} + \frac{1}{q}(y_t - y_{t-q}).$$

Libiseva keskmise sammu valik

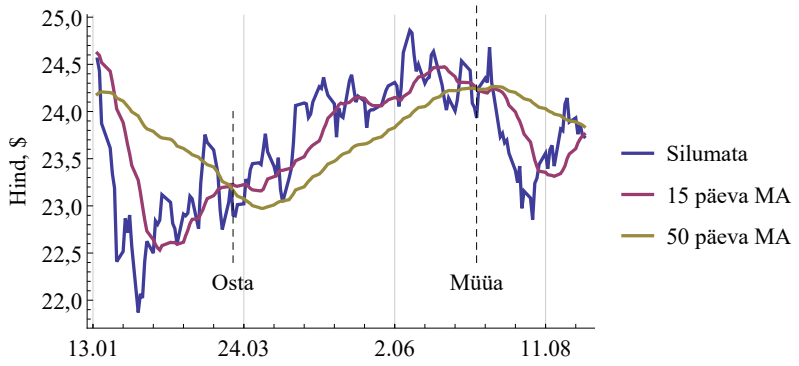
Ühe ja sama aegrea silumiseks võib kasutada erineva sammuga libisevat keskmist. Libisemissammu pikkus avaldab mõju tasandusjoone kujule. Väiksem samm silub kõikumisi vähem. Kui esineb selgesti märgatav perioodilisus, võetakse libisemissammu pikkuseks tavaliselt perioodi pikkus. Näiteks nädalasisese perioodilisuse silumiseks sobiv libisemissamm on seitse päeva, kvartalite kaupa esineva sesoonsuse silumiseks neli kvartalit jne. Joonisel 10.8 korduvad maksimumid iga kuue kuu tagant, seepärast on libiseva keskmise sammuks võetud kuus kuud. Aktsiahindade tehnilise analüüsi korral kasutatakse libisevaid keskmisi sammuga 15 päeva, 50 päeva, 200 päeva.

Näide 10.6. Libisev keskmine ja aktsiahindade tehniline analüüs

Aktsiahindade tehnilisel analüüsil prognoositakse väärtpaberi hinna muutumist tulevikus, toetudes minevikus toimunud hinna liikumisele. Investorid teevad oma ostu- ja müügiotsuseid, toetudes hinnamuutuste graafikutele ning indikaatoritele, mis vastavalt ajaloolisele kogemusele prognoosivad hinna tõusu või langust.

Üheks indikaatoriks, mille abil saab määrata hetke trendi ning jälgida selle muutumist, on libisev keskmine. Erineva pikkusega (sammude arvuga) libisevad keskmised iseloomustavad lühemaid ja pikemaid trende. Olulisteks peetakse erineva pikkusega libisevate keskmiste graafikute ristumiskohti. Näiteks kui lühema perioodi libisev keskmine lõikab alt üles pikema perioodi libisevat keskmist, on tegemist ostusignaaliga, sest siis algab kasvutrend. Müügisignaal on siis, kui lühema perioodi libisev keskmine lõikab pikema perioodi libisevat keskmist ülevalt alla. (Fontanills ja Gentile, 2001)

Joonisel on toodud ettevõtte General Electric aktsia päevase sulgemishinna graafik 15. jaanuar – 1. september 2014[a]. Lisatud on kaks libisevat keskmist: 15- ning 50-päevase sammuga. Ostusignaal on siis, kui 15-päevane libisev keskmine lõikab alt üles 50-päevast libisevat keskmist. Müügisignaal on, kui lõikamine toimub ülevalt alla.



[a]Allikas: Wolfram Research, Inc., Mathematica, Financial Data.

Prognoosimine libiseva keskmise abil

Libisevat keskmist kasutatakse ka **lühiajaliseks prognoosimiseks**: käesoleva perioodi libisev keskmine on järgmise perioodi prognoos. Niimoodi saame prognoosida ühe perioodi võrra ette.

Kasutades prognoositud väärtuse tähistamiseks tähte F (*Forecast*), võime kirjutada

$$F_{t+1} = MA_t, \tag{10.16}$$

kus F_{t+1} on väärtuse y_{t+1} prognoos. Kui soovime teha pikemat prognoosi, siis järgmised prognoositavad väärtused on kõik võrdsed:

$$F_{t+1} = F_{t+2} = F_{t+3} = \ldots = F_{t+n}. \tag{10.17}$$

N10Aegread
N10.7

Näide 10.7. Müügimahu prognoosimine libiseva keskmise abil

Tabelis on toodud tegelikud müügimahud ning nende prognoosid 3 kuu, 6 kuu ja 12 kuu libiseva keskmise järgi. Selle kuu libisev keskmine on järgmise kuu prognoos. Näiteks

$$\begin{aligned}\text{aprilli prognoos} &= \text{märtsi libisev keskmine} = \\ &= \frac{\text{jaan müük} + \text{veebr müük} + \text{märtsi müük}}{3} = \\ &= \frac{450 + 440 + 460}{3} = 450.\end{aligned}$$

Kuu	Tegelik müügi-maht	Prognoos 3 kuu libiseva keskmise järgi	Prognoos 6 kuu libiseva keskmise järgi	Prognoos 12 kuu libiseva keskmise järgi
jaanuar	450			
veebruar	440			
märts	460			
aprill	410	450		
mai	380	437		
juuni	400	417		
juuli	370	397	423	
august	360	383	410	
september	410	377	397	
oktoober	450	380	388	
november	470	407	395	
detsember	490	443	410	
jaanuar	460	470	425	424

Juulikuu prognoos on tehtud nii kolme kui ka kuue kuu libiseva keskmise alusel. Järgmise jaanuari prognoosi tegemiseks on kasutatud nii eelneva kolme kuu, eelneva kuue kuu kui ka eelneva 12 kuu libisevat keskmist.

Kasvamise korral lihtsa libiseva keskmise väärtused hindavad suuruse väärtusi alla, kahanemise korral toimub ülehindamine (vt joonis 10.8). Seetõttu kasutatakse silumisel ka **tsentreeritud libisevat keskmist** (*Central Moving Average, CMA*). Sellisel juhul kaasatakse keskmise arvutusse väärtusele y_t eelnevad ja järgnevad väärtused. Näiteks kolmesammuline tsentreeritud libisev keskmine

$$CMA_t = \frac{y_{t-1} + y_t + y_{t+1}}{3}$$

ja viiesammuline

$$CMA_t = \frac{y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2}}{5}.$$

Kuna väärtusest y_t tuleb mõlemalt poolt võtta ühesugune arv väärtusi, on tsentreeritud libiseva keskmise sammude arv paaritu arv.

Tsentreeritud libisev keskmine

Tsentreeritud libisev keskmine on aritmeetiline keskmine aegrea $2q+1$ väärtusest, väärtusele y_t lisaks on kaasa haaratud q eelmist ja q järgmist väärtust:

$$CMA_t = \frac{1}{2q+1} \sum_{i=-q}^{q} y_{t-i}. \qquad (10.18)$$

Näide tsentreeritud libiseva keskmise leidmise kohta on tabelis 10.2.

Tabel 10.2. Kolme kuu tsentreeritud libiseva keskmise CMA arvutus

Aeg kuudes	1	2	3	4	5	6
y	10	17	13	22	15	17
CMA		$\frac{10+17+13}{3} \approx 13{,}3$	$\frac{17+13+22}{3} \approx 17{,}3$	16,7	18	

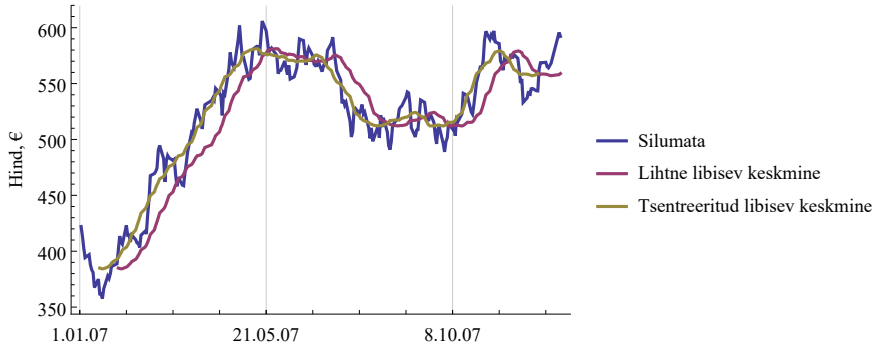
Joonisel 10.9 on näites 10.5 vaadeldud bensiini hinna aegrida silutud 21 päeva lihtsa ja tsentreeritud libiseva keskmisega.

Pikema sammuga libiseva keskmise arvutamise korral tekib küsimus, kas kõik varasemad väärtused peaksid libisevat keskmist ühepalju mõjutama. **Kaalutud libisev keskmine** (*Weighted Moving Average, WMA*) arvestab erinevaid aegrea väärtusi erineva kaaluga w_i ning kasutatakse kaalutud aritmeetilist keskmist:

Kaalutud libisev keskmine

$$WMA_t = \frac{\sum_{i=1}^{q} w_i y_{t-i+1}}{\sum_{i=1}^{q} w_i} = \frac{w_1 y_t + w_2 y_{t-1} + \ldots + w_n y_{t-q+1}}{w_1 + w_2 + \ldots + w_q}. \qquad (10.19)$$

N10Aegread
J10.9



Joonis 10.9. Ühe tonni bensiini turuhind ajavahemikul 2.01.–28.12.2007, silutud 21 päeva lihtsa ja tsentreeritud libiseva keskmisega

Millised kaalud omistada, see on analüüsija valida. Enamasti omistatakse uuritavale ajamomendile t lähemal olevatele väärtustele suurem kaal ning kaugemal olevatele väärtustele väiksem kaal:

$$w_1 > w_2 > \ldots > w_q. \tag{10.20}$$

Üks võimalus on kõige hilisema väärtuse kaal võtta võrdseks libiseva keskmise sammuga q ja eelmiste väärtuste kaalud vastavalt $q-1$, $q-2$, $\ldots, 2, 1$:

$$WMA_t = \frac{qy_t + (q-1)y_{t-1} + \ldots + 2y_{t-q+2} + y_{t-q+1}}{q + (q-1) + \ldots + 2 + 1}. \tag{10.21}$$

Sellisel juhul vähenevad kaalud lineaarselt. Jagades lugejas olevad liikmed eraldi nimetajaga läbi, võime valemi (10.21) esitada kujul, kus iga väärtuse ees on selle osakaal:

$$WMA_t = \frac{q}{\sum_{i=1}^{q} i}\, y_t + \frac{q-1}{\sum_{i=1}^{q} i}\, y_{t-1} + \ldots + \frac{1}{\sum_{i=1}^{q} i}\, y_{t-q+1}. \tag{10.22}$$

Sellist kaalude valikut kasutatakse tihti näiteks aktsiahindade tehnilisel analüüsil, kus selle meetodi ingliskeelne nimetus on *sum of the digits*.

Tabel 10.3. Sammuga 5 kaalutud libiseva keskmise WMA arvutus

Aeg	$t-4$	$t-3$	$t-2$	$t-1$	t
Aktsia hind	20	18	18	15	18
Kaal	1	2	3	4	5
Osakaal	1/15	2/15	3/15	4/15	5/15
Kaalutud väärtus	1,33	2,40	3,60	4,00	6,00
WMA					17,33

Tabelis 10.3 on toodud libiseva keskmise arvutus sellisel meetodil, kui sammuks on valitud 5. Sellisel juhul on kaalude summa $\sum_{i=1}^{q} i = 1 + 2 + \ldots + 5 = 15$.

10.6. Eksponentsilumine

Libiseva keskmise meetodist paindlikum silumismeetod on eksponentsilumine. Meetod töötati välja 1950-ndatel, autoriteks olid Robert Brown, Charles Holt ja Peter Winters. Eksponentsilumise korral omistatakse kaugemal olevatele väärtustele samuti väiksemad kaalud nagu kaalutud libiseva keskmise korral, aga kaalud vähenevad eksponentsiaalselt. Eksponentsilumise peamiseks kasutusvaldkonnaks on lühiajaline prognoosimine. Esmakordselt kasutas seda meetodit R. Brown USA mereväes varuosade nõudluse prognoosimiseks.

Eksponentsilumisel leitakse silutud väärtus E_t ajahetkele t vastava tegeliku väärtuse y_t ja eelmise silutud väärtuse E_{t-1} kaalutud keskmisena:

$$E_t = wy_t + (1-w)E_{t-1}, \tag{10.23}$$

kus w on silumiskonstant $0 \leq w \leq 1$.

Järjestikuste silutud väärtuste leidmisel võetakse esimene väärtus võrdseks tegeliku väärtusega y_1 ja edasi kasutatakse valemit 10.23:

$$\begin{aligned}
E_1 &= y_1, \\
E_2 &= wy_2 + (1-w)E_1, \\
E_3 &= wy_3 + (1-w)E_2, \\
&\ldots\ldots \\
E_t &= wy_t + (1-w)E_{t-1}.
\end{aligned}$$

Esimese silutud väärtuse E_1 leidmiseks võib kasutada ka aegrea n esimese või kõigi väärtuste aritmeetilist keskmist.

Suuremate w väärtuste korral on silumise efekt väiksem, sest y tegelik väärtus avaldab rohkem mõju. Kui $w = 1$, siis $E_t = y_t$ ja silumist ei toimu. Väiksema w korral mõjutab aegrea tegelik väärtus silutud väärtust vähem ning silumisefekt on suurem (vt näite 10.8 joonist 10.10). Pikema aegrea silumisel esineb siis aga oht, et silutud väärtused üha kaugenevad tegelikust aegreast ning prognooside tegemisel tekivad suured vead. Kui silumiskonstant $w = 0$, siis

$$E_t = E_{t-1} = E_{t-2} = \ldots = y_1,$$

s.t kõik silutud väärtused on võrdsed aegrea esimese elemendi väärtusega.

N10Aegread
N10.8,9

Näide 10.8. Indeksi S&P 500 aegrea eksponentsilumine

USA tähtsaim aktsiaindeks S&P 500 sisaldab 500 suurima USA ettevõtte aktsiaid ja katab peaaegu 80% kogu USA aktsiaturu kapitalisatsioonist (aktsia turuhinna ja aktsiate arvu korrutis). Silume selle indeksi aegrida perioodil 6.04.–1.05.2015[a] eksponentsiaalselt kahe erineva silumiskonstandiga.

		Silutud väärtus E_t	
Kuupäev	S&P 500	$w = 0{,}2$	$w = 0{,}5$
6.04.2015	2080,62	2080,62	2080,62
7.04.2015	2076,33	2079,76	2078,48
8.04.2015	2081,90	2080,19	2080,19
...	...	...	...
1.05.2015	2108,29	2103,73	2102,98

Võtame silumiskonstandiks $w = 0{,}2$. Esimene silutud väärtus 6. aprillil on võrdne sama päeva tegeliku väärtustega

$$E_1 = y_1 = 2080{,}62.$$

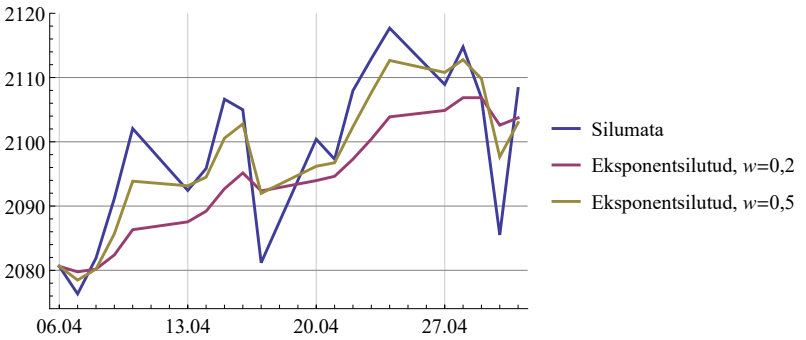
Teine silutud väärtus 7. aprillil leitakse valemist (10.23)

$$E_2 = 0{,}2 \cdot 2076{,}33 + 0{,}8 \cdot 2080{,}62 \approx 2079{,}76.$$

Kolmas silutud väärtus leitakse samast valemist:

$$E_3 = 0{,}2 \cdot 2081{,}90 + 0{,}8 \cdot 2079{,}76 \approx 2080{,}19.$$

Nii jätkatakse kuni aegrea lõpuni. Joonisel 10.10 on esitatud nii tegelikud kui ka silutud väärtused.



Joonis 10.10. Indeksi S&P 500 aegrida ja silutud väärtused 6.04.–1.05.2015

Kui silumiskonstandiks on 0,5, siis teine silutud väärtus

$$E_2 = 0{,}5 \cdot 2076{,}33 + 0{,}5 \cdot 2080{,}62 \approx 2078{,}48$$

ja kolmas silutud väärtus

$$E_3 = 0{,}5 \cdot 2081{,}90 + 0{,}5 \cdot 2078{,}48 \approx 2080{,}19.$$

Jooniselt 10.10 on näha, et suurema silumiskonstandi korral mõjutavad tegelikud väärtused silutud väärtust rohkem ning silumisefekt on väiksem.

[a] Allikas: Federal Reserve Bank of s.t. Louis, Economic Research, `https://research.stlouisfed.org`

Miks nimetatakse sellist silumist eksponentsilumiseks? Analüüsime, millised tulevad suuruse y eelnevate väärtuste kaalud, kui silutud väärtuse E_t avaldises (10.23) teeme järkjärgulise asendamise:

$$\begin{aligned}
E_1 &= y_1, \\
E_2 &= wy_2 + (1-w)y_1, \\
E_3 &= wy_3 + (1-w)\left[wy_2 + (1-w)y_1\right] = \\
&= wy_3 + w(1-w)y_2 + (1-w)^2 y_1, \\
E_4 &= wy_4 + (1-w)E_3 = \\
&= wy_4 + w(1-w)y_3 + w(1-w)^2 y_2 + (1-w)^3 y_1, \\
&\ldots\ldots \\
E_t &= w\left[y_t + (1-w)y_{t-1} + (1-w)^2 y_{t-2} + (1-w)^3 y_{t-3} + \ldots\right] + \\
&\quad + (1-w)^{t-1} y_1.
\end{aligned}$$

Aja t suurenedes ilmuvad E_t avaldisse järk-järgult varasemad y väärtused, kusjuures nende kaalud on proportsionaalsed geomeetrilise progressiooniga $\{1, (1-w), (1-w)^2, (1-w)^3, \ldots\}$. Selline geomeetriline progressioon on aga eksponentfunktsiooni $(1-w)^x$ diskreetne versioon (s.t x omandab diskreetseid väärtusi 0, 1, 2, ...).

Tuleb märkida, et mõnikord määratakse eksponentsilumisel mitte tegeliku väärtuse kaal w, vaid eelmise silutud väärtuse E_{t-1} kaal $1-w$, mida nimetatakse ka **sumbumisfaktoriks** (*damping factor*).

Juhul, kui meil on teada aegrea väärtused ajahetkeni t ja soovime **prognoosida** väärtust ajahetkel $t+1$, siis eksponentsilumise korral kasutatakse selleks eelmise ajahetke silutud väärtust E_t:

$$F_{t+1} = E_t, \tag{10.24}$$

kus F_{t+1} on väärtuse y_{t+1} prognoos. Kui soovime teha pikemat prognoosi, siis järgmised prognoositavad väärtused on kõik võrdsed:

$$F_{t+1} = F_{t+2} = F_{t+3} = \ldots = F_{t+n}. \tag{10.25}$$

Seepärast ei sobi lihtne eksponentsilumine pikemaajaliste prognoosi-de tegemiseks, eriti selliste aegridade korral, mis sisaldavad trendi ja sesoonseid kõikumisi.

Arvestades valemeid (10.23) ja (10.24), saame

$$\begin{aligned} F_{t+1} = E_t &= wy_t + (1-w)E_{t-1} = wy_t + (1-w)F_t = \\ &= F_t + w(y_t - F_t). \end{aligned}$$

Vahet $y_t - F_t$ nimetatakse **prognoosiveaks** ajahetkel t.

Prognoosimine

Prognoosimise olemus eksponentsilumisel: prognoositud väärtus ajahetkel $t+1$ on prognoositud väärtus ajahetkel t, millele on lisatud silumiskonstandiga w korrutatud prognoosiviga ajahetkel t:

$$F_{t+1} = F_t + w(y_t - F_t). \tag{10.26}$$

Tegemist on **adaptiivse prognoosimisega**: prognoositud väärtus ajahetkel $t+1$ sõltub eelmise prognoosi veast. Mida suurem on silumiskonstandi w väärtus, seda rohkem eelmise prognoosi viga arvestatakse.

N10 Aegread
N10.8,9

Näide 10.9. Indeksi S&P 500 väärtuste prognoosimine

Näites 10.8 silusime aktsiaindeksi S&P 500 aegrida kahe erineva silumiskonstandiga. Indeksi viimane väärtus aegreas oli 1. mail 2015. Prognoosime indeksi kolme järgmist väärtust ja võrdleme prognoositud väärtusi tegelike väärtustega. Kuna 2. ja 3. mail oli New Yorgi börs suletud ning nendel päevadel indeksi väärtused puuduvad, siis prognoos on tehtud 4. kuni 6. mai jaoks.
Prognoosimiseks kasutame viimast silutud väärtust. Silumiskonstandi 0,2 korral oli see 2103,73 ning konstandi 0,5 korral 2102,98.

		$w = 0{,}2$		$w = 0{,}5$	
Kuupäev	S&P 500	Prognoos	Prognoosi viga	Prognoos	Prognoosi viga
4.05.2015	2114,49	2103,73	10,76	2102,98	11,51
5.05.2015	2089,46	2103,73	−14,27	2102,98	−13,52
6.05.2015	2080,15	2103,73	−23,58	2102,98	−22,83

Eksponentsilumisel põhinev prognoosimine kokkuvõtlikult

1. Kasutades olemasolevaid vaatlusandmeid $y_1, y_2, \ldots, y_t$, leida vastavad eksponentsiaalselt silutud väärtused $E_1, E_2, \ldots, E_t$:

$$E_1 = y_1,$$
$$E_2 = wy_2 + (1-w)E_1,$$

$$E_3 = wy_3 + (1-w)E_2,$$
$$\ldots\ldots$$
$$E_t = wy_t + (1-w)E_{t-1}.$$

2. Kasutades viimast silutud väärtust, leida aegrea järgmise väärtuse prognoos:

$$F_{t+1} = E_t.$$

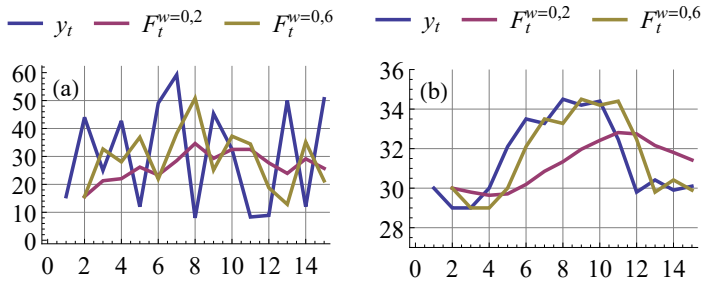
3. Eeldades, et aegrida ei sisalda trendi ega sesoonseid kõikumisi, kasutada sama prognoosi aegrea järgmiste väärtuste prognoosimiseks:

$$F_{t+2} = F_{t+1},$$
$$F_{t+3} = F_{t+1},$$
$$\vdots$$

Väga oluline on silumiskonstandi w valik. Kui w valida väike, on tulemus siledam. Suurema w korral saadakse aegrea tegelikele väärtustele lähedasem tulemus, millega kaasneb ka silumisefekti vähenemine. Konkreetse aegrea korral tuleb tavaliselt proovida mitut erinevat w väärtust, et leida sobivaim. Kui aegrida on väga varieeruv, on mõistlikum kasutada väiksemat silumiskonstanti. Seda põhjusel, et siis on prognoosi vea peamiseks põhjuseks juhuslik varieerumine ning järgmise

prognoosi tegemisel ei ole mõistlik eelmise prognoosi veale üle reageerida (valem (10.26)). Kui aegrea juhuslik varieerumine on väiksem, on prognoosi viga tingitud rohkem aegrea taseme muutumisest. Siis on parem kasutada suuremat silumiskonstanti, mis võimaldab prognoosi kiiremini kohandada aegrea taseme muutustele.

Joonisel 10.11 on mõlemal graafikul aegreale y_t lisatud eksponentsilumisel põhinev ühesammuline prognoos F_t kahe erineva silumiskonstandi kasutamisel. Ühesammuline prognoos tähendab seda, et igal ajahetkel võrdub prognoositud väärtus eelmise ajahetke silutud väärtusega, $F_t = E_{t-1}$.



Joonis 10.11. Mõlemal graafikul on aegreale y_t lisatud eksponentsilumisel põhinev ühesammuline prognoos F_t kahe erineva silumiskonstandi kasutamisel

Ühesammuline prognoosimine

Ühesammulise prognoosimise korral kasutatakse perioodile t vastava prognoosi saamiseks informatsiooni, mis on olemas perioodiks $t-1$ (vt ka tabel 10.4):

- prognoos F_t leitakse perioodile $t-1$ vastava silutud väärtuse E_{t-1} põhjal;
- minnakse samm edasi ning kasutatakse tegelikku väärtust y_t perioodile t vastava silutud väärtuse E_t saamiseks;
- prognoos F_{t+1} leitakse perioodile t vastava silutud väärtuse E_t põhjal.

Graafikul 10.11(a) olev aegrida on suure juhusliku varieeruvusega ning väiksema silumiskonstandi kasutamine annab üldiselt parema prognoosi. Graafikul (b) olev aegrida on väiksema juhusliku varieeruvusega. Prognoos, kus on kasutatud väiksemat silumiskonstanti 0,2, kohaneb aegrea taseme muutusega aeglaselt ja parem prognoos saadakse suurema silumiskonstandi 0,6 korral. Prognooside võrdlemiseks kasutatavaid arvnäitajaid vaatame alapeatükis 10.13.

Eksponentsilumiseks on Exceli andmeanalüüsi komplektis *Data Analysis* vahend *Exponential Smoothing*. Selle kasutamisel tuleb ette anda sumbumiskonstant *Damping Factor*, mis on $1-w$.

Tabel 10.4. Ühesammuline prognoosimine

t	y	Silutud väärtus E		Prognoos F
1	y_1	E_1		
			$\searrow$	
2	y_2	E_2		$F_2 = E_1$
			$\searrow$	
3	y_3	E_3		$F_3 = E_2$
			$\searrow$	
4	y_4	E_4		$F_4 = E_3$
...	...	...		...

10.7. Silumine regressioonjoonega

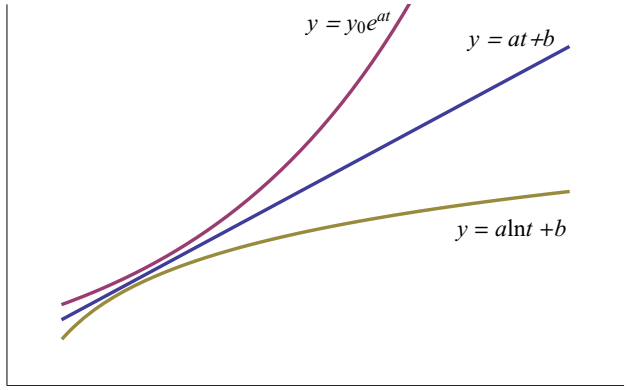
Regressioonjoonega silumise korral kasutatakse regressioonanalüüsi. Selle abil taandatakse empiiriline aegrida mingile analüütiliselt esitatavale joonele, mis iseloomustaks vastava nähtuse muutumist nn ideaalsel juhul ilma juhuslike kõrvalmõjudeta.

1. Valitakse sobiva kujuga regressioonmudel $y = f(t)$, kus argumendiks on ajamuutuja $t = 0, 1, 2, \ldots$ (mõnikord ka $t = 1, 2, \ldots$). Mudeli valikuks on kaks võimalust:
   - teooriast on teada matemaatiline mudel, mis kirjeldab uuritava suuruse muutumist ajas;
   - teoreetilist mudelit ei eksisteeri ning esialgse valiku tegemiseks võib kasutada visuaalset hindamist.
2. Leitakse regressioonmudeli parameetrite hinnangud.
3. Mudeli põhjal leitakse silutud väärtused $\hat{y}$ ning konstrueeritakse tasandusjoon.

Kui teoreetilist mudelit ei eksisteeri, siis võib konkreetse aegrea korral proovida erineva kujuga mudeleid ning sobivaim valitakse välja determinatsioonikordaja R^2 alusel. Sagedamini kasutatavad mudelid on

- lineaarne $y = at + b$;
- eksponentsiaalne $y = y_0 e^{at}$;
- logaritmiline $y = a \ln t + b$;
- n-astme polünoom $y = a_0 + a_1 t + a_2 t^2 + \ldots + a_n t^n$.

Tõlgendamise ja prognoosimise seisukohalt on esimesed kolm mudelit kõige sobivamad. Positiivse a korral kirjeldab lineaarne trend kasvamist konstantse kiirusega, võrdsetes ajavahemikes kasvab y ühepalju. Eksponentsiaalse trendi korral kasv kiireneb ja logaritmilise trendi korral aeglustub (joonis 10.12).



Joonis 10.12. Lineaarne, eksponentsiaalne ja logaritmiline kasvutrend

Ajamuutujat t kasutatakse trendi mudelites seepärast, et mudeli parameetrid oleksid paremini tõlgendatavad. Näiteks Eesti rahvaarvu muutumist aastatel 1970–1990 kirjeldab hästi lineaarne trend $\hat{N} = 1369 + 10{,}08t$, kus N on rahvaarv tuhandetes ning t aeg aastates, $t = 0$ aastal 1970 (Tiit, 1995). Mudeli vabaliige 1369 (tuhat) on rahvaarvu mudelväärtus aastal 1970. Kui me kasutaksime trendi mudelis aastaarve, oleks mudeliks $\hat{N} = -601 + 10{,}08t$ ning vabaliige -601 (tuhat) väljendaks Eesti rahavaarvu aastal 0. Ilmselt selline tõlgendus ei sobi.

Kasvukiirus

Kasvukiiruse konstantsust lineaarse trendi korral on lihtne näidata, kui arvestame, et suuruse y tuletis aja järgi on selle muutumise kiirus. Võtame lineaarse trendi mudelist tuletise aja järgi:

$$\begin{aligned} y &= at + b, \\ \frac{dy}{dt} &= a. \end{aligned} \tag{10.27}$$

Viimane avaldis näitab, et suuruse y muutumise kiirus võrdub konstandiga a.

Eksponentsiaalse kasvu korral on suuruse y muutumise kiirus võrdeline y väärtusega ajahetkel t:

$$\frac{dy}{dt} = ay. \tag{10.28}$$

Kui y suureneb, siis ka selle muutumise kiirus suureneb, mis tähendabki kasvu kiirenemist. Jagame valemi (10.28) mõlemad pooled läbi suurusega y ja korrutame diferentsiaaliga dt:

$$\frac{dy}{y} = a\,dt. \tag{10.29}$$

Viimast avaldist saame integreerida:

$$\begin{aligned}\int \frac{dy}{y} &= \int a\,dt,\\ \ln y &= at + C,\\ y &= e^{at+C},\\ y &= e^{C}e^{at}.\end{aligned}$$

Kui $t = 0$, siis $y(0) = e^C$ ja seda algväärtust tähistame y_0. Nii saame eksponentsiaalse trendi valemi

$$y = y_0 e^{at}. \tag{10.30}$$

Juhime veel tähelepanu sellele, et kui valemis (10.29) minna lõpmata väikestelt muutudelt üle lõplikele muutudele, siis saame valemi suuruse y suhtelise muudu jaoks:

$$\frac{\Delta y}{y} = a\,\Delta t. \tag{10.31}$$

Kui $\Delta t = 1$, on tegemist suhtelise muuduga ajaühikus.

Eksponentsiaalse kasvu $y = y_0 e^{at}$ korral on suuruse y suhteline muut ajaühikus konstantne ja võrdub kasvuparameetriga a:

$$\frac{\Delta y}{y} = a. \tag{10.32}$$

Eksponentsiaalse kasvu kasvumäär

Parameeter a on juurdekasvutempo, mida nimetatakse ka **kasvumääraks**.

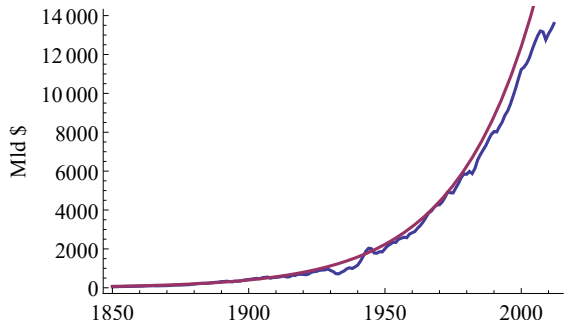
Näide 10.10. USA SKP pikaajaline kasvumäär

N10Aegread
N10.10

Leiame USA SKP pikaajalise kasvumäära, kasutades andmeid aastast 1850 (L. Johnson ja Williamson, 2016). Tegemist on SKP reaalväärtusega, s.t SKP on leitud ühe konkreetse aasta (2005) hindades ja ei ole mõjutatud dollari väärtuse muutumisest (inflatsioonist). Graafikult on näha, et kasv on eksponentsiaalne. Silumine eksponentsiaalse trendiga annab mudeliks

$$\hat{y} = 64{,}27e^{0{,}0343t},\ \ R^2 = 0{,}994,$$

kus y on SKP (miljardit 2005. aasta dollarit) ja t aeg aastates, $t = 0$ aastal 1850. Eksponentsiaalse mudeli parameeter näitab, et pikaajaliselt on SKP kasvumäär olnud keskmiselt 3,43% aastas.



Graafikult on näha, et viimasel ajal on SKP aastane kasv pidurdunud, SKP on allpool eksponentsiaalset trendi.

Leiame suuruse y muutumise kiiruse ka logaritmilise mudeli korral:

$$\begin{aligned} y &= a \ln t + b, \\ \frac{dy}{dt} &= \frac{a}{t}. \end{aligned} \tag{10.33}$$

Logaritmilise kasvu korral on kasvu kiirus pöördvõrdeline ajaga t ning aja suurenedes kasvu kiirus väheneb (joonis 10.12).


N10 Aegread
N10.11

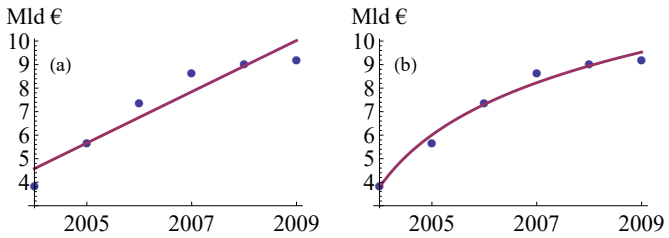
Näide 10.11. Hoiuste jääk

Kasutades Eesti Panga[a] andmeid hoiuste jäägi kohta aastatel 2004–2009, leiame sobiva trendijoone. Joonisel 10.13 on hoiuste jääki modelleeritud kahe erineva mudeliga:

(a) lineaarne mudel $\quad \hat{y} = 1{,}09t + 3{,}49, \quad R^2 = 0{,}905;$

(b) logaritmiline mudel $\quad \hat{y} = 3{,}21 \ln t + 3{,}78, \quad R^2 = 0{,}983,$

kus y on hoiuste jääk miljardites eurodes ja t on aeg aastates, $t = 1$ aastal 2004. Logaritmiline mudel on parem, sest determinatsioonikordaja R^2 on suurem ja seda on näha ka jooniselt. Logaritmilise kasvu korral kasv aeglustub. Aastal 2005 $t = 2$ ning valemist (10.33) saame, et hoiuste jäägi kasvukiirus oli ligikaudu 1,61 miljardit eurot aastas. Aastal 2009 oli aga kasvukiirus ligikaudu 0,54 miljardit eurot aastas.



Joonis 10.13. Hoiuste jäägi muutus aastatel 2004–2009. Empiirilisi andmeid on silutud kahe erineva regressioonjoonega: (a) lineaarne mudel, (b) logaritmiline mudel

[a] `http://www.eestipank.ee/`

Mõnikord tuleb aegrea silumiseks kasutada teist, kolmandat või kõrgemat järku polünoomi. Kuid polünoomi kasutamisel prognoosimiseks peab olema väga ettevaatlik, sest pikema prognoosi korral võime sattuda teisele poole lokaalset ekstreemumit ning prognoos näitab, et kasvamise asemel algab kahanemine (või vastupidi), kuigi mingit majanduslikku põhjust selliseks muutuseks pole ette näha.

Vaatame ka üht näidet, kus aegrea silumiseks kasutatava matemaatilise mudeli kuju on tuletatud teoreetilistest kaalutlustest lähtudes.

Näide 10.12. Bassi difusioon ja uue toote leviku prognoosimine

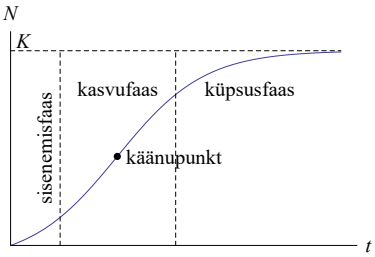
Uue toote või teenuse levimist kirjeldab tüüpiliselt S-kujuline kõver, millel eristatakse kolme faasi. Turule sisenemise faasis (*introduction phase*) on kasv aeglane, seejärel saabub kasvufaas (*growth phase*), kus müügimaht kasvab kiiresti ning lõpuks jõuab kätte n-ö küpsusfaas (*maturity phase*), millal läbimüügi kasv jätkub, ent kahaneva kiirusega (joonis 10.14). Sellist kõverat iseloomustab turu mahutavus ehk küllastuvus K — kui palju on maksimaalselt võimalik seda toodet müüa. Käänupunktis on läbimüügi kasv kõige kiirem ja selleks ajaks on saavutatud umbes pool turu mahutavusest.

Kõvera selline kuju on seletatav difusiooniteooria abil. Difusiooniteooria põhialuseks on tarbijate jagamine kahte rühma: innovaatorid ja matkijad. Innovaatorid ei ole mõjutatud sellest, kui palju see uus toode on juba tarbijate seas levinud. Matkijad ehk imitaatorid on aga mõjutatud sellest, kui paljud inimesed on vastavat toodet juba ostnud või tarbivad vastavat teenust. Matemaatiliselt on difusiooni käsitletud mitut moodi, erinevad

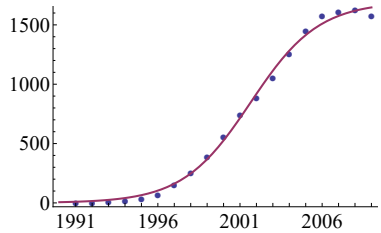
lähenemised annavad veidi erinevad mudelid. Tuntuim neist on Bassi difusioonimudel (Bass, 1969):

$$N(t) = K\frac{1 - e^{-(p+q)t}}{1 + \frac{q}{p}e^{-(p+q)t}}, \tag{10.34}$$

kus N on toote omandanute arv (kogumüük) ajahetkel t, K turu mahutavus, p innovatsioonikoefitsient ja q imitatsioonikoefitsient. Innovatsioonikoefitsient p iseloomustab tõenäosust, kas innovatiivne tarbija ostab selle toote tänu reklaamile või muudele välistele teguritele. Imitatsioonikoefitsient q iseloomustab aga tõenäosust, et matkija soetab selle toote, kuna teised on selle juba soetanud.
Mudeli (10.34) parameetrite hindamiseks kasutatakse mittelineaarset vähimruutude meetodit.



Joonis 10.14. S-kujuline kasvumudel

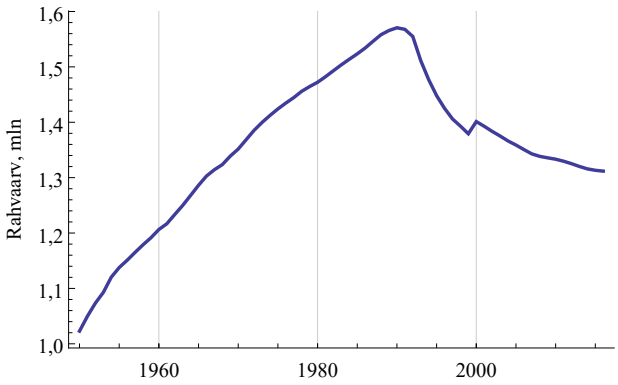


Joonis 10.15. Mobiiltelefonivõrgu kasutajad Eestis, tuh. Punktid tähistavad empiirilisi andmeid, joon mudelit

Bassi difusioonmudelit (10.34) on kasutatud Eesti mobiiltelefonivõrgu kasutajate arvu dünaamika modelleerimiseks aastatel 1991–2009 (joonis 10.15). Parameetrite hindamisel saadi, et innovatsioonikoefitsient $p = 0{,}0017$, imitatsioonikoefitsient $q = 0{,}482$ ja turu mahutavus on 1,69 miljonit (Sauga, 2011).

Pikema aegrea korral võivad majanduses või ühiskonnas toimuda suured muutused, mis mõjutavad aegrida genereerivaid protsesse. Sellisel juhul võib trend muutuda ning aegrida jagatakse perioodideks, mille jaoks leitakse erinevad trendimudelid. Joonisel 10.16 on Eesti rahvaarv aastatel 1950–2016, miljonites[4]. Hüpe aastal 2000 on tingitud sellest, et aastate 2000–2013 andmed on 17.01.2014 ümber arvutatud. See aegrida tuleb jaotada kaheks või enamaks perioodiks ja leida trend iga perioodi jaoks eraldi.

[4]Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel RV012: rahvastik soo ja vanuserühma järgi, 1. jaanuar.



Joonis 10.16. Eesti rahvaarv 1950–2016

10.8. Näide: erinevad meetodid müügimahu prognoosimisel

N10Aegread
P10.8

Elektroonikapoe müügijuht soovib prognoosida teatud tüüpi telerite müüki järgmisel kuul. Jaanuarist kuni juulini müüdi telereid vastavalt 180, 215, 225, 220, 210, 205 ja 215 tk kuus. Sobiva prognoosimismeetodi leidmiseks võrdleb ta erinevaid silumismeetodeid: libisev keskmine, kaalutud libisev keskmine, eksponentsilumine ja lineaarne regressioon. Esimese kolme meetodi korral kasutab müügijuht ühesammulist prognoosimist: perioodi t prognoos on eelmise perioodi silutud väärtus. Kuna tegelikud müügimahud on kuni juulini teada, saab neid prognoosidega võrrelda ja leida prognoosivead. Erinevatel meetoditel saadud prognooside võrdlemiseks kasutab ta prognoosivigade ruutude aritmeetilist keskmist, mida nimetatakse **keskmiseks ruutveaks** (*Mean Square Error*, MSE)

$$MSE = \frac{1}{n}\sum_{t=1}^{n} u_t^2, \tag{10.35}$$

Keskmine ruutviga

kus prognoosiviga $u_t = y_t - F_t$.

1. Lihtne libisev keskmine. Müügijuht kasutab kolmesammulist libisevat keskmist. Aprilli prognoos on märtsikuu silutud väärtus:

$$F_4 = MA_3 = \frac{1}{3}(180 + 215 + 225) \approx 206{,}67.$$

Samamoodi leiab ta prognoosid kuni augustikuuni (tabel 10.5).

Keskmine ruutviga perioodide 4 kuni 7 prognoosivigade alusel:

$$MSE = \frac{1}{4}\left(13{,}33^2 + (-10{,}00)^2 + (-13{,}33)^2 + 3{,}33^2\right) \approx 116{,}67.$$

Tabel 10.5. Prognoosimine libiseva keskmise abil

Kuu	Müük	Silutud väärtus	Prognoos	Prognoosiviga
1	180			
2	215			
3	225	206,67		
4	220	220,00	206,67	13,33
5	210	218,33	220,00	−10,00
6	205	211,67	218,33	−13,33
7	215	210,00	211,67	3,33
8			210,00	

2. Kaalutud libisev keskmine. Müügijuht kasutab kolmesammulist kaalutud libisevat keskmist ja kaalud võtab ta järgmised:

$$w_t = 0{,}5, \quad w_{t-1} = 0{,}3, \quad w_{t-2} = 0{,}2.$$

Aprilli prognoos on märtsikuu silutud väärtus, mis leitakse kaalutud libiseva keskmise valemist (10.19):

$$F_4 = WMA_3 = \frac{0{,}5 \cdot 225 + 0{,}3 \cdot 215 + 0{,}2 \cdot 180}{0{,}5 + 0{,}3 + 0{,}2} = 213.$$

Kaalutud libiseva keskmise abil saadud prognoosid ja prognoosivead on esitatud tabelis 10.6.

Tabel 10.6. Prognoosimine kaalutud libiseva keskmise abil

Kuu	Müük	Silutud väärtus	Prognoos	Prognoosi viga
1	180			
2	215			
3	225	213,0		
4	220	220,5	213,0	7,0
5	210	216,0	220,5	−10,5
6	205	209,5	216,0	−11,0
7	215	211,0	209,5	5,5
8			211,0	

Keskmine ruutviga perioodide 4 kuni 7 prognoosivigade alusel:

$$MSE = \frac{1}{4}\left(7^2 + (-10{,}5)^2 + (-11)^2 + 5{,}5^2\right) \approx 77{,}63.$$

See on väiksem kui lihtsa libiseva keskmise MA keskmine ruutviga. Järelikult annab kaalutud libiseva keskmise kasutamine prognoosimisel parema tulemuse.

3. Eksponentsilumine. Eksponentsilumise konstandiks võtab müügijuht 0,6. Jaanuarikuu silutud väärtus võrdub tegeliku väärtusega ja see on ka veebruarikuu prognoos. Märtsikuu prognoos on veebruarikuu silutud väärtus:

$$F_3 = E_2 = 0{,}6 \cdot 215 + 0{,}4 \cdot 180 = 201.$$

Aprilli prognoos on märtsi silutud väärtus:

$$F_4 = E_3 = 0{,}6 \cdot 225 + 0{,}4 \cdot 201 = 215{,}4.$$

Samamoodi leiab müügijuht prognoosid kuni augustikuuni (tabel 10.7).

Tabel 10.7. Prognoosimine eksponentsilumise abil

Kuu	Müük	Silutud väärtus	Prognoos	Prognoosi viga
1	180	180,00		
2	215	201,00	180,00	35,00
3	225	215,40	201,00	24,00
4	220	218,16	215,40	4,60
5	210	213,26	218,16	−8,16
6	205	208,31	213,26	−8,26
7	215	212,32	208,31	6,69
8			212,32	

Kuigi eksponentsilumise korral on prognoos ja ka prognoosiviga leitud juba alates veebruarist, siis keskmine ruutviga leitakse ainult perioodide 4 kuni 7 prognoosivigade alusel. Siis on see võrreldav eelmiste meetodite abil saadud prognoosidega.

$$MSE = \frac{1}{4}\left(4{,}6^2 + (-8{,}16)^2 + (-8{,}26)^2 + 6{,}69^2\right) \approx 50{,}21.$$

Näeme, et eksponentsilumise kasutamine annab veel paremad prognoosid kui kaalutud libisev keskmine.

4. Lineaarne regressioon. Kasutades müügimahtusid jaanuarist kuni juulini, leiab müügijuht lineaarse regressioonmudeli ning prognoos ajahetkel t on leitav sellest mudelist

$$\hat{y}_t = 200 + 2{,}5t.$$

Näiteks jaanuarikuu ja veebruarikuu prognoosid on vastavalt

$$\hat{y}_1 = 200 + 2{,}5 \cdot 1 = 202{,}5, \ \hat{y}_2 = 200 + 2{,}5 \cdot 2 = 205.$$

Et võrrelda teistel meetoditel saadud prognoosidega, tuleb keskmise ruutvea arvutamisel kasutada prognoose aprillist kuni juulini:

$$MSE = \frac{1}{4}\left(10^2 + (-2{,}5)^2 + (-10)^2 + (-2{,}5)^2\right) \approx 53{,}13.$$

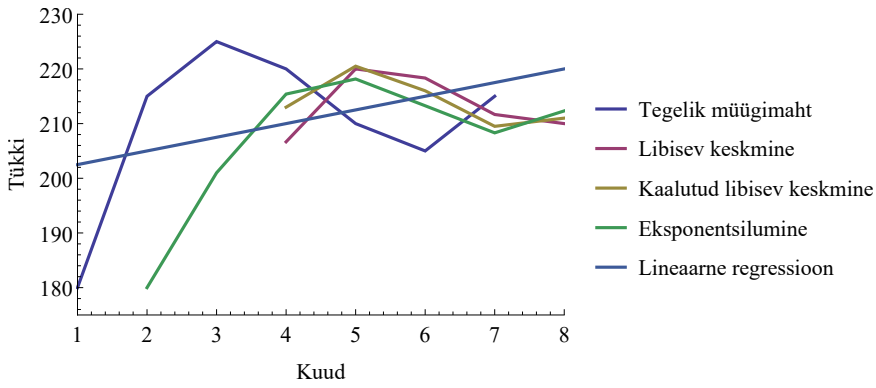
Tabel 10.8. Prognoosimine lineaarse regressioonmudeli abil

Kuu	Müük	Prognoos	Prognoosi viga
1	180	202,5	−22,5
2	215	205,0	10,0
3	225	207,5	17,5
4	220	210,0	10,0
5	210	212,5	−2,5
6	205	215,0	−10,0
7	215	217,5	−2,5
8		220,0	

Kokkuvõte erinevate prognoosimismeetodite kasutamisest ja müügimahu prognoos augustikuuks (ümardatud) on esitatud tabelis 10.9. Kuna eksponentsilumise abil tehtud prognoosil on kõige väiksem keskmine ruutviga MSE, otsustab müügijuht augustikuu müügimahu prognoosimiseks kasutada seda meetodit ning augustikuu prognoositav müük on 212 tk. Joonisel 10.17 on tegelik müügimaht ja erinevatel meetoditel saadud prognoosid.

Tabel 10.9. Kokkuvõte erinevate prognoosimismeetodite kasutamisest

Prognoosimismeetod	MSE	Prognoos augustiks
Libisev keskmine	116,67	210
Kaalutud libisev keskmine	77,63	211
Eksponentsilumine	50,21	212
Lineaarne regressioonmudel	53,13	220



Joonis 10.17. Müügimahu prognoosimine erinevatel meetoditel

10.9. Aegridade kompleksanalüüs

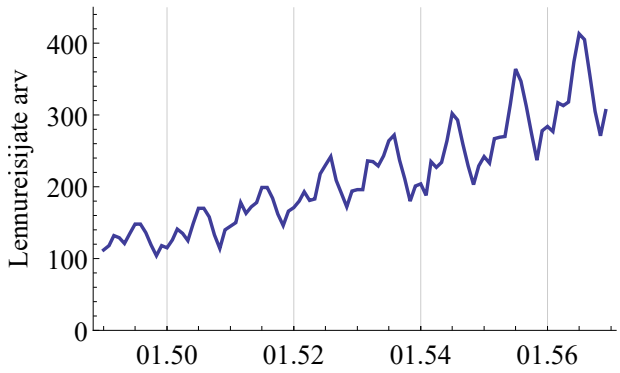
Aegridade kompleksanalüüsi korral jaotatakse ajas muutuva suuruse muutumine mitmeks komponendiks. Süstemaatiliselt muutuvaid komponente on kolm.

Aegrea komponendid

- **Trend** T_i on suuruse väärtuses pikema aja jooksul ilmnev tendents.
- **Sesoonsed muutused** S_i on seaduspärased perioodilised muutused trendi ümbruses. Periood on tavaliselt aasta või lühem. Näiteks hooajaliste kaupade käive sõltub aastaajast, keskmine brutopalk on tavaliselt suurem juunis ja detsembris jms.
- **Tsüklilised muutused** C_i on aastast pikema perioodiga toimuvad perioodilised muutused trendi ümbruses. Näiteks Inglismaa majanduses esines IX sajandil 9 aasta pikkune aktiivsuse muutumise tsükkel. Ka tänapäeva majanduses on tuntud majanduskasvu perioodilised muutused, mida nimetatakse majandustsükliteks (*business cycle*).

Juhuslik komponent ε_i on põhjustatud paljude tegurite koosmõjust vaadeldavale suurusele ja pole prognoositav.

Joonisel 10.18 toodud aegreal esineb nii trend kui ka sesoonne komponent. Erinevate aegridade korral võib üks või teine komponent puududa. Näiteks intressimäärad, vahetuskursid, nafta hind maailmaturul aja jooksul küll muutuvad, kuid nende dünaamikas üldreeglina puudub pikaajaline trend. Selliseid suurusi, mille aegread ei sisalda kindlaid trende, vaid kõiguvad keskmise taseme ümber, nimetatakse **statsionaarseteks** suurusteks. Suurused, mille aegread sisaldavad pikaajalisi trende, on **mittestatsionaarsed** suurused. Mittestatsionaarsed on näiteks riigi sisemajanduse koguprodukt (SKP), tarbijahinnaindeks, reaalpalk, rahvaarv.



Joonis 10.18. Lennureisijate arv USA rahvusvahelistel lennuliinidel 1949–1956, kuised andmed (Makridakis, Wheelwright ja J. R. Hyndman, 1998). Esineb trend ja sesoonne komponent

Sesoonsust võib analüüsida kahel erineval eesmärgil:

- sesoonsuse silumine eesmärgiga muuta võrreldavaks erinevate perioodide andmed (joonised 10.7, 10.8);
- sesoonsuse modelleerimine prognoosimise eesmärgil (näited 10.13, 10.14, 10.16).

Komponentide eraldamiseks kasutatakse erinevaid mudeleid:

- **aditiivne** mudel, mille korral tunnuse väärtus avaldatakse üksikute komponentide summana;
- **multiplikatiivne** mudel, mille korral tunnuse väärtus avaldatakse üksikute komponentide korrutisena.

Sesoonsuse modelleerimiseks kasutatakse ka fiktiivsete tunnuste lisamist regressioonmudelisse (vt alapeatükk 9.18). Kvartaalse sesoonsuse korral on vaja kolme fiktiivset tunnust, kuise sesoonsuse korral 11 fiktiivset tunnust.

10.10. Aditiivne mudel

Mõiste aditsioon tuleneb ladinakeelsest sõnast *addere*, mis tähendab lisama. Aditiivne tähendab liitmisel põhinevat ning aditiivse mudeli korral aegrea komponendid liidetakse.

Aditiivse mudeli korral vaadeldakse aegrida üksikute komponentide summana:

$$y_t = T_t + C_t + S_t + \varepsilon_t, \tag{10.36}$$

kus T_t on trend, C_t tsükliline komponent, S_t sesoonne komponent ja ε_t juhuslik komponent.

Seda mudelit on sobiv kasutada, kui absoluutne kõrvalekalle trendist on erinevatel perioodidel ligikaudu ühesugune (joonis 10.19):

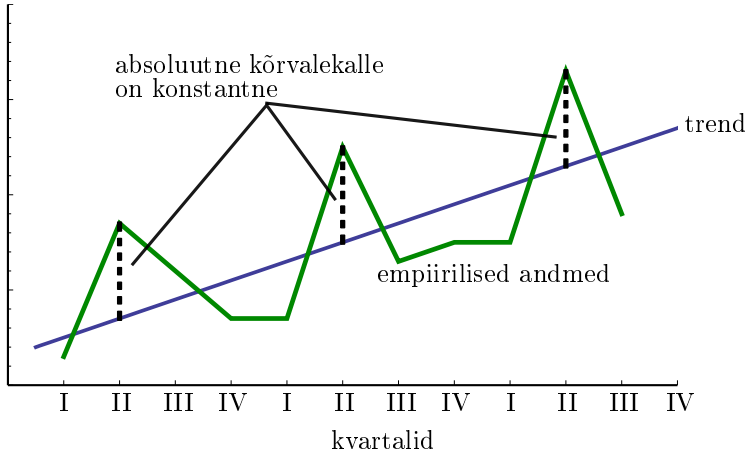
$$y_t - T_t \approx const. \tag{10.37}$$

Aditiivse mudeli üksikute komponentide eraldamiseks on mitu võimalust. Üks sagedasti kasutatav meetod on fiktiivsete tunnustega (*dummy*) regressioonmudeli kasutamine. Siin tutvume ka teise võimalusega. Lähemalt vaatame ainult sesoonse komponendiga aegrea analüüsi, s.t

$$y_t = T_t + S_t + \varepsilon_t. \tag{10.38}$$

Mudeli koostamiseks läbitakse järgmised sammud:

1) regressioonanalüüsi abil leitakse trendifunktsioon;
2) trendifunktsiooni kasutades leitakse trendi väärtused T_i;



Joonis 10.19. Sesoonsed kõrvalkalded trendist on konstantsed

3) leitakse sesoonsed komponendid $S_t = y_t - T_t$;
4) leitakse keskmised sesoonsed komponendid:
   - kuise aegrea korral kõikide aastate jaanuarikuude sesoonsete komponentide aritmeetiline keskmine, veebruarikuude sesoonsete komponentide aritmeetiline keskmine jne. Kokku saadakse 12 keskmist sesoonset komponenti, iga kuu jaoks üks;
   - kvartaalse aegrea korral kõikide aastate esimeste kvartalite sesoonsete komponentide aritmeetiline keskmine, teiste kvartalite sesoonsete komponentide aritmeetiline keskmine jne. Saadakse neli keskmist sesoonset komponenti, iga kvartali jaoks üks.

Sellega ongi komponendid eraldatud: trend ja keskmised sesoonsed komponendid. Vajadusel järgneb prognoos. Prognoosimiseks

1) leitakse trendi T_t väärtus prognoositava ajaperioodi t jaoks;
2) prognoosi saamiseks liidetakse vastava perioodi trendi väärtusele keskmine sesoonne komponent: $F_t = T_t + \bar{S}_t$.

Näide 10.13. Ettevõtte käive ja aditiivse mudeli kasutamine


N10Aegread
N10.13

Tabelis on toodud ühe ettevõtte käive (tuhat eurot), kvartalite kaupa. Vaatluse all on kolme aasta andmed.

1. Trendi leidmiseks silutakse aegrida lineaarse regressioonjoonega, mudeliks on

$$T_t = 5{,}89t + 15{,}62, \tag{10.39}$$

kus t on aeg kvartalites (joonis 10.20 (a)).

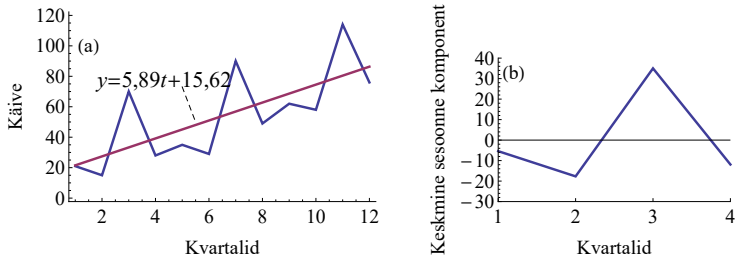
2. Trendi väärtuste leidmiseks pannakse valemisse (10.39) järjest aja t väärtused $1, 2, \ldots$.
3. Vaatlusandmete ja trendi vahe on sesoonne komponent S_t.

Aeg t kvartalites	Käive y_t	Trend T_t	Sesoonne komponent $S_t = y_t - T_t$
1	21	21,5	−0,5
2	15	27,4	−12,4
3	70	33,3	36,7
4	28	39,2	−11,2
5	35	45,1	−10,1
6	29	51,0	−22,0
7	90	56,9	33,1
8	49	62,8	−13,8
9	62	68,6	−5,8
10	58	74,5	−18,6
11	114	80,4	35,1
12	76	86,3	−10,8

Sesoonsed komponendid grupeeritakse kvartalite kaupa ja leitakse iga kvartali jaoks sesoonsete komponentide aritmeetiline keskmine. Paneme tähele, et keskmiste sesoonsete komponentide summa on 0.

	Sesoonsed komponendid			Keskmine sesoonne
Kvartal	1. aasta	2. aasta	3. aasta	komponent
1	−0,5	−10,1	−5,8	−5,4
2	−12,4	−22,0	−18,6	−17,7
3	36,7	33,1	35,1	35,0
4	−11,2	−13,8	−10,8	−11,9

Keskmise sesoonse komponendi võib esitada eraldi graafikul (joonis 10.20 (b)).



Joonis 10.20. (a) Käive (tuhat eurot) ja trend. (b) Keskmine sesoonne komponent

Esimeses kvartalis on käive keskmiselt 5,4 tuhande euro võrra trendist väiksem. 3. kvartalis on käive keskmiselt 35 tuhande euro võrra trendist suurem.
Neljanda aasta 1. kvartali väärtuse prognoosimiseks leiame kõigepealt trendi väärtuse valemist (10.39), kui $t = 13$:

$$T_{13} = 5{,}89 \cdot 13 + 15{,}62 \approx 92{,}2.$$

Sellele liidame aasta esimese kvartali keskmise sesoonse komponendi ja saame prognoositud väärtuse:

$$F_{13} = 92{,}2 + (-5{,}4) = 86{,}8.$$

Nii saame leida prognoositud väärtused kõigi järgnevate kvartalite jaoks.

Aeg t, kvartalites	Trend T	Keskmine sesoonne komponent $\bar{S}$	Prognoos $F = T + \bar{S}$
13	92,2	−5,4	86,8
14	98,1	−17,7	80,5
15	104,0	35,0	139,0
16	109,9	−11,9	98,0

Aditiivne mudel sobib aegrea modelleerimiseks hästi, kui keskmiste sesoonsete komponentide summa on ligikaudu null. Keskmiste sesoonsete komponentide leidmiseks peab kõiki ühesuguse sesoonsusega perioode olema vähemalt kolm. Järelikult peaks kuise või kvartaalse sesoonsuse modelleerimiseks olema aegrida vähemalt kolme aasta pikkune.

10.11. Multiplikatiivne mudel

Mõiste multiplikatsioon (ladina *multiplicātio*) tähendab matemaatikas korrutamist. Multiplikatiivse mudeli korral korrutatakse trendi väärtus tsüklilisele, sesoonsele ja juhuslikule komponendile vastavate kordajatega.

Multiplikatiivse mudeli korral vaadeldakse aegrida üksikute komponentide korrutisena:

$$y_t = T_t \cdot C_t \cdot S_t \cdot \varepsilon_t. \tag{10.40}$$

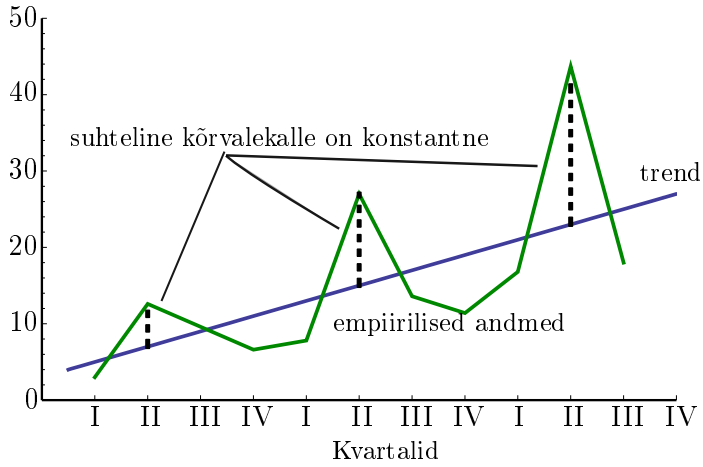
kus T_t on trend, C_t – tsükliline komponent, S_t – sesoonne komponent ja ε_t – juhuslik komponent.

Kui tsükliline või sesoonne komponent puudub, siis vastav kordaja võrdub ühega.

Multiplikatiivset mudelit on sobiv kasutada, kui tunnuse väärtuste y_t ja trendi T_t suhe on erinevatel perioodidel ligikaudu konstantne:

$$\frac{y_t}{T_t} \approx const. \tag{10.41}$$

Sellisel juhul on suhtelised erinevused trendist ühesugused (joonis 10.21).



Joonis 10.21. Sesoonsete kõrvalekallete suhteline erinevus trendist on konstantne

Lähemalt vaatame ainult trendi ja sesoonse komponendiga aegrea analüüsi:

$$y_t = T_t \cdot S_t \cdot \varepsilon_t. \tag{10.42}$$

Märgime, et trendi ühik on sama, mis analüüsitaval suurusel, ja sesoonne ning juhuslik komponent on ühikuta suurused.

Mudeli koostamiseks on vaja läbida järgmised sammud:

1) regressioonanalüüsi abil leitakse sobiv trendifunktsioon;
2) trendifunktsiooni kasutades leitakse trendi väärtused T_t;
3) leitakse sesoonsed komponendid $S_t = y_t/T_t$;

4) leitakse keskmised sesoonsed komponendid:
   - kuise aegrea korral jaanuari keskmine sesoonne komponent on kõikide aastate jaanuarikuude sesoonsete komponentide **geomeetriline** keskmine, veebruarikuu keskmine sesoonne komponent on geomeetriline keskmine kõikide veebruarikuude sesoonsetest komponentidest jne. Kokku saadakse 12 keskmist sesoonset komponenti, iga kuu jaoks üks;
   - kvartaalse aegrea korral kõikide esimeste kvartalite sesoonsete komponentide geomeetriline keskmine, teiste kvartalite sesoonsete komponentide geomeetriline keskmine jne. Saadakse 4 keskmist sesoonset komponenti, iga kvartali jaoks üks.

Sellega on komponendid eraldatud: trend ja keskmised sesoonsed komponendid. Vajadusel järgneb prognoos. Prognoosimiseks:

1) leitakse trendi T_t väärtus prognoositava ajaperioodi t jaoks;
2) trendi väärtust T_t korrutatakse vastava perioodi keskmise sesoonse komponendiga: $F_t = T_t \cdot S_t$.

Näide 10.14. Lennureisijate arvu dünaamika ja multiplikatiivne mudel


N10Aegread
N10.14

Leiame mudeli joonisel 10.18 toodud lennureisijate arvu dünaamika modelleerimiseks. Modelleerimiseks kasutame andmeid kuni 1955. aastani. Sellisel juhul on meil võimalus mudeli abil leida prognoos 1956. aasta jaoks ning võrrelda seda 1956. aasta tegelike andmetega.

Paneme jooniselt 10.18 tähele, et on olemas kasvutrend ja selle ümber muutused, mis korduvad perioodiliselt igal aastal. Maksimumid esinevad iga aasta juulis. Kuna siin kõikumine ümber trendi pidevalt suureneb, on sobiv kirjeldada seda aegrida multiplikatiivse mudeliga.

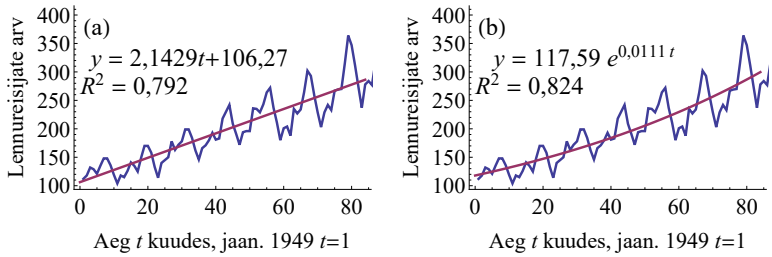
Esimese sammuna määrame kindlaks trendi. Visuaalse hinnangu põhjal võiks kasutada kas lineaarset või eksponentsiaalset regressioonmudelit. Leiame mõlemad mudelid ja võrdleme, kumb sobib paremini. Mudeli sobivuse hindamiseks kasutame determinatsioonikordajat R^2. Joonistelt 10.22 on näha, et eksponentsiaalne mudel sobib paremini, sest determinatsioonikordaja R^2 on suurem.

Arvutame nüüd trendi komponendi T_t ja sesoonse komponendi S_t iga kuu jaoks. Trendi väärtuste leidmiseks kasutame leitud eksponentsiaalset mudelit

$$T_t = 117{,}95e^{0{,}0111t}. \qquad (10.43)$$

Näiteks trendi väärtus $t = 1$ korral

$$T_1 = 117{,}95e^{0{,}0111 \cdot 1} \approx 119{,}27.$$



Joonis 10.22. Lennureisijate arvu muutumise trendi leidmiseks on kasutatud kaht erinevat mudelit: (a) lineaarne mudel, (b) eksponentsiaalne mudel

Sesoonsed komponendid leiame valemist

$$S_t = \frac{y_t}{T_t}.$$

Ruumi kokkuhoiu mõttes esitame tabelis 10.10 tulemused vaid esimese nelja kuu jaoks. Trendi ja sesoonse komponendi väärtused on ümardatud.

Tabel 10.10. Trendi ja sesoonse komponendi leidmine

Kuud	t, alates jaan 1949	Vaatlusandmed y_t	Trend T_t	Sesoonne komponent $S_t = y_t/T_t$
jaan 49	1	112	119,27	0,94
veebr 49	2	118	120,60	0,98
märts 49	3	132	121,94	1,08
apr 49	4	129	123,30	1,05
...	...	...	...	...

Keskmise sesoonse muutuse leidmiseks grupeerime sesoonsed komponendid kuude kaupa ja leiame iga kuu korral geomeetrilise keskmise. Näiteks jaanuarikuu keskmine sesoonne komponent on

$$\bar{S}_1 = \sqrt[7]{0{,}94 \cdot 0{,}84 \cdot 0{,}93 \cdot 0{,}96 \cdot 0{,}96 \cdot 0{,}88 \cdot 0{,}91} \approx 0{,}92.$$

Jaanuari, veebruari, märtsi ja aprilli keskmine sesoonne komponent on esitatud tabelis 10.11.

Tabel 10.11. Keskmiste sesoonsete komponentide leidmine

Kuud	1949	1950	1951	1952	1953	1954	1955	Keskmine sesoonne komponent $\bar{S}_t$
1	0,94	0,84	0,93	0,96	0,97	0,88	0,91	0,92
2	0,98	0,91	0,95	1,00	0,95	0,80	0,87	0,92
3	1,08	1,01	1,12	1,06	1,14	0,99	0,99	1,05
4	1,05	0,96	1,01	0,98	1,12	0,95	0,98	1,01
...	...	...	...	...	...	...	...	...

Saadud tulemusi kasutame 1956. aasta väärtuste prognoosimiseks. Prognoosi F_t arvutamiseks leiame algul eksponentsiaalse mudeli (10.43) põhjal trendi väärtused T_t ja korrutame need keskmiste sesoonsete komponentidega $\bar{S}_t$:

$$F_t = T_t \bar{S}_t.$$

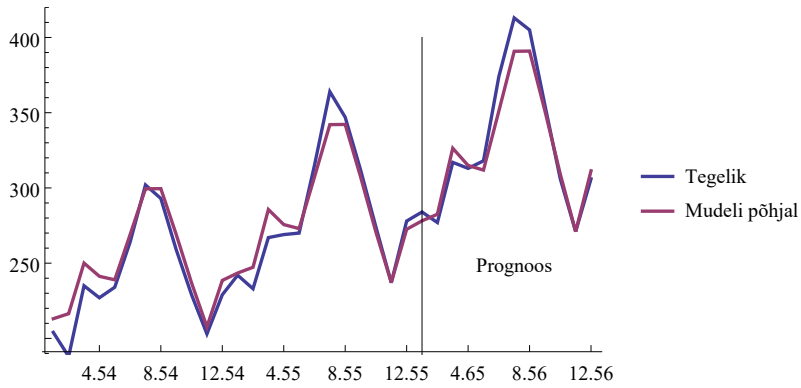
Kuna on olemas ka vaatlusandmed 1956. aasta kohta, saame võrrelda prognoositud väärtusi tegelike väärtustega. Võrdlemiseks on leitud suhtelised vead:

$$E_t = \frac{y_t - F_t}{y_t} \tag{10.44}$$

Tabel 10.12. Prognoosi leidmine 1956. aastaks

1956. a kuud	t	Keskmine sesoonne komponent $\bar{S}_t$	Trend T_t	Prognoos F_t	Tegelik väärtus y_t	Suhteline viga E_t, %
1	85	0,92	303,0	278	284	2,1
2	86	0,92	306,4	283	277	−2,0
3	87	1,05	309,8	326	317	−3,0
4	88	1,01	313,3	315	313	−0,6
...	...	...	...	...	...	...

Joonisel 10.23 on esitatud nii tegelikud väärtused kui ka mudeli põhjal leitud väärtused. Aastad 1954–55 on valimisisene võrdlus. Aasta 1956 on aga prognoos valimist välja, sest mudeli leidmisel me 1956. aasta andmeid ei kasutanud.



Joonis 10.23. Lennureisijate arvu muutumine ja mudeli põhjal leitud väärtused. Prognoos on tehtud aastaks 1956

Multiplikatiivne mudel sobib hästi, kui keskmiste sesoonsete komponentide korrutis on ligikaudu üks. Näites 10.14 tuli see 0,993.

Nagu aditiivset mudelit, saab ka multiplikatiivset mudelit kasutada siis, kui meil iga keskmise sesoonse komponendi leidmiseks on vähemalt kolm erinevat väärtust. Kui soovime modelleerida kvartaalset või kuist sesoonsust, peab kasutatav aegrida olema vähemalt kolme aasta pikkune.

10.12. Trendi ja sesoonsusega eksponentsilumine

Eksponentsilumisel, mida vaatlesime alapeatükis 10.6, ei arvestatud trendi ega sesoonseid muutusi. Kui aga aegrida neid sisaldab, siis tuleks need silumisel ka arvesse võtta. See võimaldab saada täpsemaid prognoose. Eksponentsilumist koos trendi ja sesoonsuse arvestamisega nimetatakse ka **Holti-Wintersi mudeliks**.

Topelt eksponentsilumine

Topelt eksponentsilumise (*double exponential smoothing*) korral võetakse arvesse ainult trend T. Eksponentsiaalselt silutud väärtus E_t ja trend T_t arvutatakse kumbki eraldi:

$$E_t = wy_t + (1-w)(E_{t-1} + T_{t-1}), \tag{10.45}$$
$$T_t = v(E_t - E_{t-1}) + (1-v)T_{t-1}. \tag{10.46}$$

Selline lähenemine nõuab juba kaht tasandusparameetrit w ja v, mõlema väärtused peavad olema 0 ja 1 vahel.

Paneme tähele, et trendi komponendi hindamine on **adaptiivne**: trendi hinnang T_t on kaalutud aritmeetiline keskmine eksponentsiaalselt silutud väärtuse viimasest muutusest $E_t - E_{t-1}$ ja eelmisest trendi hinnangust T_{t-1}.

Prognoos ajahetkeks $t+1$ on viimase silutud komponendi E_t ja trendi T_t summa:

$$F_{t+1} = E_t + T_t. \tag{10.47}$$

Pikema prognoosi korral eeldatakse, et trend on lineaarne ja trendi väärtus T_t korrutatakse läbi prognoositavate perioodide arvuga k:

$$F_{t+2} = E_t + 2T_t, \tag{10.48}$$
$$F_{t+k} = E_t + kT_t. \tag{10.49}$$

Näide 10.15. Ettevõtte netokäive aastas


N10Aegread
N10.15

Kasutame trendiga eksponentsilumist ettevõtte netokäibe jaoks ja teeme prognoosi kolm aastat edasi. Silumiskonstantide väärtusteks võtame $w = 0{,}7$ ja $v = 0{,}5$.
Alustada tuleb ajaperioodist $t = 2$, sest selle jaoks saame leida trendi:

$$E_2 = y_2 = 4{,}0$$
$$T_2 = y_2 - y_1 = 4{,}0 - 4{,}8 = -0{,}8.$$

Aeg t aastates	Netokäive, tuh €	E_t	T_t
1	4,8		
2	4,0	4,0	−0,8
3	5,5	4,8	0,005
4	15,6	12,4	3,8
5	23,1	21,0	6,2
...	...	...	...
35	150,9	152,9	2,7

Ajaperioodi $t = 3$ jaoks kasutame juba valemeid (10.45) ja (10.46):

$$E_3 = wy_3 + (1-w)(E_2 + T_2) = 0{,}7 \cdot 5{,}5 + 0{,}3 \cdot (4{,}0 - 0{,}8) \approx 4{,}81,$$
$$T_3 = v(E_3 - E_2) + (1-v)T_2 = 0{,}5 \cdot (4{,}81 - 4{,}0) + 0{,}5 \cdot (-0{,}8) \approx$$
$$\approx 0{,}005.$$

Nii jätkame aegrea lõpuni.
Käibe prognoosi leidmiseks kasutame valemeid (10.47)–(10.49).
Prognoos järgmiseks kolmeks aastaks:

Aeg t aastates	k	F_{t+k}
36	1	155,6
37	2	158,4
38	3	161,1

Trendiga eksponentsilumine kokkuvõtlikult

1. Valida eksponentsilumise konstant w ($0 \leq w \leq 1$). Väiksema w korral on aegrea viimane väärtus väiksema kaaluga, suurema w korral arvestatakse viimast väärtust rohkem.
2. Valida trendi silumiskonstandi v väärtus ($0 \leq v \leq 1$). Väiksema v korral on hilisem trendi väärtus väiksema kaaluga, suurema v korral arvestatakse hilisemaid trendi väärtusi rohkem.
3. Leida komponendid E_t ja T_t, alustades ajahetkest $t = 2$ (E_1 ja T_1 ei ole defineeritud).

$$
\begin{aligned}
E_2 &= y_2, \\
T_2 &= y_2 - y_1, \\
E_3 &= wy_3 + (1-w)(E_2 + T_2), \\
T_3 &= v(E_3 - E_2) + (1-v)T_2, \\
&\vdots
\end{aligned}
$$

$$E_t = wy_t + (1-w)(E_{t-1} + T_{t-1}),$$
$$T_t = v(E_t - E_{t-1}) + (1-v)T_{t-1}.$$

4. Prognoosimine:

$$F_{t+1} = E_t + T_t,$$
$$F_{t+2} = E_t + 2T_t,$$
$$F_{t+k} = E_t + kT_t.$$

Kui aegreal esineb lisaks trendile ka **sesoonne komponent**, kasutatakse nn kolmekordset eksponentsiaalset silumist (*triple exponential smoothing*). Eraldi leitakse silutud väärtus E_t, trendi komponent T_t ja sesoonne komponent S_t. Eesmärgiks on arvestada nii trendi kui ka sesoonsust, mis mõlemad võivad aja jooksul muutuda.

Kui sesoonne kõikumine trendi ümber on ligikaudu konstantse amplituudiga, kasutatakse **aditiivset mudelit**, mille korral liidetakse sesoonne komponent silutud väärtusele E_t+S_t. Silutud väärtuse leidmisel asendatakse valemis (10.45) y_t väärtus vahega $y_t - S_{t-p}$, kus p on sesoonse tsükli pikkus:

$$E_t = w(y_t - S_{t-p}) + (1-w)(E_{t-1} + T_{t-1}). \tag{10.50}$$

Näiteks kui aegrida on esitatud kuude kaupa ja esineb aastane sesoonsus, siis $p = 12$. Kui aegrida on esitatud kvartalite kaupa, siis $p = 4$. Nädalase sesoonsuse korral on aegrida esitatud enamasti päevade kaupa, siis $p = 7$. Trendi komponent leitakse analoogselt eelmise mudeliga:

$$T_t = v(E_t - E_{t-1}) + (1-v)T_{t-1}. \tag{10.51}$$

Sesoonne komponent S_t on kaalutud aritmeetiline keskmine vahest $y_t - E_t$ ja eelmise tsükli vastavast sesoonsest komponendist S_{t-p}:

$$S_t = \alpha(y_t - E_t) + (1-\alpha)S_{t-p}, \tag{10.52}$$

kus α on kolmas silumiskonstant, $0 \leq \alpha \leq 1$. Suurema α korral on suurem kaal kõige hilisemal sesoonsel komponendil, väiksema α korral saavad suurema kaalu varasemalt hinnatud sesoonsed komponendid.

Kui sesoonse kõikumise amplituud on ligikaudu võrdeline aegrea keskmise tasemega, kasutatakse **multiplikatiivset mudelit**, mille korral sesoonne komponent korrutatakse silutud väärtusega: $E_t \cdot S_t$. Sellisel juhul on sesoonne komponent ühikuta kordaja, nii nagu regressioonanalüüsiga leitud trendi ja multiplikatiivse sesoonsuse korral. Silutud väärtuse leidmisel asendatakse valemis (10.45) y_t

väärtus jagatisega y_t/S_{t-p}, kus p on sesoonse tsükli pikkus:

$$E_t = w\frac{y_t}{S_{t-p}} + (1-w)(E_{t-1} + T_{t-1}), \tag{10.53}$$

$$T_t = v(E_t - E_{t-1}) + (1-v)T_{t-1}, \tag{10.54}$$

$$S_t = \alpha\frac{y_t}{E_t} + (1-\alpha)S_{t-p}. \tag{10.55}$$

Trendi ja sesoonsusega eksponentsilumine kokkuvõtlikult

1. Valida konstantide w, v ja α väärtused, kusjuures $0 \leq w \leq 1$, $0 \leq v \leq 1$ ja $0 \leq \alpha \leq 1$.

2. Määrata kindlaks sesoonsust iseloomustav ajaperioodide arv p. Kvartaalsete andmete korral on see 4, kuiste andmete korral 12.

3. Esimesed väärtused leitakse aja $t = 2$ jaoks. E_1, T_1 ja S_1 ei ole defineeritud:

aditiivne	multiplikatiivne
$E_2 = y_2$	$E_2 = y_2$
$T_2 = y_2 - y_1$	$T_2 = y_2 - y_1$
$S_2 = 0$	$S_2 = 1.$

4. Seni, kuni üks terve tsükkel (p perioodi) pole läbitud, ei ole võimalik leida sesoonse komponendi valemis esinevat komponenti S_{t-p}. Seega, väärtuste $t = 3, \ldots, p+1$ jaoks on vaja kasutada järgmisi valemeid:

aditiivne	multiplikatiivne
$E_t = wy_t + (1-w)(E_{t-1} + T_{t-1})$	$E_t = wy_t + (1-w)(E_{t-1} + T_{t-1})$
$T_t = v(E_t - E_{t-1}) + (1-v)T_{t-1}$	$T_t = v(E_t - E_{t-1}) + (1-v)T_{t-1}$
$S_t = y_t - E_t$	$S_t = \frac{y_t}{E_t}.$

5. Alates ajamuutuja t väärtusest $p+2$ muutuvad E_t ja S_t valemid, trendi T_t valem jääb samaks:

aditiivne	multiplikatiivne
$E_t = w(y_t - S_{t-p}) +$	$E_t = w\frac{y_t}{S_{t-p}} +$
$+(1-w)(E_{t-1} + T_{t-1})$	$+(1-w)(E_{t-1} + T_{t-1})$
$T_t = v(E_t - E_{t-1}) + (1-v)T_{t-1}$	$T_t = v(E_t - E_{t-1}) + (1-v)T_{t-1}$
$S_t = \alpha(y_t - E_t) + (1-\alpha)S_{t-p}$	$S_t = \alpha\frac{y_t}{E_t} + (1-\alpha)S_{t-p}.$

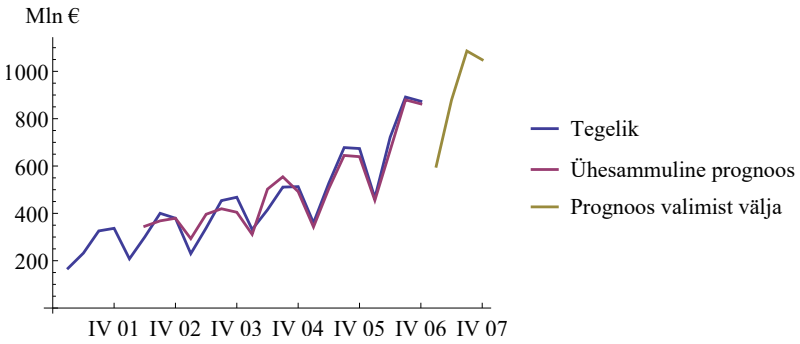
6. Prognoosimine:

aditiivne	multiplikatiivne
$F_{t+1} = E_t + T_t + S_{t+1-p}$	$F_{t+1} = (E_t + T_t)S_{t+1-p}$
$F_{t+2} = E_t + 2T_t + S_{t+2-p}$	$F_{t+2} = (E_t + 2T_t)S_{t+2-p}$
$F_{t+k} = E_t + kT_t + S_{t+k-p}$	$F_{t+k} = (E_t + kT_t)S_{t+k-p}.$

Näide 10.16. Eestis tehtud ehitustööd ja eksponentsilumine trendi ning sesoonsusega


N10Aegread
N10.16,20

Joonisel on aegrida, mis kirjeldab Eestis tehtud ehitustööde mahtusid aastatel 2001–2006 (miljonit eurot)[a]. Tegemist on kvartaalsete andmetega ja näha on tugev sesoonsus. Aegrea silumiseks kasutatakse multiplikatiivset eksponentsilumist trendi ja sesoonsusega. Perioodide $t = 6$ kuni $t = 24$ jaoks on leitud ühesammuline prognoos ning prognoosivead. Prognoosi keskmist ruutviga minimeerides on silumiskonstantideks saadud $w = 0{,}52$, $v = 0{,}24$ ja $\alpha = 0{,}49$, mille korral $MSE = 1716{,}2$. Keskmise ruutvea minimeerimiseks kasutati Exceli lahendajat *Solver*. Seejärel on tehtud prognoos aastaks 2007.



[a] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel EH012: ehitustööd jooksevhindades.

Eksponentsilumise võimalused ei piirdu vaadeldud mudelitega. Kombineerides erinevaid trendi ja sesoonsuse arvestamise võimalusi, on kokku 15 erinevat eksponentsilumise meetodit (R. Hyndman ja Athanasopoulos, 2013). Erinevalt regressioonmudelitest on eksponentsilumise mudelid võimelised peegeldama aegridades aja jooksul toimuvaid muutusi ja suudavad nendele muutustele vastavalt reageerida. Seepärast nimetatakse neid ka **adaptiivseteks mudeliteks**.

Silumiskonstantide ehk tasandusparameetrite optimaalseks valikuks pakuvad erinevad autorid erinevaid võimalusi, eesmärgiks on saada prognoosimisel võimalikult väike viga. Üldine soovitus on, et lühiajaliste prognooside korral tuleks anda suurem kaal aegrea hilisematele väärtustele ning võtta suurem silmumiskonstant. Prognoosiperioodi pikkuse suurenemisel tuleks lühiajalisi kõikumisi rohkem tasandada ning anda suurem kaal eelmiste perioodide informatsioonile, s.t tasandusparameeter võib olla väiksem (Vainu,

2006). Excelis on võimalik kasutada lahendajat *Solver*, mille abil otsitakse parameetritele sobivad väärtused, nii et prognoosi keskmine ruutviga MSE (või vigade ruutude summa SSE) oleks minimaalne. Selle abil on leitud tasandusparameetrite väärtused näites 10.16.

Eksponentsilumisel põhinevat prognoosimist on mugav teha sobiva tarkvara abil. Näiteks võib kasutada veebipõhist tarkvara aadressil `http://www.wessa.net/`, mille töö põhineb statistikapaketil R (Wessa, 2016).

10.13. Prognooside hindamine

Alapeatükis 10.8 tutvusime ühe sagedasti prognooside hindamiseks kasutatava suurusega, keskmise ruutveaga MSE. Toome siin uuesti ära selle valemi:

$$MSE = \frac{1}{n}\sum_{t=1}^{n} u_t^2, \tag{10.56}$$

kus n on prognoositavate väärtuste arv ehk prognoosi pikkus ja $u_t = y_t - F_t$ on prognoosiviga. Kuna keskmise ruutvea ühikuks on analüüsitava tunnuse ühiku ruut, siis tihti leitakse ruutjuur keskmisest ruutveast, mida võib nimetada **juuritud keskmiseks ruutveaks** (ka ruutkeskmine viga) (*Root Mean Square Error*):

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} u_t^2}. \tag{10.57}$$

Lisaks on kasutusel veel mitmed teised prognoosimisvõime näitajad.

Keskmine viga (*Mean Error*)

$$ME = \frac{1}{n}\sum_{t=1}^{n} u_t. \tag{10.58}$$

Keskmine absoluutviga (*Mean Absolute Deviation*)

$$MAD = \frac{1}{n}\sum_{t=1}^{n} |u_t|. \tag{10.59}$$

Keskmine suhteline viga (*Mean Percent Error*)

$$MPE = \frac{1}{n}\sum_{t=1}^{n} \frac{u_t}{y_t}. \tag{10.60}$$

Keskmine suhteline absoluutviga (*Mean Absolute Percent Error*)

$$MAPE = \frac{1}{n}\sum_{t=1}^{n} \frac{|u_t|}{y_t}. \tag{10.61}$$

Keskmine viga ME ja keskmine suhteline viga MPE võivad olla nii positiivsed kui ka negatiivsed ja iseloomustavad prognoosi nihet. Positiivse keskmise vea (või keskmise suhtelise vea) korral on tegelikud väärtused y_t keskmiselt suuremad kui prognoositud väärtused, s.t prognoos alahindab tegelikkust. Negatiivse keskmise vea korral on tegelikud väärtused keskmiselt väiksemad kui prognoositud väärtused ja prognoos ülehindab aegrea väärtusi. Hea prognoosi korral $MPE \leq 5\%$.

Näide 10.17. Ülehindavad prognoosid jaekaubanduses

Jaekaubanduses lähtutakse varude soetamisel müügimahu prognoosist. Kui see prognoos ülehindab tegelikkust, on tulemuseks varude liigne suurenemine. Praktika näitab, et kui varud kasvavad kiiremini kui müük, siis kasum langeb. Kasumi vähenemine on tingitud sellest, et varude vähendamiseks tuleb müüki suurendada, milleks alandatakse hinda. Lisaks vähendavad jaekauplejad siis järgmiste tellimuste suurust, mis mõjub negatiivselt kogu tarneahelale. Näiteks 1997. aasta algul paljud USA jaemüügiketid prognoosisid suuremat müüki ja tellisid kaupa vastavalt sellele. Jaemüügiketi Gap varud suurenesid 28%, kuid müük ainult 12%. Müügiketil Kmart suurenesid varud 8,9% ja müük 4,1%. Investorid reageerisid sellele ning mõlema aktsia hind langes. (Pulliam, 1997)

Keskmine absoluutviga MAD, keskmine ruutviga MSE (samuti selle juur $RMSE$) ja keskmine suhteline absoluutviga $MAPE$ ei sõltu prognoosivigade märgist. Need sõltuvad ainult sellest, kui lähedal on tegelikud väärtused prognoositud väärtustele.

Juuritud keskmine ruutviga $RMSE$ on prognoosivigade standardhälbe s hinnanguks. Kui me eeldame, et prognoosivead alluvad normaaljaotusele, siis ligikaudu 95% vigadest langeb vahemikku $0 \pm 2s$ ja ligikaudu 99,7% vahemikku $0 \pm 3s$. Kui see nii ei ole, siis tuleks prognoosimist korrigeerida.

Näitajaid, mis sõltuvad uuritava tunnuse ühikutest, nagu keskmine viga, keskmine absoluutne viga ja keskmine ruutviga, on mõtet kasutada vaid erinevate prognoosimismudelite võrdlemisel ühe ja sama aegrea korral. Erinevate aegridade puhul neid võrrelda ei saa. Siis tuleb kasutada suhtelisi vigu MPE ja $MAPE$. Prognoosi täpsus on väga hea, kui $MAPE$ ei ületa 10%, ning hea, kui jääb 10% ja 20% vahele.

N10Aegread
N10.18

Näide 10.18. Müügimahu prognooside täpsus

Alapeatükis 10.8 vaatasime näidet erinevate meetodite kasutamisest müügimahu prognoosimisel. Vaatluse all olid lihtne libisev keskmine, kaalutud libisev keskmine, eksponentsilumine ning lineaarne regressioon. Keskmine ruutviga MSE oli kõige väiksem eksponentsilumise korral. Nüüd on leitud nende meetodite korral ka teised prognoosimisvõime näitajad.

	ME	$RMSE$	MAD	MPE	$MAPE$
Lihtne libisev keskmine	−1,67	10,80	10,00	−0,91%	4,72%
Kaalutud libisev keskmine	−2,25	8,81	8,50	−1,16%	4,03%
Eksponentsilumine	−1,28	7,09	6,93	−0,68%	3,28%
Lineaarne regressioonmudel	−1,25	7,29	6,25	−0,67%	2,94%

Keskmine viga ME näitab, et kõigi meetodite korral prognoos ülehindab tegelikke väärtusi, sest keskmine viga on negatiivne. Sama näitab ka keskmine suhteline viga MPE. Kõige suurem on ülehindamine kaalutud libiseva keskmise korral. Keskmise absoluutvea MAD järgi on kõige parem prognoos lineaarse regressioonmudeli kasutamisel. Nii $RMSE$ kui ka MAD järgi saab kõige halvema prognoosi lihtsa libiseva keskmise põhjal.

Tabelis 10.13 on esitatud ettevõtjate hulgas läbiviidud küsitluse tulemused erinevate näitajate kasutamisest müügiprognooside täpsuse hindamisel. Vastav uuring viidi läbi USA-s ja vastajaid oli 86 ettevõttest (McCarthy jt, 2006). Vastajad võisid märkida mitu näitajat.

Tabel 10.13. Erinevate näitajate kasutamine prognooside hindamisel (McCarthy jt, 2006)

Täpsuse mõõt	Kasutajate protsent
$MAPE$	45
MPE	45
MAD	20
MSE	6

Aegrea silumiseks ja prognoosimiseks kasutatavat mudelit võib põhimõtteliselt testida kahes piirkonnas:

- valimi sees (*within sample*);
- valimist väljaspool (*out of sample*).

Valimiks nimetatakse siin neid aegrea väärtusi, mida kasutati mudeli saamiseks: eksponentsilumise tasandusparameetrite või regressioonmudeli parameetrite hindamiseks. Libiseva keskmise ning eksponentsilumise korral kasutatakse valimi sees ühesammulist prognoosimist. Regressioonmudeli korral on prognoositud väärtusteks mudelväärtused. Kui mudeli prognoosimisvõimet hinnatakse ainult valimi sees, s.t ainult nende aegrea väärtuste korral, millele tuginedes leiti sobiv mudel ja selle parameetrid, on oht **liigsobitamisele** (*overfitting*). See on olukord, kus mudel annab valimi sees tunduvalt parema prognoosi kui valimist väljas.

Kui me tahame hinnata mudeli prognoosimisvõimet valimist väljaspool, on meil vaja ka selles piirkonnas teada aegrea tegelikke väärtusi, millega prognoosi võrrelda. Libiseva keskmise ja eksponentsilumise korral kasutatakse valimist väljaspool mitmesammulist prognoosimist: leitakse F_{t+1}, F_{t+2}, $F_{t+3}, \ldots$, nii et viimane tegelik väärtus, mida prognooside leidmisel kasutatakse, on y_t. Seejärel saab väärtusi y_{t+1}, y_{t+2}, $y_{t+3}, \ldots$ teades leida prognoosivead ja prognoosi keskmise ruutvea.

Näide 10.19. Indeksi S&P 500 prognooside võrdlemine

N10Aegread
N10.19

Näites 10.8 silusime aktsiaindeksi S&P 500 aegrida kahe erineva silumiskonstandiga. Kasutasime indeksi väärtusi perioodil 6.04.–1.05.2015, need moodustavad valimi. Näites 10.9 prognoosisime indeksi kolme järgmist väärtust ja võrdlesime prognoositud väärtusi tegelike väärtustega. See oli prognoos väljaspool valimit. Leiame nüüd selle prognoosi keskmise ruutvea MSE mõlema silumiskonstandi korral.

		$w = 0{,}2$		$w = 0{,}5$	
Kuupäev	S&P 500	Prognoos	Prognoosi-viga	Prognoos	Prognoosi-viga
4.05.2015	2114,49	2103,73	10,76	2102,98	11,51
5.05.2015	2089,46	2103,73	−14,27	2102,98	−13,52
6.05.2015	2080,15	2103,73	−23,58	2102,98	−22,83

Kui $w = 0{,}2$, siis

$$MSE = \frac{1}{3}\left(10{,}76^2 + (-14{,}27)^2 + (-23{,}58)^2\right) \approx 291{,}8$$

ja silumiskonstandi $w = 0{,}5$ korral

$$MSE = \frac{1}{3}\left(11{,}51^2 + (-13{,}52)^2 + (-22{,}83)^2\right) \approx 278{,}8.$$

Näeme, et silumiskonstandi 0,5 kasutamine annab valimist väljaspool parema prognoosi.

Trendiga ning trendi ja sesoonsusega eksponentsilumise korral peab enne valimi sees toimuvat ühesammulist prognoosimist eelnema algväärtustamine. See on ajavahemik, mille korral pole veel trendi ja/või sesoonseid komponente. Näiteks trendi ja sesoonsusega eksponentsilumise korral toimub algväärtustamine vahemikus $t = 1, \ldots, p+1$, kus p on sesoonsust iseloomustav ajaperioodide arv. Olemasolev aegrida jagatakse sel juhul kolmeks nn aknaks:

1) algväärtustamise ehk initsialiseerimise aken;
2) mudeli kalibreerimise aken (mudeli parameetrite leidmine);
3) mudeli testimise aken (prognoosimine valimist väljaspool).

Kui andmeid on piisavalt, siis algväärtustamisest ülejääv aegrida jagada kaheks, nii et viimane neljandik oleks mudeli testimise aken.

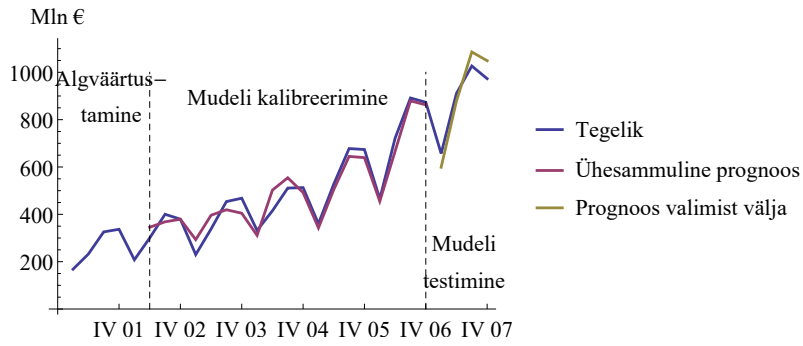
Näide 10.20. Eestis tehtud ehitustööde aegrea mudeli testimine

N10 Aegread
N10.16,20

Näites 10.16 kasutasime aastaid 2001 kuni 2006 sobiva eksponentsilumise mudeli leidmiseks Eestis tehtud ehitustööde (miljonit eurot) prognoosimiseks. Minimeerides prognoosi keskmist ruutviga, saime silumiskonstantideks $w = 0{,}52$, $v = 0{,}24$ ja $\alpha = 0{,}49$. Keskmine ruutviga $MSE = 1716{,}21$ (miljon eurot)^2. Kuna on teada ka vastava suuruse 2007. aasta kvartaalsed andmed, võrdleme prognoosi tegelike väärtustega ning leiame keskmise ruutvea valimist väljaspool.

Kvartal	Tegelik	Prognoos	Prognoosiviga
1	657,1	598,6	58,5
2	911,2	878,7	32,5
3	1026,8	1086,4	−59,6
4	973,1	1049,5	−76,4

Prognoosimisel valimist välja saame keskmiseks ruutveaks 3466,2 (miljon eurot)^2, mis on *ca* kaks korda suurem kui valimi sees toimunud ühesammulise prognoosimise korral.



Regressioonmudeli korral algväärtustamise aken puudub. Näites 10.14 kasutasime kaht akent: aastaid 1949–1955 mudeli kalibreerimiseks (trendi ja keskmiste sesoonsete komponentide leidmine) ning aastat 1956 mudeli testimiseks.

10.14. Ülevaade prognoosimismeetoditest

Prognoosimisega on inimene tegelenud juba iidsetest aegadest saadik. Ürginimese elu sõltus väga palju loodusest ja ilmastikust ning halvad ajad elasid paremini üle need, kes oskasid neid ette näha ja varusid soetada. Antiikajal mõõdeti Egiptuses Niiluse jõe veetaset, et prognoosida saaki ja sellest tulenevalt määras valitseja oma alluvatele maksud. Enne 1950-ndaid aastaid olid regressioonanalüüs ja aegridade kompleksanalüüs olemas, aga neid kasutati vaid akadeemilistes ja valitsusasutustes. Ärinduses süstemaatilist prognoosimist ei toimunud. 1960-ndatel aastatel võimaldas arvutusvõimsuse odavnemine hakata kasutama arvutuslikult mahukamaid meetodeid. Akadeemilistes ringkondades otsiti üldist prognoosimismeetodit. 1970-ndail töötasid George Box ja Gwilym Jenkins välja ARIMA mudeleid kasutava aegridade süstemaatilise analüüsi meetodi, mida nimetatakse nüüd Boxi-Jenkinsi meetodiks (Box ja Jenkins, 1970). See tugineb aegridade autokorrelatsioonil ja võimaldab analüüsida väga erineva mustriga aegridu.

1970-ndate lõpus sai selgeks, et prognoosimise metoodika võib olla ükskõik kui hea, aga prognoosimine on mõttetu, kui tulemusi ei kasutata. Erinevad uuringud näitasid, et otsuste vastuvõtjad tihti ei kasutanud analüütikute prognoose, sõltumata nende täpsusest. Otsuste vastuvõtmisel ignoreeriti objektiivset reaalsust, toetuti nägemustele, illusioonidele ja tihti poliitilistele teguritele ning omakasule.

Tänapäeval on kasutusel hulgaliselt erinevaid prognoosimismeetodeid, nii kvantitatiivseid kui ka kvalitatiivseid.

Kvantitatiivsed ja kvalitatiivsed prognoosimismeetodid

- **Kvantitatiivsed meetodid** kasutavad andmete töötlemiseks matemaatikat ja statistikat:
  - aegridade analüüs;
  - regressioonmudelid seletavate tunnustega (ptk 9);
  - toote elutsükli analüüs (Bassi mudel näites 10.12).
- **Kvalitatiivsed meetodid** — puudub matemaatiline mudel, kuna mineviku andmeid pole tuleviku jaoks võimalik kasutada. Prognooside tegijateks on protsessis osalejad või eksperdid.
  - Juhtimisotsustuslik (*Jury of Excecutives*) — juhtivtöötajate kokkusaamine eesmärgiga genereerida prognoose, kui minevikuandmed on ebapiisavad, kvantitatiivsed tulemused vasturäägivad jms.

- Müügitöötajate prognoosid — müügiprotsessile kõige lähemal olevatel isikutel on kõige rohkem informatsiooni. Prognoose on lihtne rühmitada müügipiirkonna, toote, kliendi järgi.
- Delphi meetod — teatud arv eksperte jõuab konsensuseni. Konsensuseni jõudmiseks võib läbida mitu tsüklit: tehakse vahekokkuvõtted, iga ekspert loeb need läbi ja korrigeerib oma prognoosi. Eesmärgiks on saada võimalikult väike prognooside hajuvus.
- Tarbijauuringud.

Tabelis 10.14 on ülevaade erinevate müügimahu prognoosimiseks kasutatavate meetodite tuntusest ettevõtjate hulgas (McCarthy jt, 2006).

Tabel 10.14. Prognoosimismeetodite tuntus ettevõtjate seas

Prognoosimismeetod	Tuttav, %	Mõnevõrra tuttav, %	Ei tea, %
Kvalitatiivsed			
Juhtimisotsustuslik	57	17	26
Müügitöötajate prognoos	66	18	17
Tarbijauuring	62	21	17
Kvantitatiivsed			
Libisev keskmine	84	16	0
Eksponentsilumine	76	20	4
Silumine regressioonjoonega	69	23	7
Regressioonmudelid seletavate tunnustega	73	24	3
Aegridade kompleksanalüüs	38	16	46
Toote elutsükli analüüs	49	25	25
Boxi-Jenkinsi meetod	30	22	48

Näeme, et libisev keskmine on kõige tuntum kvantitatiivne prognoosimismeetod, sellele järgneb eksponentsilumine.

Meetod, mida kasutatakse järgmise kuu käibe prognoosimiseks (lühiajaline prognoos), ei pruugi sobida järgmise viie aasta käibe prognoosimiseks. Lühiajaliste prognooside korral kasutatakse rohkem kvantitatiivseid meetodeid. Pikaajaliste prognooside korral analüüsitakse ärivõi majandussituatsioone ning kasutatakse rohkem kvalitatiivseid meetodeid. Prognooside jaotus pikkuse järgi on toodud tabelis 10.15.

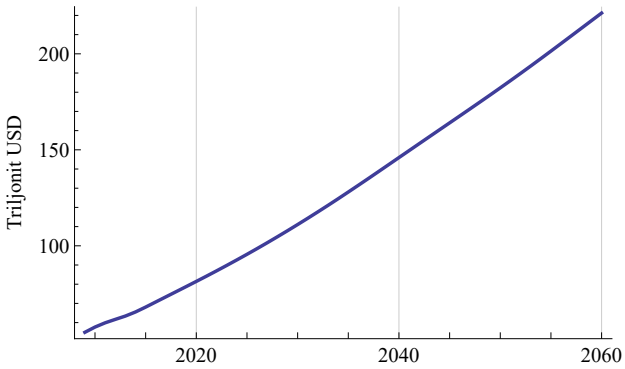
Prognooside jaotus pikkuse järgi on suhteline. Näiteks riigi energiatarbimise prognoos 5–6 aasta peale on lühiajaline, 50 aasta peale pikaajaline.

Kasutatakse ka kvalitatiivsete ja kvantitatiivsete meetodite kombineerimist. Tavaliselt alustatakse kvantitatiivsest prognoosimisest ja tulemusi korrigeeritakse mõne kvalitatiivse meetodi abil. Näiteks OECD

Tabel 10.15. Prognooside jaotus pikkuse järgi

Lühiajalised (*short-term*) 3–6 kuud	Operatiivsed	Nõudluse, varude prognoosimine, tootmise planeerimine
Keskmise pikkusega (*medium term*) 6 kuud kuni 2 aastat	Taktikalised	Rahavoogude prognoosimine laenu- või liisingumakseteks
Pikaajalised (*long-term*) üle kahe aasta	Strateegilised	Uuringud, tootearendus

pikaajaline maailma SKP prognoos (joonis 10.24) põhineb hinnangutel üksikute riikide ja maailmamajanduse majanduskliima kohta, kvantitatiivsete mudelite abil saadud tulemusi kombineeritakse eksperthinnangutega (*GDP long-term forecast* 2016).



Joonis 10.24. Maailma SKP prognoos aastani 2060 (*GDP long-term forecast* 2016)

Mida arvestada prognoosimismeetodi valikul?

- **Ajahorisont**. Kui pikk prognoos on vajalik?
- **Andmete muster**. Kas esineb trend, tsüklilisus, sesoonsus? Erinevad meetodid sobivad erineva mustriga aegridadele.
- **Prognoosimiskulud**. Arenduskulud (mudeli väljatöötamine), andmete kogumine ja ettevalmistamine, tegevuskulud.
- **Täpsus**. Mõnikord piisab suhtelisest veast ±10%, teisel juhul on ka viga ±5% liiga suur.
- **Rakendamise lihtsus**. Otsuse vastuvõtja vastutab oma otsuse eest. Kui ta toetub prognoosile, siis peab ta seda usaldama, järelikult soovib ta kasutatavast meetodist aru saada. Mida keerulisem meetod, seda vähem on neid, kes seda mõistavad. Seepärast kasutatakse keerulisemaid meetodeid harvemini.
- **Tarkvara olemasolu**.

Prognoosimiseks kasutatava tarkvara võib jagada kolme suurde rühma. Esiteks tavaline tabelarvutus nagu Excel või LibreOffice Calc. Teise rühma kuuluvad statistikapaketid nagu SPSS, SAS, Statgraphics, vabavara Gretl, R jt. Need võimaldavad kasutada suurt hulka erinevaid statistilisi meetodeid, mille hulgas on ka prognoosimine. Kolmas rühm on spetsiaalne prognoosimistarkvara: Forecast Pro, Autobox jpt. Tüüpiliselt sisaldab seda liiki tarkvara regressioonanalüüsi, eksponentsilumist, Boxi-Jenkinsi ARIMA modelleerimist, aga ka mitmeid spetsiifilisi meetodeid, mida tavalistes statistikapakettides ei pruugi olla. Spetsiaalne tarkvara on tavaliselt ka rohkem automatiseeritud. See analüüsib andmeid, soovitab kasutada sobivat protseduuri või mudelit, leiab mudeli parameetrid ning väljastab prognoosid, graafikud ja prognooside statistilised näitajad. Kasutaja kas aktsepteerib soovitusi või mitte. Vähem automatiseeritud tarkvara nõuab kasutajalt rohkem statistika-alaseid teadmisi. Aegridade sesoonseks korrigeerimiseks võib kasutada Eurostati väljatöötatud programmi DEMETRA[5], mis sisaldab nii USA Rahvaloenduse büroo loodud programmi X12-ARIMA kui ka Hispaania Keskpangas loodud tarkvara TRAMO/SEATS (Täht, 2007).

10.15. Ülesanded

Elementaaranalüüs

10.1. 19. novembril 2014. aastal oli Tartu turul[6] banaanide kilohind 1,1 eurot ja 25. novembril oli see 1,4 eurot. Kui suur oli banaanide hinna keskmine absoluutne juurdekasv päevas? VASTUS lk 692.

10.2. Tabelis on kaupade jaemüük jooksevhindades Eestis aastatel 2009–2011[7].

Aasta	Kaupade jaemüük, mln eurot
2009	4596,0
2010	4631,9
2011	5189,3

Leida aastate 2010 ja 2011 jaoks
a) absoluutne aheljuurdekasv;

[5] European Comission, Joinup `https://joinup.ec.europa.eu/software/demetraplus/home`

[6] `http://www.tartuturg.ee/`

[7] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel KM0011: kaupade ja mootorsõidukite hoolduse ja remonditeenuse müük jooksevhindades.

b) absoluutne alusjuurdekasv;
c) aheljuurdekasvutempo;
d) tagasivaatav aheljuurdekasvutempo;
e) ahelkasvutempo ehk ahelindeks;
f) aluskasvutempo ehk alusindeks.

VASTUS lk 692.

10.3. Eesti SKP väärtus jooksevhindades oli 1996. aastal 3637,6 miljonit eurot ja 2005. aastal 11 181,7 miljonit eurot[8].

1. Leida SKP keskmine kasvutempo aastas perioodil 1996–2005.
2. Võttes aluseks SKP väärtuse 2005. aastal ja leitud keskmise kasvutempo, prognoosida SKP väärtust aastal 2008.
3. Võrrelda prognoosi tegeliku väärtusega. SKP tegelik väärtus 2008. aastal oli 16 235,1 miljonit eurot.

VASTUS lk 692.

10.4. Leida keskmine kvartalikäive perioodil 2013–2014 järgmiste andmete korral:

Periood	Käive perioodil, mln €
2013. a I kvartal	0,5
2013. a II kvartal	0,55
2013. a II poolaasta	1,3
2014. a	3

VASTUS lk 692.

10.5. Esitatud on ettevõtte kassaseis iga kuu alguses. Leida I ja II kvartali keskmine kassaseis ning I poolaasta keskmine kassaseis. VASTUS lk 692.

Kuupäev	1.01	1.02	1.03	1.04	1.05	1.06	1.07
Kassaseis, tuh €	40	190	30	660	580	420	620

Aegridade korrigeerimine

10.6. Pereisa jälgib pidevalt pere elektritarbimist. Tal on olemas elektrienergia tarbimise andmed iga kuu kohta ja ta soovib võrrelda tarbimist kuude lõikes. Probleemiks on kuude erinev pikkus.

1. Kui aastas on 365 päeva, siis kui suur on keskmine päevade arv kuus?
2. Millise kordajaga tuleb korrutada juuni, juuli ja augusti päevade arvu, et saada keskmine päevade arv kuus?

[8] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel RAA012: sisemajanduse koguprodukt ja kogurahvatulu.

3. Juunis oli pere elektrienergia tarbimine 808 kWh, juulis 810 kWh ja augustis 812 kWh. Korrigeerida kuude elektritarbimist nii, et need vastaksid keskmisele päevade arvule kuus. Millisel kuul oli korrigeeritud tarbimine kõige suurem ja millisel kuul kõige väiksem?

VASTUS lk 692.

10.7. Pood on lahti ainult tööpäevadel. 2015. aastal oli keskmine tööpäevade arv kuus 21,0. Juulis oli 23 tööpäeva, augustis 20 ja septembris 22. Poe käive oli juulis 41,2 tuhat eurot, augustis 40,5 tuhat eurot ning septembris 40,8 tuhat eurot. Korrigeerida kuude käibeid nii, et need vastaksid keskmisele tööpäevade arvule kuus. Millisel kuul oli tööpäevade arvuga korrigeeritud käive kõige suurem? VASTUS lk 692.

Eksponentsilumine

10.8. Börsianalüütik kasutab aktsiahinna muutuste silumiseks eksponentsilumist konstandiga 0,2. Esmaspäeval oli aktsiahind 12,5 eurot ning silutud väärtus 9,25 eurot. Tesipäeval oli aktsiahind 11,3 eurot. Kui suur on teisipäeva silutud väärtus? VASTUS lk 692.

10.9. Jaanuarikuu käibeks prognoositi 28 tuhat eurot, tegelik käive oli 26 tuhat eurot. Leida veebruarikuu prognoos, kui kasutada eksponentsilumist

a) konstandiga 0,1;
b) konstandiga 0,3.

VASTUS lk 692.

10.10. Keemilises puhastuses kasutatakse seadmete koormuse prognoosimiseks eksponentsilumist konstandiga 0,1. Augustikuuks prognoositi koormuseks 88 protsenti, tegelik koormus oli 89,6%.

1. Milline on septembrikuu prognoos?
2. Kui septembris oli tegelik koormus 92%, siis milline on prognoos oktoobrikuuks?

VASTUS lk 692.

10.11. Ettevõtte analüütik prognoosis teatud kauba nõudlust 2013. aastaks 520 tuhat tükki. Tegelik nõudlus oli 550 tuhat tükki. Järgmiseks aastaks prognoosis ta nõudluseks 530,5 tuhat tükki. Kui suurt silumiskonstanti analüütik kasutas, kui on teada, et ta kasutas prognoosimiseks eksponentsilumist? VASTUS lk 692.

Silumine regressioonjoonega

10.12. Eesti rahvaarvu muutumist aastatel 1970–1990 kirjeldab hästi lineaarne trend $\hat{N} = 1369 + 10{,}08t$, kus N on rahvaarv tuhandetes ning t aeg aastates, $t = 0$ aastal 1970 (Tiit, 1995).

1. Mida näitab aja t ees olev kordaja?
2. Kuidas näeks välja see mudel siis, kui aeg t oleks kuudes ja $t = 0$ jaanuaris 1970?

Vastus lk 692.

10.13. 2011. aastal viidi Tartu Ülikoolis koostöös SA Poliitikauuringute Keskusega Praxis läbi energeetika tööjõu uuring, mille tellijaks oli Eesti Elektritööstuse Liit (Eamets jt, 2011). Uuringu üheks eesmärgiks oli prognoosida tööjõu vajadust energeetikasektoris kuni aastani 2020. Selleks prognoositi põlevkiviõlide tootmises kasutatava põlevkivi koguse muutumist aastani 2020. Mudeli leidmiseks kasutati aastate 1999–2010 andmeid. Ühe stsenaariumi korral eeldati, et põlevkivi kasutamisel jätkub lineaarne trend ning mudeliks oli $\hat{y} = 192{,}29t + 1466{,}8$, kus y on kasutatava põlevkivi hulk (tuhat tonni) ja t aeg aastates, $t = 0$ aastal 1999. Teise stsenaariumi kohaselt kasutatava põlevkivi mahu kasv tulevikus aeglustuks ning kasutati logaritmilist trendi $\hat{y} = 848{,}32 \ln t + 1270{,}8$.

1. Kui palju suureneks põlevkivi kasutamine aastatel 2010 kuni 2020 esimese stsenaariumi järgi? Leida tonnides ja protsentides.
2. Kui palju suureneks põlevkivi kasutamine aastatel 2010 kuni 2020 teise stsenaariumi järgi? Leida tonnides ja protsentides.

Vastus lk 692.

10.14. RFID (*radio-frequency identification*) on raadiolaineid kasutav tehnoloogia esemete ja ka elusolendite märgistamiseks. RFID-kiipe on võimalik panna kaubasaadetistele ja ka üksikutele toodetele, lemmikloomadele, neid kasutatakse pääslates ja e-piletites. Alates 2007. aastast on Eesti Vabariigi passid varustatud RFID-kiibiga. Prognooside kohaselt on 2021. aastaks 25% mittetoidukaupadest ja 5% toidukaupadest varustatud RFID kiibiga. B. Stedron ja H. Bínová analüüsisid RFID tehnoloogia kasutamise arengutrende aastatel 2004–2009 ning tegid prognoose aastani 2020 (Stedron ja Bínová, 2015). Nad leidsid, et RFID-kiipide hind kahaneb eksponentsiaalselt mudeli $\hat{y} = 0{,}6213e^{-0{,}274t}$ järgi, kus y on hind USA dollarites ja t aeg aastates, $t = 1$ aastal 2004.

1. Kui suur oli RFID-kiipide hind 2004. aastal?
2. Kui palju kiipide hind aastas väheneb?
3. Mis aastaks langeb RFID-kiipide hind 1 USD sendini?

Vastus lk 692.

10.15. Tuletõrje ja päästeriskide analüüsi raames uuriti Eesti Sisekaitseakadeemias tulekahjudes hukkunutega seonduvat statistikat. Kalendrikuude lõikes on tulesurmades hukkunute arv ebaühtlane. Kasutades andmeid aastast 2005–2010 leiti mudel, mis kirjeldab hukkunute arvu protsentuaalset jaotust kalendrikuude lõikes: $\hat{y} = 0{,}3251t^2 - 4{,}3734t + 19{,}419$, kus y on protsent hukkunute arvust aastas ja t on

kuu number aastas (Randoja ja Käerdi, 2011). Millisel kuul aastas on hukkunute arv kõige väiksem? VASTUS lk 692.

Prognoosi viga

10.16. Kauba hinda mõõdetakse eurodes. Hinna prognoosi täpsuse iseloomustamiseks on kasutatud erinevaid näitajaid: keskmine ruutviga MSE, juuritud keskmine ruutviga $RMSE$, keskmine viga ME, keskmine absoluutviga MAD, keskmine suhteline viga MPE ja keskmine suhteline absoluutviga $MAPE$. Mis ühikutes on need keskmised vead? VASTUS lk 692.

10.17. Prognoosi keskmine suhteline viga oli 7,5%. Kas prognoos alahindab või ülehindab tegelikke väärtusi? VASTUS lk 692.

10.18. Huong Ngo Higgins analüüsis oma 1998. aastal ilmunud artiklis „Analyst Forecasting Performance in Seven Countries“ (Higgins, 1998), kuidas on seotud ettevõtte finantsiline „läbipaistvus“ (*disclosure*) ja analüütikute võime prognoosida tulu aktsia kohta. Analüüsi aluseks oli üle 11 tuhande ettevõtte USA-s, Jaapanis ja viies Euroopa riigis. Läbipaistvuse hindamiseks kasutati finantsalase läbipaistvuse indeksit, mille leidmisel kasutati 142 eksperdi hinnanguid (ettevõtete juhid, investorid, pankurid, teadlased). Higgins sai järgmise mudeli:

$$MAPE = 0{,}55 - 0{,}01DISCL + \varepsilon,$$

kus $MAPE$ on prognooside keskmine suhteline absoluutviga ja $DISCL$ läbipaistvuse indeks. Kuidas mõjutab ettevõtte finantsiline läbipaistvus prognooside täpsust? VASTUS lk 692.

ÜL10Aegread

Järgmiste ülesannete andmed on failis ÜL10Aegread

Elementaaranalüüs

A.10.1. Kasutades andmeid USA dollari ja euro kursi kohta aprillis 2004, leida

a) absoluutne aheljuurdekasv, kasvutempo ja juurdekasvutempo;
b) kuu keskmine kurss, keskmine absoluutne aheljuurdekasv, keskmine kasvutempo ja keskmine juurdekasvutempo.

VASTUS lk 692.

A.10.2. Tabelis on toodud keskmine brutokuupalk Eestis vahemikus 2002. aasta I kvartal kuni 2011. aasta IV kvartal[9]. Leida alusindeks ja ahelindeks ning konstrueerida diagramm, kus oleks esitatud mõlema indeksi dünaamika. VASTUS lk 692.

A.10.3. Eesti Pank on Statistikaameti kõrval teine riikliku statistika

[9] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel PA010: keskmine brutokuupalk ja brutotunnipalk põhitegevusala (EMTAK2008) järgi (kvartalid).

tegija Eestis. Eesti Panga ülesandeks on finantssektori statistika kogumine ja avaldamine. Tabelis on kodumajapidamistele antud laenude jääk iga kuu lõpus aastal 2013, kõik laenud kokku[10]. Leida kvartali keskmine laenujääk kõikide kvartalite jaoks ja aasta keskmine laenujääk. VASTUS lk 692.

A.10.4. Lähtudes õhutemperatuuri mõõtmisandmetest Tallinnas 1. juunil 2013[11], leida päeva keskmine õhutemperatuur. VASTUS lk 692.

A.10.5. Eesti Kontsert korraldab erinevat liiki kontserte, mille kuulajate arvud (tuhat) perioodil 1995–2005 on tabelis[12]. Millist liiki kontserdil oli kuulajate arvu keskmine kasvutempo aastas kõige suurem ja millist liiki kontserdil kõige väiksem? VASTUS lk 693.

A.10.6. Lossimine on kauba laevalt maha laadimine sadamas. Tabelis on kaupade lossimine Eesti sadamates, tuhat tonni[13]. Leida

a) millisel aastal oli kuu keskmine kasvutempo kõige väiksem;
b) millisel kalendrikuul oli kuu keskmine absoluutne aheljuurdekasv kõige suurem (üle kõikide aastate).

VASTUS lk 693.

A.10.7. Tabelis on toodud ühe tonni bensiini turuhind ajavahemikul 2.01.–28.12.2007, hinnad on teisendatud eurodesse. Turuhind on määratud vaid kauplemispäevadel, s.o tööpäevadel.

1. Leida absoluutne aheljuurdekasv päevas ja keskmine absoluutne aheljuurdekasv.
2. Leida keskmine absoluutne aheljuurdekasv nädalapäevade kaupa.
3. Millisel nädalapäeval tuleks bensiinitanklal oma varusid täiendada?

VASTUS lk 693.

A.10.8. Tabelis on USD/EUR kuu keskmine kurss jaanuarist 2008 kuni detsemberini 2011[14]. *Libisev keskmine*

1. Siluda seda aegrida 10 kuu ja 20 kuu libiseva keskmisega.
2. Prognoosida 2012. aasta jaanuarikuu keskmist kurssi, võttes aluseks nii 10 kuu libisev keskmine kui 20 kuu libisev keskmine.
3. Tegelik kuu keskmine kurss oli 2012. aasta jaanuaris 1,2905. Kumb libisev keskmine andis täpsema prognoosi?
4. Luua diagramm, kus on nii tegelikud väärtused kui ka mõlemad libisevad keskmised.

[10] Allikas: Eesti Pank `https://www.eestipank.ee/`

[11] Allikas: `http://www.ilm.ee`

[12] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel KU112: Eesti Kontserdi Eestis korraldatud kontserdid liigi järgi (1995–2005).

[13] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel TS185: kaupade lastimine ja lossimine Eesti sadamates (kuud).

[14] Allikas: Euroopa Keskpank `http://sdw.ecb.europa.eu/`

VASTUS lk 693.

A.10.9. Toodud on andmed laevapiletite müügiga tegeleva ettevõtte päevakäibe kohta 1997. aasta aprillis. Kroonid on teisendatud eurodeks.

1. Luua diagramm ja määrata sellelt libiseva keskmise jaoks sobiv samm. Samm valida selline, et silutaks ära sesoonsed kõikumised. Milline samm sobib?
2. Lisada diagrammile vastav libiseva keskmise joon.
3. Kui suur on selle libiseva keskmise väärtus 30. aprillil?

VASTUS lk 693.

A.10.10. Tabelis on toodud Arco Vara aktsia sulgemishind 2. jaanuarist – 3. veebruarini 2015. Leida Bollingeri koridori (vt näide 5.28) piirid $\mu - 2\sigma$ ja $\mu + 2\sigma$ 30. jaanuari ning 1. ja 2. veebruari jaoks. Mingi päeva sulgemishinna Bollingeri piiride arvutamisel on nii keskväärtus μ kui ka standardhälve σ libisevad ja leitakse eelneva 20 päeva hindade põhjal. VASTUS lk 693.

Eksponentsilumine

A.10.11. Toodud on Tallinna börsi OMXT indeksi väärtuste aegrida ajavahemikul 1.09.–30.12.2015. Siluda aegrida eksponentsiaalselt kahe erineva silumiskonstandi väärtuse korral: 0,2 ja 0,7. Lisada vastav diagramm. VASTUS lk 693.

A.10.12. Käibe prognoosimiseks sobib tihti eksponentsilumine. Tabelis on USA Illinoisi osariigis asuva Palatine toidupoe kalatoodete käive nädalas ($) 1.01.–16.12.1990, kokku 50 nädalat (*Dominick's Database* 2016). Kuna jõulude ja aastavahetuse ajal ostetakse toitu oluliselt rohkem, on aasta viimased kaks nädalat välja jäetud.

1. Siluda aegrida eksponentsiaalselt silumiskonstandiga 0,4.
2. Kasutada silutud väärtusi prognoosimiseks, nii et nädala prognoos on eelneva nädala silutud väärtus.
3. Leida prognoosivead.
4. Leida prognoosi keskmine ruutviga MSE, mis on prognoosivigade ruutude aritmeetiline keskmine.
5. Muuta silumiskonstanti, võttes väärtusteks 0,3 ja 0,5 ning mõlemal juhul fikseerida MSE. Millise silumiskonstandi korral on MSE kõige väiksem?
6. Minimeerida MSE väärtust, muutes silumiskonstante. Programmis Excel kasutada selleks lahendajat *Solver*. Sihifunktsiooniks (*Set Objective*) on MSE (lahter, kus asub arvutatud MSE väärtus). Seda tuleb minimeerida (*To Min*), muutes silumiskonstandi väärtust (*By Changing Variable Cells*), s.t viidata lahtrile, kus asub silumiskonstant.
7. Konstrueerida diagramm, kus on käibe tegelikud ja prognoositud väärtused. Kasutada eelmises punktis leitud silumiskonstanti.

Vastus lk 693.

Silumine lineaarse regressioonjoonega

A.10.13. Tabelis on toodud majanduslikult aktiivsete ettevõtete arv kahes Eesti maakonnas aastatel 2000 kuni 2005.

1. Kasutades lineaarset regressiooni, leida mõlema aegrea jaoks trendijoone parameetrid.
2. Kui palju lisandus aastas keskmiselt majanduslikult aktiivseid ettevõtteid Ida-Viru maakonnas ja kui palju Lääne-Viru maakonnas?
3. Kasutades leitud lineaarset mudelit, prognoosida majanduslikult aktiivsete ettevõtete arvu kummaski maakonnas 2006. aastal.
4. Võrrelda prognoosi tegeliku ettevõtete arvuga 2006. aastal (Ida-Viru maakonnas 2438, Lääne-Viru maakonnas 1486). Leida, mitu protsenti erineb tegelik prognoositust.

Vastus lk 695.

A.10.14. On teada, et puuviljade käive sõltub väga tugevasti aastaajast. Tabelis on ühe hulgimüügifirma banaanide müügikäive kvartalis aastatel 1999–2001 ning sama perioodi Eesti SKP. Kroonid on teisendatud eurodeks.

1. Leida korrelatsioonikordaja SKP ja käibe vahel.
2. Siluda käivet lineaarse regressioonjoonega.
3. Luua diagramm, kus on SKP, tegelik käive ja silutud käive.
4. Leida korrelatsioonikordaja SKP ja silutud käibe vahel ning võrrelda 1. osas leitud korrelatsioonikordajaga.
5. Milline on järeldus?

Vastus lk 695.

Silumine mittelineaarse regressioonjoonega

A.10.15. Vaba aja kulutuste hulka kuuluvad kulutused spordile, meelelahutusele ja kultuurile, reisimisele, hobidele, söömisele väljaspool kodu jms. Tabelis on toodud kaubagrupi „Vaba aeg“ tarbijahinnaindeksi (THI) dünaamika Eestis aastatel 2004–2007[15]. 1997. aastal oli indeksi väärtus 100.

1. Siluda seda aegrida kolme erineva regressioonjoonega: lineaarne, ruutpolünoom ja eksponentsiaalne.
2. Milline mudel sobib kõige paremini? Miks?
3. Kasutada parimat mudelit THI 2008. aasta jaanuari väärtuse prognoosimiseks.

Vastus lk 695.

A.10.16. Raha pakkumine hõlmab kogu ringluses oleva raha, mida üksikisikud ja ettevõtted saavad kasutada. Tabelis on raha pakkumine

[15] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel IA02: tarbijahinnaindeks, 1997 = 100 (kuud).

Eestis 1994. aasta I kvartalist kuni 1998. aasta II kvartalini[16]. Luua aegrea diagramm ning selgitada välja, millise kujuga trendijoon silub seda aegrida kõige paremini. VASTUS lk 695.

A.10.17. Tabelis on Eesti rahvaarv aastatel 1950–2016.

1. Luua diagramm, mis kujutab Eesti rahvaarvu muutumist aastatel 1950–2016.
2. Perioodi 1950–1990 jaoks leida lineaarne trend. Kui suur oli sellel perioodil keskmine rahvaarvu muutus aastas?
3. Leida perioodi 1950–1990 jaoks ka sobiv mittelineaarne trend. Kui suur oli selle mudeli järgi rahvaarvu aastane muutus 1960. ja 1990. aastal?
4. Perioodi 2000–2016 jaoks leida lineaarne trend. Kui suur oli sellel perioodil keskmine rahvaarvu muutus aastas?
5. Leida perioodi 2000–2016 jaoks ka sobiv mittelineaarne trend. Kui suur oli selle mudeli järgi rahvaarvu aastane muutus 2000. ja 2015. aastal?

VASTUS lk 695.

A.10.18. Eksponentsiaalset mudelit ja mitmeid teisi mudeleid on võimalik lineariseerida ning seejärel kasutada lineaarse mudeli parameetrite leidmise tehnikat. See on oluliselt täpsem ning võimaldab leida ka parameetrite usalduspiirid.

Tabelis on kodumajapidamiste laenujääk Eesti pankades aastatel 1997–2007, miljonit eurot[17].

1. Luua diagramm ja lisada eksponentsiaalne trendijoon $y_t = y_0 e^{at}$ koos valemi ja determinatsioonikordajaga.
2. Leida laenujäägi naturaallogaritmid.
3. Hinnata vastavat lineaarset mudelit $\ln y_t = \ln y_0 + at$.
4. Lineaarse mudeli parameetrite hinnangute põhjal arvutada eksponentsiaalse mudeli parameetrid.
5. Võrrelda tulemusi diagrammil toodud eksponentsiaalse mudeli parameetritega.

VASTUS lk 695.

Erinevad silumismeetodid

A.10.19. USA Põllumajandusministeerium USDA avaldab detailseid andmeid erinevate põllumajandussaaduste tootmise, tarbimise ja hindade kohta. Tabelis on kanaliha tarbimine elaniku kohta perioodil 1990–1999, kg aastas[18].

1. Koostada tarbimise muutumist kirjeldav diagramm.

[16] Allikas: Eesti Pank `http://www.eestipank.ee/`

[17] Allikas: Eesti Pank `http://www.eestipank.ee/`

[18] Allikas: USDA Economics, Statistics and Market Information System `http://usda.mannlib.cornell.edu`. Table081: Young chicken: Per capita consumption, ready-to-cook weight basis.

2. Prognoosida 2000. aasta tarbimist neljal erineval meetodil:
   (a) lineaarne trend;
   (b) viie aasta lihtne libisev keskmine;
   (c) kolme aasta kaalutud libisev keskmine kaaludega 0,6, 0,3 ja 0,1;
   (d) eksponentsilumine konstandiga 0,8.
3. Aastal 2000 oli tegelik tarbimine 40,63 kg aastas. Milline meetod andis kõige parema prognoosi? Miks?

VASTUS lk 695.

A.10.20. Harju Elekter toodab elektriseadmeid alates 1968. aastast. Tabelis on ettevõtte müügitulu aastatel 2008–2014, miljonit eurot[19].

1. Koostada müügitulu muutumist kirjeldav diagramm.
2. Prognoosida 2015. aasta müügitulu neljal erineval meetodil:
   (a) lineaarne trend;
   (b) viie aasta lihtne libisev keskmine;
   (c) kolme aasta kaalutud libisev keskmine kaaludega 0,5, 0,3 ja 0,2;
   (d) eksponentsilumine konstandiga 0,4.
3. Aastal 2015 oli Harju Elektri tegelik müügitulu 60,7 miljonit eurot. Milline meetod andis kõige parema prognoosi?

VASTUS lk 695.

Kompleksanalüüs, aditiivne mudel

A.10.21. Tabelis on Eesti meretranspordi ettevõtete sõitjate vedu perioodil 2005 I kvartal kuni 2013 IV kvartal[20]. Sõitjate arv on tuhandetes. Kasutades andmeid aastatest 2005–2012, teha aegrea kompleksanalüüs (trendi ja sesoonse komponendi eraldamine). Tulemusi kasutada 2013. aasta sõitjate arvu prognoosimiseks ning võrrelda tehtud prognoosi 2013. aasta tegelike andmetega.

1. Leida sobiva kujuga trendijoon. Valida lineaarse, eksponentsiaalse ja logaritmilise trendijoone vahel.
2. Leida trendi väärtused.
3. Kasutades aditiivset mudelit, eraldada sesoonne komponent.
4. Leida kvartalite keskmised sesoonsed komponendid.
5. Prognoosida sõitjate arvu 2013. aasta I–IV kvartalis, kasutades trendi ja keskmisi sesoonseid komponente.
6. Leida prognoosi keskmine ruutviga MSE.
7. Konstrueerida diagramm, kus on toodud sõitjate arv aastatel 2005–2012, 2013. aasta prognoos ja tegelikud väärtused.

VASTUS lk 695.

A.10.22. Tabelis on ühe ettevõtte käive perioodil jaanuar 1997 – juuli

[19] Allikas: Harju Elekter, majandusaruanded 2011–2015 `http://www.harjuelekter.ee/et/content/aruanded`

[20] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel TS101: sõitjatevedu transpordiliigi järgi (kvartalid).

Multiplikatiivne mudel

2001, tuhat eurot. Teha aegrea kompleksanalüüs (trendi ja sesoonse komponendi eraldamine), kasutades andmeid aastatest 1997–2000. Tulemusi kasutada 2001. aasta käibe prognoosimiseks ning võrrelda tehtud prognoosi 2001. aasta tegelike andmetega.

1. Leida sobiva kujuga trendijoon. Valida lineaarse, eksponentsiaalse ja logaritmilise trendijoone vahel.
2. Eraldada sesoonne komponent, kasutades multiplikatiivset mudelit.
3. Leida keskmised sesoonsed komponendid.
4. Prognoosida käibe väärtusi 2001. aasta juulini, kasutades trendi ja keskmist sesoonset komponenti.
5. Leida 2001. a pronoosi keskmine ruutviga MSE.
6. Konstrueerida diagramm, kus on toodud nii käibe väärtused aastatel 1997–2000, 2001. a prognoos ja 2001. aasta tegelikud väärtused.

VASTUS lk 696.

Eksponentsilumine trendi ja sesoonsusega

A.10.23. Ülesandes A.10.21 analüüsiti Eesti meretranspordi ettevõtete sõitjate veo andmeid perioodil 2005 I kvartal kuni 2012 IV kvartal, kasutades regressioonanalüüsi abil leitud trendi ja aditiivset sesoonsust. Saadud mudeli järgi tehti prognoos aastaks 2013. Nüüd siluda samu andmeid eksponentsiaalselt trendi ja sesoonsusega, kasutades aditiivset mudelit.

1. Leida silutud väärtused, trend ja sesoonne komponent kuni 2012. aasta IV kvartalini. Silumiskonstandid võtta järgmised: $w = 0{,}1$, $v = 0{,}5$, $\alpha = 0{,}6$.
2. Leida ühesammulised prognoosid ja prognoosivead 2006. aasta III kuni 2012. aasta IV kvartalini.
3. Arvutada prognoosi keskmine ruutviga MSE.
4. Minimeerida keskmist ruutviga MSE, kasutades Exceli lahendajat *Solver*. Muudetavateks väärtusteks (*By Changing Variable Cells*) on kolm silumiskonstanti.
5. Leida prognoos aastaks 2013, kasutades eelmises punktis leitud silumiskonstante.
6. Leida 2013. aasta prognoosi keskmine ruutviga MSE ja võrrelda seda ülesandes A.10.21 leitud prognoosi ruutkeskmise veaga.
7. Kumb meetod sobib selle aegrea prognoosimiseks rohkem?

VASTUS lk 696.

A.10.24. Ülesandes A.10.22 analüüsiti ühe ettevõtte käivet perioodil jaanuar 1997 – detsember 2000, kasutades regressioonanalüüsi abil leitud trendi ja multiplikatiivset sesoonsust. Saadud mudeli järgi tehti prognoos 2001. aasta juulini. Nüüd siluda sama aegrida eksponentsiaalselt trendi ja sesoonsusega, kasutades multiplikatiivset mudelit.

1. Leida silutud väärtused, trend ja sesoonne komponent kuni 2000. aasta detsembrini. Silumiskonstandid võtta järgmised: $w = 0{,}2$, $v = 0{,}1$, $\alpha = 0{,}7$.
2. Leida ühesammulised prognoosid ja prognoosivead märtsist 1998 kuni detsembrini 2000.
3. Arvutada prognoosi keskmine ruutviga MSE.
4. Minimeerida keskmist ruutviga MSE, kasutades Exceli lahendajat *Solver*. Muudetavateks väärtusteks (*By Changing Variable Cells*) on kolm silumiskonstanti.
5. Leida prognoos 2001. aasta juulini, kasutades eelmises punktis leitud silumiskonstante.
6. Leida 2001. aasta prognoosi keskmine ruutviga MSE ja võrrelda seda ülesandes A.10.22 leitud prognoosi ruutkeskmise veaga.
7. Kumb meetod sobib selle aegrea prognoosimiseks rohkem?

VASTUS lk 696.

Prognooside hindamine

A.10.25. Näites 10.14 kasutati multiplikatiivset mudelit lennureisijate arvu prognoosimiseks USA rahvusvahelistel lennuliinidel aastaks 1956. Leida prognoosi keskmine ruutviga MSE, juuritud keskmine ruutviga $RMSE$ ja keskmine suhteline absoluutviga $MAPE$. VASTUS lk 697.

A.10.26. Vaida Pilinkienė analüüsis kvartaalset nõudlust Leedu mööbliturul aastatel 1998–2006. Ta kasutas libisevat keskmist sammuga 3, 5 ja 7, eksponentsilumist silumiskonstandiga 0,3, 0,5 ja 0,8 ning lineaarset regressiooni. Kõikide meetodite korral tegi ta prognoosi aastaks 2007 (Pilinkiene, 2008). Leida iga prognoosimismeetodi korral keskmine suhteline absoluutviga $MAPE$ ning otsustada selle järgi, milline meetod andis kõige parema tulemuse. VASTUS lk 697.

Peatükk 11

Indeksid

Jalatsipoe käive vähenes eelmise aastaga võrreldes 10%. Millest oli see tingitud? Kas inimesed ostsid jalatseid vähem või ostsid nad sama palju, aga odavamaid? Tööstusettevõtte kulud suurenesid eelmise aastaga võrreldes 5%. Mis selle tingis? Kas suurenes tootmismaht või suurenesid hoopis tooraine, energia ja tööjõu hinnad? Võib-olla tootmismaht hoopis langes?

Selliste nihete ja muutuste uurimiseks, mis pole empiiriliselt küllalt täpselt määratud ja jääksid muidu analüüsija pilgu eest varju, kasutatakse mitmesuguseid indekseid. Indeksite abil analüüsitakse muutuva suuruse dünaamikat.

11.1. Indeksi mõiste, rakendusalad ja liigitus

Eesti keeles on sõnal „indeks“ palju erinevaid tähendusi. Näiteks matemaatikas tähendab see mingile arvule madalamale või kõrgemale väiksemas kirjas lisatavat arvu või tähtmärki. Raamatu lõpus olev indeks tähendab märksõnade loendit. Majanduses on see suhtarv, mis väljendab mingi majandussuuruse muutumist.

Indeksi mõiste

Indeks (*index number*) on kahe arvu suhe, mis on leitud spetsiaalse metoodika järgi ja mis iseloomustab mingi majandusalase suuruse muutumist ajas.

Suure panuse indeksmeetodi rakendamisele majandusanalüüsis on andnud Eesti majandusteadlase akadeemik Uno Mereste (1928–2009) tööd.

Indeksite **rakendusalad** võib jagada kahte suurde rühma:

- rahvamajanduse, maailmamajanduse analüüs. Nende indeksite konstrueerimise ja avaldamisega tegelevad peamiselt riiklikud ja

rahvusvahelised statistikaorganisatsioonid. Statistilistes kogumikes avaldatud indeksid on oluliseks informatsiooniallikaks makromajanduse analüüsimisel. Näiteks tarbijahinnaindeks, tootjahinnaindeks, reaalpalgaindeks;
- ettevõttesisene analüüs. Neid indekseid konstrueeritakse ettevõtte majandusliku või finantstegevuse üksikasjalikul uurimisel, sel juhul on kõik lähteandmed uurijale teada. Näiteks käibe indeks, kulude indeks.

Indeksite **liigitamine** võib toimuda mitmel erineval viisil.

Indeksite liigitus

- Objekti struktuuri järgi:
  - individuaalindeksid (ühelaadsed komponendid, algkogum);
  - üldindeksid (erinevad komponendid, liitkogum).
- Valemi kuju järgi:
  - lihtindeksid (kahe arvu suhe);
  - liitindeksid (agregaatindeksid või lihtindeksite keskmised).
- Baasperioodi alusel:
  - alusindeksid;
  - ahelindeksid.
- Tunnetusliku funktsiooni alusel:
  - kirjeldavad (kõik individuaalindeksid);
  - üldistav-analüütilised.

Järgnevalt vaatleme neid kõiki põhjalikumalt. Indeksitega tegelemisel puutume kokku kvantitatiivsete ja kvalitatiivsete majandusalaste suurustega. **Kvantitatiivsed** suurused — neid saab summeerida. Näiteks toodete hulk, tööpäevade arv, palgakulu. **Kvalitatiivsed** suurused — neid ei saa summeerida. Näiteks toote hind, tööjõu tootlikkus, seadmete jõudlus.

11.2. Alusindeks ja ahelindeks

Indeksi leidmisel on baasperioodiks see periood, mille suhtes muutus leitakse. Baasperioodi väärtust nimetatakse baasväärtuseks. Sõltuvalt baasperioodi valikust liigitatakse indeksid alus- ja ahelindeksiteks.

Alusindeks

Alusindeks (*fixed base index*) on indeks mingi kindla baasväärtuse suhtes teatud ajamomendil (või perioodil):

$$i^{alus} = \frac{y_t}{y_0}, \tag{11.1}$$

kus y_t on suuruse y väärtus ajamomendil või perioodil t ja **baasväärtus** y_0 on selle suuruse väärtus baasiks võetud ajamomendil (perioodil).

Näide 11.1. Sealiha ostukoguse ja hinna alusindeks

Tabelis on toodud sealiha keskmine ostukogus leibkonnaliikme kohta ja sealiha hind aastatel 2002–2007[a].
Et võrrelda nende suuruste dünaamikat, tuleb leida alusindeksid. Sest kilogrammides mõõdetud kogust ja eurodes mõõdetud hinda ühes ja samas teljestikus esitada ei saa. Baasperioodiks on 2002. aasta ning selle aasta väärtused on baasväärtused. Indeksid on ümardatud. Alusindeksite muutus on esitatud joonisel 11.1.

	2002	2003	2004	2005	2006	2007
Sealiha ostukogus, kg	0,70	0,82	0,85	0,79	0,72	0,70
Koguse alusindeks	$\frac{0,7}{0,7}=1$	$\frac{0,82}{0,7}\approx 1,17$	$\frac{0,85}{0,7}\approx 1,21$	1,13	1,03	1,00
Sealiha hind, €/kg	2,71	2,34	2,41	2,72	2,97	3,29
Hinna alusindeks	$\frac{2,71}{2,71}=1$	$\frac{2,34}{2,71}\approx 0,86$	$\frac{2,41}{2,71}\approx 0,89$	1,00	1,10	1,21

Joonis 11.1. Sealiha ostukoguse ja hinna alusindeksid. On näha, et kui hind langeb, siis ostukogus tõuseb ja vastupidi

[a] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel LE13: toidukaupade keskmine ostukogus ja hind leibkonnaliikme kohta kuus (1996–2007).

Alusindeks on ühikuta suhtarv ning seda nimetatakse ka **aluskasvutempoks**. Indeksi interpreteerimiseks ning kasvu või kahanemise määramiseks tuleb seda võrrelda ühega.

Alusindeksi interpreteerimine

Alusindeksi interpreteerimine:

$i^{alus} > 1$ baasperioodiga võrreldes toimus kasvamine;
$i^{alus} = 1$ baasperioodiga võrreldes jäi väärtus samaks;
$i^{alus} < 1$ baasperioodiga võrreldes toimus kahanemine.

Mõnikord korrutatakse indeksi arvutamisel suhe y_t/y_0 läbi arvuga 100 ning sellisel juhul on indeksi väärtus baasperioodil 100. Eesti Statistikaamet avaldab tarbijahinnaindeksit ning tootjahinnaindeksit just sellisel kujul.

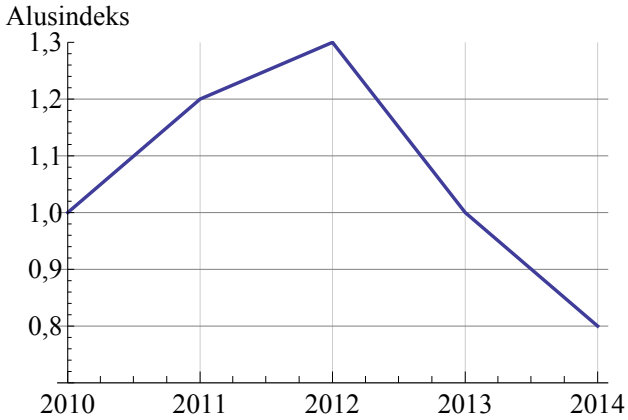
Baasi muutmine

Pikkade aegridade korral on mõnikord otstarbekas baasi muuta, sest indeksi väärtused lähevad liiga suureks. Indeksi väärtuse leidmiseks uue baasi korral jagatakse vanad väärtused baasperioodil olnud vana väärtusega. Tabelis 11.1 on baasaastalt 1970 üle mindud baasaastale 2005 ja selleks jagatakse kõik väärtused läbi 2005. aasta väärtusega.

Tabel 11.1. Baasi muutmine

	2005	2006	2007	2008	2009
Vana indeks, 1970. aastal 1	9,1	10,3	11,5	13,4	12,1
Uus indeks, 2005. aastal 1	$\frac{9,1}{9,1} = 1$	$\frac{10,3}{9,1} \approx 1,13$	$\frac{11,5}{9,1} \approx 1,26$	1,47	1,33

Kuidas leida alusindeksi graafikult protsentuaalset muutust? Analüüsime graafikut, mis on toodud joonisel 11.2.



Joonis 11.2. Suuruse Y alusindeks

1. Mitu protsenti oli suurus Y 2012. aastal suurem kui 2010. aastal? 2012. aastal oli alusindeksi väärtus 1,3, järelikult 30% suurem.
2. Mitu protsenti kasvas Y 2012. aastal, võrreldes aastaga 2011? Aastal 2012 oli alusindeks 1,3 ja aastal 2011 oli see 1,2. Järelikult, indeks 2011. aasta suhtes on

$$\frac{1{,}3}{1{,}2} = 1{,}083$$

ning kasv oli 8,3%.

Kui soovime alusindeksi põhjal leida kasvu mingi teise, baasaastast erineva aasta suhtes, tuleb see teine aasta võtta baasaastaks ja leida indeks selle suhtes.

Ahelindeks (*chain index*) on suuruse y väärtus ajahetkel (ajaperioodil) t jagatud väärtusega eelmisel ajahetkel (ajaperioodil) $t-1$:

$$i^{ahel} = \frac{y_t}{y_{t-1}}. \tag{11.2}$$

Ahelindeks

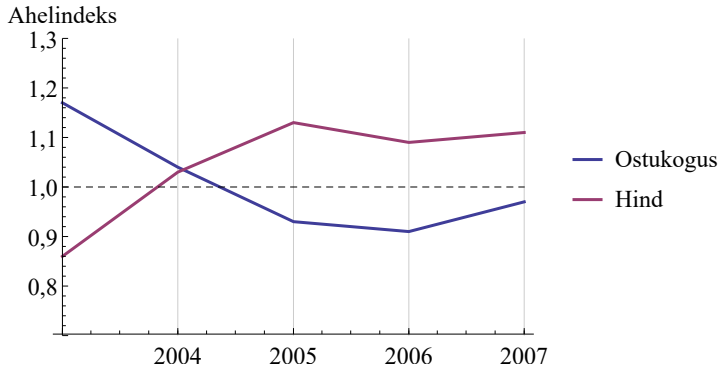
Ahelindeksit nimetatakse väga tihti ka **kasvutempoks**.

Näide 11.2. Sealiha ostukoguse ja hinna ahelindeksid

Leiame näites 11.1 toodud andmete põhjal sealiha ostukoguse ja hinna ahelindeksid perioodil 2003–2007.

	2002	2003	2004	2005	2006	2007
Sealiha ostukogus, kg	0,70	0,82	0,85	0,79	0,72	0,70
Koguse ahelindeks		$\frac{0{,}82}{0{,}7} \approx 1{,}17$	$\frac{0{,}85}{0{,}82} \approx 1{,}04$	0,93	0,91	0,97
Sealiha hind, €/kg	2,71	2,34	2,41	2,72	2,97	3,29
Hinna ahelindeks		$\frac{2{,}34}{2{,}71} \approx 0{,}86$	$\frac{2{,}41}{2{,}34} \approx 1{,}03$	1,13	1,09	1,11

Ahelindeksite dünaamika on esitatud joonisel 11.3. Graafikule on lisatud väärtusele 1 vastav kriipsjoon. Kui indeksi väärtus on sellest allpool (väiksem kui 1), siis vastav suurus sellel perioodil kahanes. Kui indeksi väärtus on sellest joonest kõrgemal (suurem kui 1), siis vastav väärtus kasvas. Aastatel 2005–2007 on hind kogu aeg kasvanud ja ostukogus kahanenud. Kogus kasvas aastatel 2003 ja 2004. Hind on kahanenud ainult aastal 2003.



Joonis 11.3. Sealiha ostukoguse ja hinna ahelindeksid

Ahelindeksi interpreteerimine

Ahelindeksi interpreteerimine:

$i^{ahel} > 1$ eelmise perioodiga võrreldes toimus kasvamine;
$i^{ahel} = 1$ eelmise perioodiga võrreldes jäi väärtus samaks;
$i^{ahel} < 1$ eelmise perioodiga võrreldes toimus kahanemine.

Alus- ja ahelindeksid on omavahel seotud. Järgmise perioodi alusindeksi saame, kui eelmise perioodi alusindeksi korrutame läbi ahelindeksiga.

Näide 11.3. Sealiha koguse ahel- ja alusindeks

Avaldame näites 11.1 leitud sealiha koguse alusindeksid näites 11.2 leitud ahelindeksite kaudu.

Aasta	Koguse ahelindeks	Koguse alusindeks
2002		1
2003	1,17	$1{,}17 = 1 \cdot 1{,}17$
2004	1,04	$1{,}21 = 1{,}17 \cdot 1{,}04$
2005	0,93	$1{,}13 = 1{,}21 \cdot 0{,}93 = 1{,}17 \cdot 1{,}04 \cdot 0{,}93$
2006	0,91	$1{,}03 = 1{,}13 \cdot 0{,}91 = 1{,}17 \cdot 1{,}04 \cdot 0{,}93 \cdot 0{,}91$

Üldiselt saame m-nda perioodi alusindeksi järjestikuste ahelindeksite korrutamisel:

$$i_m^{alus} = i_0^{alus} \cdot i_1^{ahel} \cdot i_2^{ahel} \cdot \ldots \cdot i_m^{ahel}. \tag{11.3}$$

11.3. Individuaalindeksid ja üldindeksid

Kogumit, mis koosneb ühelaadsetest elementidest, nimetatakse **algkogumiks**. Näiteks ühte sorti kaubad, üht ja sama tööd tegevad töötajad, ühe ettevõtte aktsiad.

Individuaalindeksiga (*simple index number*) väljendatakse algkogumi või kvalitatiivse üksiktunnuse ajalist muutumist.

Mingi tooteliigi toodangu mahu indeks (kvantitatiivse suuruse indeks):

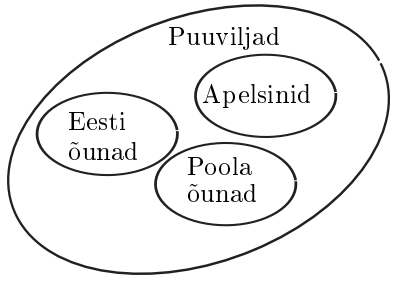
$$i_q = \frac{q_1}{q_0}, \tag{11.4}$$

kus q_1 on toodangu maht aruandeperioodil ja q_0 on toodangu maht baasperioodil.
Toote hinna (kvalitatiivse suuruse) muutumise indeks:

$$i_p = \frac{p_1}{p_0}, \tag{11.5}$$

kus p_1 on toote hind aruandeperioodil ja p_0 hind baasperioodil.

Erinevaid algkogumeid ühendades saadakse **liitkogum** (joonis 11.4). Liitkogumiteks on näiteks mitut toodet väljastava tootja tootmismaht, tööstusharu kogutoodang, kaubagrupp. Liitkogumite analüüsimisel kasutatakse **üldindekseid**.



Joonis 11.4. Liitkogum „puuviljad“ moodustub ühelaadsetest algkogumitest

Üldindeks (*composite index number*) on dünaamika suhtarv, millega väljendatakse ebaühtlase koostisega kogumi suhtelist ajalist muutumist.

Üldindeks

Näiteid üldindeksitest:

- tarbijahinnaindeks koosneb järgmistest komponentidest: toit, alkohoolsed joogid ja tubakatooted, riietus ja jalatsid, eluase, majapidamine, tervishoid, transport ja side, vaba aeg, mitmesugused kaubad ja teenused;
- tootjahinnaindeksisse kuuluvad energeetika, mäetööstuse ja töötleva tööstuse hinnaindeksid;
- börsiindeksisse kuuluvad erinevate ettevõtete aktsiad.

Ühelaadsed komponendid → algkogum → **individuaalindeks**.
Erinevad komponendid → liitkogum → **üldindeks**.

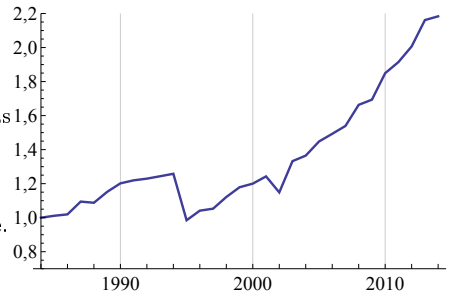
Näide 11.4. Jõulukingi indeks

Alates aastast 1984 arvutatakse nn jõulukingi indeksit (*Christmas Price Index*, CPI), mis põhineb tuntud jõululaulul „The Twelve Days of Christmas". Laulus saadab kallim esimesel päeval nurmkana, teisel päeval nurmkana ja kaks turteltuvi, kolmandal päeval nurmkana, kaks turteltuvi ja kolm prantsuse kana jne.
Indeksi arvutamisel leitakse jõulukingi kogumaksumus jooksva aasta hindades. Kui 1984. aastal maksid kõik laulus üles loetletud asjad kokku 12 673,56 $, siis 2014. aastal 27 673,22 $[a]. Indeksi lühend CPI vastab ingliskeelsele tarbijahinnaindeksi lühendile (*Consumer Price Index*). Graafikul on jõulukingi hinnaindeksi muutus 1984–2014.

On the first day of Christmas
my true love sent to me:
A Partridge in a Pear Tree.

On the second day of Christmas
my true love sent to me:
2 Turtle Doves
And a Partridge in a Pear Tree.

On the third day of Christmas
my true love sent to me:
3 French Hens,
2 Turtle Doves
And a Partridge in a Pear Tree....



[a] Allikas: PNC Christmas Price Index `http://www.pncchristmaspriceindex.com`

11.4. Keskmised indeksid

Keskmisi indekseid kasutatakse juhul, kui on teada üksikute komponentide indeksid.

Oletame, et jaanuaris oli Eesti õunte hinnaindeks 1,3 ja Poola õunte hinnaindeks 1,1 (lihtsuse mõttes vaatleme siin ainult kahte erinevat sorti õunu). Eesti õunu on müügil vähe, sel aastaajal ainult 10% õunte kogumüügist. Ülejäänud 90% õunte müügist moodustavad Poola õunad. Kui suur oli jaanuaris kaubagrupi „õunad“ hinnaindeks?

Üks võimalus on leida nende kahe indeksi lihtne aritmeetiline keskmine:

$$\frac{1{,}3 + 1{,}1}{2} = 1{,}2,$$

mis tähendaks et õunte hind tõusis keskmiselt 20%. Aga kuna Poola õunte osakaal õunte kogumüügis on oluliselt suurem kui Eesti õunte osakaal, kas siis on õige keskmise arvutamisel arvestada mõlema kauba indeksit ühesuguse kaaluga? Ilmselt ei ole. Kaup, mida müüakse vähe, peaks hinnaindeksit vähem mõjutama. Järelikult tuleb kasutada kaalutud aritmeetilist keskmist, kus liitkogumi erinevatel komponentidel võivad olla erinevad kaalud.

Keskmine indeks leitakse individuaalindeksite kaalutud aritmeetilise keskmisena (vt valem (2.6)):

$$i = \frac{\sum_k f_k i_k}{\sum_k f_k}, \tag{11.6}$$

kus f_k on k-nda komponendi kaal ja i_k on i-nda komponendi indeks. Summeerimine toimub üle kõigi komponentide.

Näiteid kaaludest:

- hinnaindeksi korral on kaaluks tavaliselt maht või kulutused;
- mahuindeksi korral on kaaluks tavaliselt hind või kulutused;
- tootlikkuse (tootmismaht töötaja kohta) indeksi korral on kaaluks tavaliselt töötajate arv.

Näide 11.5. Keskmise omahinna indeks

Ettevõte toodab nelja erinevat toodet: A, B, C ja D. Võrreldes eelneva aastaga, on kõikide toodete omahind tõusnud. Keskmise omahinna indeksi leidmiseks kasutatakse kaalutud keskmist, kus kaaluks on vastava toote tootmismaht aastas.

Toode	Omahinna muutus	Individuaal-indeks i_k	Tootmismaht ehk kaal f_k	Korrutised $f_k i_k$
A	10%	1,1	4100	4510
B	20%	1,2	5300	6360
C	20%	1,2	3000	3600
D	10%	1,1	1200	1320
KOKKU			13600	15790

Keskmise omahinna indeks

$$i = \frac{15790}{13600} \approx 1{,}16.$$

Kui on teada iga komponendi osakaal liitkogumis $w_k = f_k / \sum_i f_i$, siis üldindeksi valem on sama, mis osakaalude abil leitud kaalutud aritmeetiline keskmine (2.7):

$$i = \sum_k w_k i_k. \tag{11.7}$$

Niimoodi arvutatakse näiteks tarbijahinnaindeksit THI. Leitakse erinevate kaubagruppide hinnaindeksid i_k ning kaubagruppide osakaalud kogutarbimises w_k. Kuna inimeste tarbimisharjumused muutuvad, tuleb neid osakaalusid pidevalt uuendada ja alates 2001. aastast tehakse seda igal aastal. Kaalusüsteemi uuendamisel lähtutakse leibkondade hulgas läbiviidud uuringutest. Lisaks uuendatakse ka arvutuste aluseks olevaid baaskaupu.

Näide 11.6. Kaubagruppide oskaalud tarbijahinnaindeksis

Tabelis on toodud erinevate kaubagruppide osakaal tarbijahinnaindeksis aastatel 2007 ja 2011 ning osakaalude muutus[a]. Osakaalud on antud promillides. On näha, et oluliselt on suurenenud eluasemele ja toidule tehtavate kulutuste osakaal. See on tingitud sellest, et majanduslanguse ajal inimeste sissetulekud vähenesid ja tuli vähendada ka kulutusi. Kuna toidu ja eluaseme pealt eriti kokku hoida ei saanud, siis vähendati muid kulusid. Eriti palju on vähenenud kulutused transpordile, mis tähendab, et vähendati autoga sõitmist.

	Osakaal, ‰		
	2007	2011	Muutus, ‰
KOKKU	1 000	1 000	
Toit ja mittealkohoolsed joogid	217,0	242,1	25,1
Alkohoolsed joogid ja tubakas	71,7	77,7	6,0
Riietus ja jalatsid	67,2	54,1	−13,1
Eluase	144,2	177,3	33,1
Majapidamine	56,1	44,9	−11,2
Tervishoid	37,8	41,2	3,4
Transport	152,1	132,2	−19,9
Side	49,2	49,0	−0,2
Vaba aeg	87,0	81,3	−5,7
Haridus ja lasteasutused	19,3	17,2	−2,1
Söömine väljaspool kodu, majutus	43,1	31,6	−11,5
Mitmesugused kaubad ja teenused	55,3	51,4	−3,9

[a] Allikas: Eesti Statistikaamet

11.5. Ühismõõdustamine ja agregeerimine

Üldindeksid jagunevad kaheks:

- **koondindeksid** väljendavad korraga mitme teguri muutumist;
- **teguriindeksid** ainult ühe teguri muutumist, teiste tegurite mõju on elimineeritud:
  - kvantitatiivse teguri indeksid;
  - kvalitatiivse teguri indeksid.

Ühismõõdustamine

Ebaühtlase kogumi mahu väljendamiseks tuleb osakogumid ühismõõdustada, s.t avaldada ühtsetes mõõtühikutes (näiteks rahalistes ühikutes). **Ühismõõdustamine** on indekseeritava suuruse läbikorrutamine ühismõõdustajaga. Tulemuseks peab olema korrutis, millel on majanduslik tähendus. Ühismõõdustamise käigus uuritava suuruse maht (näiteks toodangu naturaalne maht) väljendatakse teise suuruse mahu (näiteks toodangu maksumus) kaudu.

Agregeerimine

Agregeerimine on ühismõõdustatud suuruste ühendamine. Üldiselt tähendab agregatsioon millegi kokku kogunemist või kuhjumist. Näiteks ettevõtete esitatud statistiliste aruannete põhjal agregeerib Eesti Statistikaamet erinevate ettevõtete kogutoodangu, mille tulemusel saadakse tööstusharu kogutoodang. Erinevate harude kogutoodangute agregeerimisel saadakse sisemajanduse koguprodukt SKP.

Näide 11.7. Ühismõõdustamine ja agregeerimine mööblivabrikus

Mööblivabrikus toodetakse erinevaid mööbliesemeid. Kogutoodangu leidmiseks ei ole mõttekas kokku liita kuu aja jooksul toodetud toole, laudasid ja diivaneid, sest nende tegemine nõuab erinevas mahus aega ja materjali. Erinevate mööbliesemete kogused ühismõõdustatakse hinnaga. Saadakse maksumused, mis seejärel agregeeritakse.

Toode	Kogus, tk	Hind, € (ühismõõdustaja)	Maksumus, € (ühismõõdustatud)
Toolid	600	10	6 000
Lauad	500	20	10 000
Diivanid	50	100	5 000
	Agregeerimine	KOKKU	21 000

Ühismõõdustatud suuruste agregeerimisel saadakse **agregaatsummad**. Näiteks eri toodete maksumuste liitmisel saadakse kaks avaldist (k loendab tooteid):

$$\text{agregeeritud maksumus baasperioodil} \qquad Q_0 = \sum_k p_{0,k} q_{0,k}, \tag{11.8}$$

$$\text{agregeeritud maksumus aruandeperioodil} \qquad Q_1 = \sum_k p_{1,k} q_{1,k}, \tag{11.9}$$

kus p tähistab hinda ja q kogust. Nende maksumuste põhjal leiame toodangu maksumuse üldindeksi:

$$I_Q = \frac{Q_1}{Q_0} = \frac{\sum_k p_{1,k} q_{1,k}}{\sum_k p_{0,k} q_{0,k}}. \tag{11.10}$$

Valemite lihtsustamiseks summeerimisindeksit k edaspidi välja ei kirjutata. Arvestame, et summeerimine tähendab agregeerimist, s.t toimub alati üle liitkogumi üksikute komponentide. Toodangu maksumuse üldindeksi valem on siis järgmine:

$$I_Q = \frac{\sum p_1 q_1}{\sum p_0 q_0}. \tag{11.11}$$

Näide 11.8. Mööblivabriku kogutoodangu kogumuutus

Näites 11.7 toodud mööblivabrikus muutusid järgmisel perioodil nii toodete hinnad kui ka kogused. Leiame toodangu maksumuse muutumist kirjeldava indeksi.

Toode	Baaasperiood 0			Aruandeperiood 1		
	Kogus q_0	Hind p_0	p_0q_0	Kogus q_1	Hind p_1	p_1q_1
Toolid	600	10	6 000	800	11	8 800
Lauad	500	20	10 000	500	20	10 000
Diivanid	50	100	5 000	45	120	5 400
KOKKU			21 000			24 200

Tabelis on leitud toodangu maksumus baasperioodil:

$$Q_0 = \sum p_0q_0 = 21\ 000 \tag{11.12}$$

ja aruandeperioodil

$$Q_1 = \sum p_1q_1 = 24\ 200. \tag{11.13}$$

Maksumuse indeks

$$I_Q = \frac{\sum p_1q_1}{\sum p_0q_0} = \frac{Q_1}{Q_0} = \frac{24200}{21000} \approx 1{,}15. \tag{11.14}$$

Mööblivabriku kogutoodangu maksumuse indeks on 1,15, järelikult kogutoodang kasvas 15%.

11.6. Koondindeks ja teguriindeksid

Teades vaid maksumuse üldindeksit (11.11), ei ole võimalik kindlaks teha, kui suur osa muutusest oli põhjustatud hindade p muutusest ja kui suur osa oli põhjustatud toodangu mahu q muutustest. Muutuda võivad nii hinnad ($p_0 \to p_1$) kui ka kogused ($q_0 \to q_1$). Mööblivabriku näites 11.8 kasvas nii toolide kui ka diivanite hind. Samuti kasvas toolide kogus, kuid diivanite kogus kahanes. Maksumuse indeks 1,15 (valem (11.14)) näitab kõigist neist teguritest põhjustatud kogumuutust.

Muutuste eraldamiseks moodustatakse murrust (11.11) kahe murru korrutis:

$$I_Q = \frac{\sum p_1q_1}{\sum p_0q_0} = \frac{\sum p_1q_1}{\sum p_0q_1} \cdot \frac{\sum p_0q_1}{\sum p_0q_0}. \tag{11.15}$$

Näeme, et murdudes, mida me korrutame, muutub korraga ainult üks tegur: esimeses murrus $\sum p_1q_1/\sum p_0q_1$ on erinevad hinnad p_0 ja p_1 ning kogus q_1 on sama. Teises murrus $\sum p_0q_1/\sum p_0q_0$ on erinevad kogused q_1 ja q_0, aga hind p_0 on sama. Sellega oleme koguste ja hindade muutuse eraldanud.

Kuna maksumus Q, mille indeksit me analüüsime, on kahe teguri korrutis $Q = pq$, siis räägitakse teguriindeksitest. Muutuste eraldamiseks püstitatakse hüpotees, et **korraga muutub vaid üks tegur**, kas p või q.

Teguriindeks mõõdab kahest koos toimivast tegurist ainult ühe muutumist.

Valemis (11.15) on meil kolm murdu: vasakul maksumuse koondindeks, paremal kahe murru $\sum p_1q_1/\sum p_0q_1$ ja $\sum p_0q_1/\sum p_0q_0$ korrutis. Kummalgi murrul on oma tähendus: need on teguriindeksid.

Hinnaindeks (kvalitatiivse teguri indeks) näitab, kui palju oleks maksumus muutunud, kui muutunud oleksid ainult hinnad:

$$I_p = \frac{\sum p_1q_1}{\sum p_0q_1}. \tag{11.16}$$

Koguseindeks (kvantitatiivse teguri indeks) näitab, kui palju oleks maksumus muutunud, kui muutunud oleksid ainult kogused:

$$I_q = \frac{\sum p_0q_1}{\sum p_0q_0}. \tag{11.17}$$

Koondindeks on kvalitatiivse teguri ja kvantitatiivse teguri indeksite korrutis

$$I_Q = I_p \cdot I_q. \tag{11.18}$$

Koondindeks näitab nii hindade kui koguste muutumisest tingitud kogumaksumuse muutust.

Korrutis (11.18) tuleb valemist (11.15). Need kolm indeksit moodustavad teguriindeksite süsteemi ja paneme need kolm koos uuesti kirja.

Teguriindeksite süsteem, I versioon:

$$I_Q = \frac{\sum p_1 q_1}{\sum p_0 q_0}, \tag{11.19}$$

Teguriindeksite süsteem

$$I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1} \quad \text{kvantitatiivse teguri aruandeperioodi väärtused } q_1, \tag{11.20}$$

$$I_q = \frac{\sum p_0 q_1}{\sum p_0 q_0} \quad \text{kvalitatiivse teguri baasperioodi väärtused } p_0. \tag{11.21}$$

Teguriindeksite süsteemi I versiooni korral kasutatakse nn **Satunovski reeglit**:

- kvalitatiivne muutuv suurus ühismõõdustatakse **kvantitatiivse** teguri **aruandeperioodi** väärtustega;
- kvantitatiivne muutuv suurus ühismõõdustatakse **kvalitatiivse** teguri **baasperioodi** väärtustega.

Hinnaindeksi I_p nimetajas ja koguseindeksi I_q lugejas on summa $\sum p_0 q_1$. Milline on selle tõlgendus?

Tinglik maksumus on maksumus, mis oleks esinenud aruandeperioodil, kui hinnad oleksid baasperioodi omad (ei oleks muutunud):

$$Q_{tingl} = \sum p_0 q_1. \tag{11.22}$$

Tinglik maksumus

Praktilistes arvutustes ei ole mõistlik lähtuda otseselt teguriindeksite süsteemi definitsioonvalemitest (11.19)–(11.21). Paneme tähele, et teguriindeksite süsteemi kõigi kolme indeksi arvutamiseks on vaja leida kolm summat:

Praktilised arvutused

aruandeperioodi maksumus	$Q_1 = \sum p_1 q_1,$	(11.23)
baasperioodi maksumus	$Q_0 = \sum p_0 q_0,$	(11.24)
tinglik maksumus	$Q_{tingl} = \sum p_0 q_1.$	(11.25)

Nende kaudu teguriindeksite süsteemi indeksid:

$$I_Q = \frac{O_1}{Q_0}, \tag{11.26}$$

$$I_p = \frac{Q_1}{Q_{tingl}}, \tag{11.27}$$

$$I_q = \frac{Q_{tingl}}{Q_0}. \tag{11.28}$$

Näide 11.9. Mööblivabriku kogutoodangu osamuutused

Näites 11.8 leidsime, et mööblivabriku kogutoodangu maksumuse koondindeks oli 1,15, s.t kogutoodang kasvas 15%. Kui suur osa sellest muutusest oli tingitud hindade muutumisest ja kui suur osa toodete arvu muutumisest? Nende osamuutuste arvutamiseks peame leidma hinnaindeksi I_p (11.16) ja koguseindeksi I_q (11.17). Arvutamiseks kasutame valemeid (11.27) ja (11.28). Paneme tähele, et näites 11.8 on meil leitud maksumus aruandeperioodil $Q_1 = 21\ 000$ (valem (11.12)) ja maksumus baasperioodil $Q_0 = 24\ 200$ (valem (11.13)). Lisaks tuleb vaid arvutada tinglik maksumus Q_{tingl} valemist (11.25).

Toode	Baasperiood 0 Hind p_0	Aruandeperiood 1 Kogus q_1	p_0q_1
Toolid	10	800	8 000
Lauad	20	500	10 000
Diivanid	100	45	4 500
KOKKU			22 500

Tabelis on real KOKKU leitud tinglik maksumus:

$$Q_{tingl} = \sum p_0q_1 = 22500.$$

Nüüd on hinnaindeks

$$I_p = \frac{Q_1}{Q_{tingl}} = \frac{24200}{22500} \approx 1{,}08$$

ja koguse indeks

$$I_q = \frac{Q_{tingl}}{Q_0} = \frac{22500}{21000} \approx 1{,}07.$$

Hindade muutumisest tingituna kasvas kogutoodang 8% ja koguste muutustest põhjustatud kogutoodangu muutus oli 7%.

Indeksisüsteemid võimaldavad tuletada teadaolevate indeksite alusel teisi süsteemseid indekseid. Näiteks maksumuse koondindeksi ja mahu teguriindeksi kaudu võime avaldada hinna teguriindeksi:

$$I_p = \frac{I_Q}{I_q} = \frac{\frac{\sum p_1 q_1}{\sum p_0 q_0}}{\frac{\sum p_0 q_1}{\sum p_0 q_0}} = \frac{\sum p_1 q_1}{\sum p_0 q_1}.$$

Sõltuvalt sellest, kuidas toimub ühismõõdustamisel aruande- ja baasperioodide väärtuste kasutamine, on võimalikud kaks erinevat versiooni teguriindeksite süsteemist. Vaikimisi kasutatakse eespool toodud I versiooni, kuid on võimalik koostada ka II versioon.

Teguriindeksite süsteem, II versioon:

$$I_Q = \frac{\sum p_1 q_1}{\sum p_0 q_0}, \tag{11.29}$$

$$I_p = \frac{\sum p_1 q_0}{\sum p_0 q_0} \quad \text{kvantitatiivse teguri baasperioodi väärtused } q_0, \tag{11.30}$$

$$I_q = \frac{\sum p_1 q_1}{\sum p_1 q_0} \quad \text{kvalitatiivse teguri aruandeperioodi väärtused } p_1. \tag{11.31}$$

Konkreetsuse mõttes kasutasime teguriindeksite süsteemi tundmaõppimisel maksumust, hinda ja kogust. Analoogset teguriindeksite süsteemi võib kasutada ka teistel juhtudel, kui mingi majandussuurus on kvalitatiivse ja kvantitatiivse suuruse korrutis:

- palga P ja töötajate arvu n korrutis annab palgakulu $C = Pn$;
- tootlikkuse Y ja vastava tootlikkusega töötavate töötajate arvu n korrutis annab kogutoodangu $Q = Yn$;
- põllukultuuri saagikuse S ja külvipinna k korrutis on kogusaak: $Q = Sk$.

11.7. Muutuva ja püsiva struktuuri ning struktuurinihete indeksid

Struktuurinihete uurimise probleem kerkib üles **kvalitatiivsete** tegurite keskmiste tasemete suhtes. Näiteks

- keskmine hind;
- keskmine tootlikkus;
- keskmine palk.

Agregeerimisel saadakse kolm indeksit:

- muutuva struktuuri indeks;

- püsiva struktuuri indeks;
- struktuurinihete indeks.

Olgu meil kvalitatiivseks teguriks hind p ja kvantitatiivseks suuruseks kogus q. Keskmise hinna indeks on aruandeperioodi keskmise hinna $\bar{p}_1$ ja baasperioodi keskmise hinna $\bar{p}_0$ suhe:

$$I^{\bar{p}} = \frac{\bar{p}_1}{\bar{p}_0}. \tag{11.32}$$

Arvestame, et aruandeperioodi keskmine hind on kaalutud keskmine, kus kaaludeks on aruandeperioodi toodangumahud q_1:

$$\bar{p}_1 = \frac{\sum p_1 q_1}{\sum q_1}. \tag{11.33}$$

Baasperioodi keskmine hind on kaalutud keskmine, kus kaaludeks on baasperioodi toodangumahud q_0:

$$\bar{p}_0 = \frac{\sum p_0 q_0}{\sum q_0}. \tag{11.34}$$

Pannes valemid (11.33) ja (11.34) valemisse (11.32), saame keskmise hinna indeksi, kus muutuvad nii hind kui ka kogus.

Muutuva struktuuri indeks

Muutuva struktuuri indeks iseloomustab kvalitatiivse suuruse p keskmise muutumist, mis on tingitud nii kvantitatiivse teguri q muutustest kui ka kvalitatiivse teguri p enda muutustest:

$$I^{\bar{p}}_{m.str} = \frac{\sum p_1 q_1}{\sum q_1} : \frac{\sum p_0 q_0}{\sum q_0}. \tag{11.35}$$

Et elimineerida kvantitatiivse teguri q muutusi, kasutatakse valemis (11.35) ainult aruandeperioodi toodangumahtusid q_1 ja saadakse keskmise hinna püsiva struktuuri indeks.

Püsiva struktuuri indeks

Püsiva struktuuri indeks iseloomustab kvalitatiivse suuruse p muutumist, mis on tingitud ainult kvalitatiivse teguri enda muutustest:

$$I^{\bar{p}}_{p.str} = \frac{\sum p_1 q_1}{\sum q_1} : \frac{\sum p_0 q_1}{\sum q_1}. \tag{11.36}$$

Et keskmise hinna indeksis (11.35) elimineerida kvalitatiivse teguri p muutused, kasutatakse nüüd ainult baasperioodi hindasid p_0. Saadakse keskmise hinna struktuurinihete indeks.

Struktuurinihete indeks iseloomustab kvalitatiivse suuruse muutumist, mis on tingitud ainult kvantitatiivse teguri muutustest, s.t muutustest kogumi struktuuris:

$$I^{\bar{p}}_{str.n} = \frac{\sum p_0 q_1}{\sum q_1} : \frac{\sum p_0 q_0}{\sum q_0}. \tag{11.37}$$

Struktuurinihete indeks

Miks nimetatakse viimast struktuurinihete indeksiks? Vaatame näiteks ettevõtte keskmise omahinna kujunemist. Keskmine omahind võib alaneda kahel põhjusel: alanevad üksikute toodete omahinnad või hakatakse rohkem tootma väiksema omahinnaga tooteid. Viimane tähendab muutust tootmise struktuuris. Samamoodi on tarbimise korral. Tarbimiskulud võivad väheneda hindade languse tõttu või seetõttu, et toimuvad muutused tarbimise struktuuris: ostetakse odavamaid kaupu.

Paneme uuesti kirja kolm indeksit, mis iseloomustavad kvalitatiivse teguri (hinna) muutust ja moodustavad struktuuriindeksite süsteemi.

Struktuuriindeksite süsteem, I versioon:

Struktuuriindeksite süsteem

Muutuva struktuuri indeks
$$I^{\bar{p}}_{m.str} = \frac{\sum p_1 q_1}{\sum q_1} : \frac{\sum p_0 q_0}{\sum q_0}. \tag{11.38}$$

Püsiva struktuuri indeks
$$I^{\bar{p}}_{p.str} = \frac{\sum p_1 q_1}{\sum q_1} : \frac{\sum p_0 q_1}{\sum q_1}. \tag{11.39}$$

Struktuurinihete indeks
$$I^{\bar{p}}_{str.n} = \frac{\sum p_0 q_1}{\sum q_1} : \frac{\sum p_0 q_0}{\sum q_0}. \tag{11.40}$$

Seos struktuuriindeksite vahel:

$$I^{\bar{p}}_{m.str} = I^{\bar{p}}_{p.str} \cdot I^{\bar{p}}_{str.n}. \tag{11.41}$$

Suhet $\sum p_0 q_1 / \sum q_1$, mis esineb nii püsiva struktuuri indeksi valemis kui ka struktuurinihete indeksi valemis, nimetatakse aruandeperioodi **tinglikuks hinnaks**.

Tinglik hind

Aruandeperioodi **tinglik hind** oleks kujunenud siis, kui üksikud hinnad oleksid jäänud samaks nagu baasperioodil, kogused q aga vastavad aruandeperioodile:

$$\bar{p}_{tingl} = \frac{\sum p_0 q_1}{\sum q_1}. \tag{11.42}$$

Kasutades aruandeperioodi tinglikku hinda, võime avaldada nii püsiva struktuuri kui ka muutuva struktuuri hinna. Struktuuriindeksite arvutamisel ongi mõistlik leida algul kolm hinda:

aruandeperioodi keskmine hind $$\bar{p}_1 = \frac{\sum p_1 q_1}{\sum q_1}, \tag{11.43}$$

baasperioodi keskmine hind $$\bar{p}_0 = \frac{\sum p_0 q_0}{\sum q_0}, \tag{11.44}$$

tinglik hind $$\bar{p}_{tingl} = \frac{\sum p_0 q_1}{\sum q_1} \tag{11.45}$$

ja seejärel kolm struktuuriindeksit:

$$I^{\bar{p}}_{m.str} = \frac{\bar{p}_1}{\bar{p}_0}, \tag{11.46}$$

$$I^{\bar{p}}_{p.str} = \frac{\bar{p}_1}{\bar{p}_{tingl}}, \tag{11.47}$$

$$I^{\bar{p}}_{str.n} = \frac{\bar{p}_{tingl}}{\bar{p}_0}. \tag{11.48}$$

Näide 11.10. Mööblivabriku toodangu keskmise hinna muutuse analüüs

Leiame, kuidas muutus näites 11.8 toodud mööblivabrikus toodangu keskmine hind. Kasutame valemeid (11.43)–(11.48). Hindade leidmiseks valemitest (11.43)–(11.45) tuleb leida viis summat:

- kogused kokku baasperioodil $\sum q_0$ ja aruandeperioodil $\sum q_1$;
- maksumused kokku baasperioodil $\sum p_0 q_0$ ja aruandeperioodil $\sum p_1 q_1$;
- tinglik maksumus $\sum p_0 q_1$.

Kuigi maksumused on meil eespool toodud näidetes juba leitud, esitame siin uuesti tabeli koos andmete ja vajalike summadega real KOKKU.

Toode	p_0	q_0	p_0q_0	p_1	q_1	p_1q_1	p_0q_1
Toolid	10	600	6000	11	800	8800	8000
Lauad	20	500	10 000	20	500	10 000	10 000
Diivanid	100	50	5000	120	45	5400	4500
KOKKU		1150	21 000		1345	24 200	22 500

Leiame kolm hinda:

$$\bar{p}_1 = \frac{\sum p_1q_1}{\sum q_1} = \frac{24200}{1345} \approx 17{,}99,$$
$$\bar{p}_0 = \frac{\sum p_0q_0}{\sum q_0} = \frac{21000}{1150} \approx 18{,}26,$$
$$\bar{p}_{tingl} = \frac{\sum p_0q_1}{\sum q_1} = \frac{22500}{1345} \approx 16{,}73.$$

Nüüd arvutame välja kolm struktuuriindeksit:

$$I^{\bar{p}}_{m.str} = \frac{\bar{p}_1}{\bar{p}_0} = \frac{17{,}99}{18{,}26} \approx 0{,}985,$$
$$I^{\bar{p}}_{p.str} = \frac{\bar{p}_1}{\bar{p}_{tingl}} = \frac{17{,}99}{16{,}73} \approx 1{,}076,$$
$$I^{\bar{p}}_{str.n} = \frac{\bar{p}_{tingl}}{\bar{p}_0} = \frac{16{,}73}{18{,}26} \approx 0{,}916.$$

Keskmise hinna muutuva struktuuri indeks on 0,985, mis tähendab, et mööblivabriku toodete keskmine hind alanes 1,5%. Püsiva struktuuri indeks 1,076 näitab, et hindade muutus põhjustas keskmise hinna tõusu 7,6%. Muutuva struktuuri indeks 0,916 näitab, et muutused toodangu mahtudes põhjustasid keskmise hinna alanemise 8,4%.

Struktuuriindeksite süsteemi I versiooni korral, mida kasutatakse vaikimisi, on

- püsiva struktuuri indeksis kvantitatiivse teguri aruandeperioodi väärtused;
- struktuurinihete indeksis kvalitatiivse teguri baasperioodi väärtused.

On võimalik kasutada ka struktuuriindeksite süsteemi II versiooni:

- püsiva struktuuri indeksis kvantitatiivse teguri baasperioodi väärtused;
- struktuurinihete indeksis kvalitatiivse teguri aruandeperioodi väärtused.

Toome ära ka selle versiooni valemid.

Struktuuriindeksite süsteem, II versioon:

$$\text{muutuva struktuuri indeks} \quad I^{\bar{p}}_{m.str} = \frac{\sum p_1 q_1}{\sum q_1} : \frac{\sum p_0 q_0}{\sum q_0}, \tag{11.49}$$

$$\text{püsiva struktuuri indeks} \quad I^{\bar{p}}_{p.str} = \frac{\sum p_1 q_0}{\sum q_0} : \frac{\sum p_0 q_0}{\sum q_0}, \tag{11.50}$$

$$\text{struktuurinihete indeks} \quad I^{\bar{p}}_{str.n} = \frac{\sum p_1 q_1}{\sum q_1} : \frac{\sum p_1 q_0}{\sum q_0}. \tag{11.51}$$

Kasutades kvantitatiivse teguri osakaalusid

$$w_i = \frac{q_i}{\sum_i q_i}, \tag{11.52}$$

võib struktuuriindeksid esitada nende osakaalude abil.

Struktuuriindeksid kaaludega, I versioon:

$$\text{muutuva struktuuri indeks} \quad I^{\bar{p}}_{m.str} = \frac{\sum p_1 w_1}{\sum p_0 w_0}, \tag{11.53}$$

$$\text{püsiva struktuuri indeks} \quad I^{\bar{p}}_{p.str} = \frac{\sum p_1 w_1}{\sum p_0 w_1}, \tag{11.54}$$

$$\text{struktuurinihete indeks} \quad I^{\bar{p}}_{str.n} = \frac{\sum p_0 w_1}{\sum p_0 w_0}. \tag{11.55}$$

Struktuuriindeksid kaaludega, II versioon:

$$\text{muutuva struktuuri indeks} \quad I^{\bar{p}}_{m.str} = \frac{\sum p_1 w_1}{\sum p_0 w_0}, \tag{11.56}$$

$$\text{püsiva struktuuri indeks} \quad I^{\bar{p}}_{p.str} = \frac{\sum p_1 w_0}{\sum p_0 w_0}, \tag{11.57}$$

$$\text{struktuurinihete indeks} \quad I^{\bar{p}}_{str.n} = \frac{\sum p_1 w_1}{\sum p_1 w_0}. \tag{11.58}$$

11.8. Tegurite absoluutne mõjuulatus

Majandusprotsesside uurimisel pakub huvi ka tegurite absoluutne mõjuulatus (eurodes, kilogrammides, meetrites). Tegurite absoluutse mõjuulatuse leidmine eeldab koondtulemuse absoluutse juurdekasvu jaotamist. Selleks kasutatakse mitmeid erinevaid jaotusmeetodeid, laialdasemalt on kasutusel **ahelasendusmeetod**.

Ahelasendusmeetodi kasutamisel leitakse igale indeksile vastav absoluutne mõjuulatus indeksi lugeja ja nimetaja vahena.

Ahelasendusmeetod

Ahelasendusmeetod koond- ja teguriindeksite korral

Toodangu maksumuse koondindeks oli $I_Q = \sum p_1q_1 / \sum p_0q_0$. Koondtulemuse absoluutne juurdekasv avaldub koondindeksi lugeja ja nimetaja vahena:

$$\Delta Q(p,q) = \sum p_1q_1 - \sum p_0q_0 = Q_1 - Q_0. \tag{11.59}$$

Kvalitatiivse teguri indeks oli $I_p = \sum p_1q_1 / \sum p_0q_1$. Kvalitatiivse teguri muutumise absoluutne mõju avaldub vastava teguriindeksi lugeja ja nimetaja vahena:

$$\Delta Q(p) = \sum p_1q_1 - \sum p_0q_1 = Q_1 - Q_{tingl}. \tag{11.60}$$

Kvantitatiivse teguri indeks oli $I_q = \sum p_0q_1 / \sum p_0q_0$. Kvantitatiivse teguri muutumise absoluutne mõju avaldub vastava teguriindeksi lugeja ja nimetaja vahena:

$$\Delta Q(q) = \sum p_0q_1 - \sum p_0q_0 = Q_{tingl} - Q_0. \tag{11.61}$$

Tegurite absoluutsed mõjud annavad kokku koondtulemuse absoluutse muutuse:

$$\Delta Q(p,q) = \Delta Q(p) + \Delta Q(q). \tag{11.62}$$

Ahelasendusmeetod struktuuriindeksite korral

Keskmise hinna muutuva struktuuri indeks oli

$$I^{\bar{p}}_{m.str} = \frac{\sum p_1q_1}{\sum q_1} : \frac{\sum p_0q_0}{\sum q_0} = \frac{\bar{p}_1}{\bar{p}_0}.$$

Keskmise hinna absoluutne kogumuutus on jagatava ja jagaja vahe ehk aruandeperioodi keskmine hind miinus baasperioodi keskmine hind:

$$\Delta\bar{p}(p,q) = \frac{\sum p_1q_1}{\sum q_1} - \frac{\sum p_0q_0}{\sum q_0} = \bar{p}_1 - \bar{p}_0. \tag{11.63}$$

Püsiva struktuuri indeks oli

$$I^{\bar{p}}_{p.str} = \frac{\sum p_1q_1}{\sum q_1} : \frac{\sum p_0q_1}{\sum q_1} = \frac{\bar{p}_1}{\bar{p}_{tingl}}.$$

Kvalitatiivse teguri osamuutus, mis on tingitud kvalitatiivse teguri enda muutusest, on samuti jagatava ja jagaja vahe ehk aruandeperioodi keskmine hind miinus tinglik hind:

$$\Delta\bar{p}(p) = \frac{\sum p_1 q_1}{\sum q_1} - \frac{\sum p_0 q_1}{\sum q_1} = \bar{p}_1 - \bar{p}_{tingl}. \tag{11.64}$$

Struktuurinihete indeks oli

$$I^{\bar{p}}_{str.n} = \frac{\sum p_0 q_1}{\sum q_1} : \frac{\sum p_0 q_0}{\sum q_0} = \frac{\bar{p}_{tingl}}{\bar{p}_0}.$$

Kvalitatiivse teguri osamuutus, mis on tingitud kvantitatiivse teguri muutusest, on jagatava ja jagaja vahe ehk tinglik hind miinus baasperioodi keskmine hind:

$$\Delta\bar{p}(q) = \frac{\sum p_0 q_1}{\sum q_1} - \frac{\sum p_0 q_0}{\sum q_0} = \bar{p}_{tingl} - \bar{p}_0. \tag{11.65}$$

Absoluutsed osamuutused annavad kokku keskmise hinna absoluutse kogumuutuse:

$$\Delta\bar{p}(p,q) = \Delta\bar{p}(p) + \Delta\bar{p}(q). \tag{11.66}$$

11.9. Näide: käibe ja keskmise hinna indeksanalüüs


N11Indeksid

Autode müügiga tegeleval ettevõttel on müügiesindused kolmes erinevas linnas. Tabelis 11.2 on esitatud teatud sõiduauto mudeli X müügikäivet iseloomustavad andmed erinevates linnades.

Tabel 11.2. Müügikäive ja müüdud autode arv

Linn	Müügikäive, tuh €		Müüdud autode arv	
	Baasperiood	Aruandeperiood	Baasperiood	Aruandeperiood
Tallinn	2 508	2 950	220	250
Tartu	1 010	990	100	99
Pärnu	441	396	45	40

Esitatud andmete alusel tuleb leida:

- müügikäibe kogumuutus ja osamuutused baasperioodiga võrreldes nii protsentides kui ka eurodes;
- autode keskmise müügihinna kogumuutus ja osamuutused baasperioodiga võrreldes nii protsentides kui ka eurodes.

Käive on kvantitatiivne suurus, kasutada tuleb teguriindeksite süsteemi. Kasutame versiooni I ja teeme arvutused valemite (11.23)–(11.28) põhjal. Hind on kvalitatiivne suurus ja selle muutuse

analüüsimiseks tuleb kasutada struktuuriindekseid ning lähtume samuti I versioonist. Arvutusteks kasutame valemeid (11.38)–(11.40). Müügikäibe ja keskmise hinna absoluutsete muutuste (eurodes) leidmiseks kasutame alapeatükis 11.8 toodud valemeid (11.59)–(11.61) ja (11.63)–(11.65). Mugav on kasutada tabelarvutust.

Valemite (11.23)–(11.28) ja (11.38)–(11.40) järgi arvutamiseks on meil vaja leida järgmised summad:

$$\sum q_0, \quad \sum q_1, \quad \sum p_0q_0, \quad \sum p_1q_1, \quad \sum p_0q_1. \tag{11.67}$$

Esimesed neli summat saame leida tabeli 11.2 veergude põhjal, kus müüdud autode arv on q ja müügikäive pq. Tingliku käibe $\sum p_0q_1$ leidmiseks arvestame seda, et baasperioodi hinnad p_0 saab leida baasperioodi käivete ja koguste kaudu:

$$p_0 = \frac{p_0q_0}{q_0} \tag{11.68}$$

ning siis on võimalik leida ka korrutised p_0q_1.

Tabel 11.3. Indeksanalüüsi arvutused tabeli 11.2 põhjal

	p_0q_0	p_1q_1	q_0	q_1	p_0	p_0q_1
Tallinn	2508	2950	220	250	11,4	2850,0
Tartu	1010	990	100	99	10,1	999,9
Pärnu	441	396	45	40	9,8	392,0
KOKKU	3959	4336	365	358		4241,9

Arvutuste organiseerimiseks lisame tabelisse 11.2 veerud p_0 ja p_0q_1, kus p_0 arvutamiseks kasutame valemit 11.68. Tabeli viimasele reale KOKKU leiame vajaminevad viis summat. Selguse mõttes paneme tabeli 11.3 veergude pealkirjadeks valemites kasutatud tähistused.

1.Käibe indeksanalüüs

1.1. Kogumuutus

Autode müügikäibe kogumuutuse annab meile koondindeks I_Q:

$$I_Q = \frac{\sum p_1q_1}{\sum p_0q_0} = \frac{4336}{3959} \approx 1{,}095.$$

Autode müügikäibe kogukasv tuhandetes eurodes on

$$\Delta pq(pq) = \sum p_1q_1 - \sum p_0q_0 = 4336 - 3959 = 377.$$

Võime järeldada, et sõiduautode X müügikäive kasvas aruandeperioodil baasperioodiga võrreldes nii hindade kui ka müüdud autode arvu muutumise tõttu 9,5% ehk 377 tuhande euro võrra.

Niiviisi arvutatud kogumuutuse saame jagada osamuutusteks — hindade muutumisest ja müüdud autode arvu muutumisest tingitud käibe muutus.

1.2. Hindade muutumisest tingitud käibe osamuutuse annab hinnaindeks:

$$I_p = \frac{\sum p_1 q_1}{\sum p_0 q_1} = \frac{4336}{4241{,}9} \approx 1{,}022.$$

Autode hindade muutumisest tingitud käibe osamuutus tuhandetes eurodes (absoluutne mõjuulatus) on

$$\Delta pq(p) = \sum p_1 q_1 - \sum p_0 q_1 = 4336 - 4241{,}9 = 94{,}1.$$

1.3. Autode arvu muutusest tingitud käibe osamuutuse annab mahuindeks:

$$I_q = \frac{\sum p_0 q_1}{\sum p_0 q_0} = \frac{4241{,}9}{3959} \approx 1{,}071.$$

Müüdud autode arvu muutumisest tingitud käibe osamuutus tuhandetes eurodes on

$$\Delta pq(q) = \sum p_0 q_1 - \sum p_0 q_0 = 4241{,}9 - 3959 = 282{,}9.$$

Kokkuvõttes saame osamuutuste kohta öelda, et sõiduautode müügikäive kasvas hindade arvelt 2,2% ehk 94,1 tuhat eurot ning müüdud autode arvelt 7,1% ehk 282,9 tuhat eurot.

1.4. Kontroll

Arvutuste kontrollimiseks kasutame järgmisi seoseid.

1. Osamuutusi väljendavate indeksite korrutis peab võrduma kogumuutust näitava indeksiga:

$$I_Q = I_p I_q = 1{,}022 \cdot 1{,}071 = 1{,}095.$$

2. Absoluutsete osamuutuste summa peab võrduma absoluutse kogumuutusega:

$$\Delta pq(pq) = \Delta pq(p) + \Delta pq(q) = 94{,}1 + 282{,}9 = 377.$$

2. Keskmise hinna indeksanalüüs

Nüüd kasutame struktuuriindeksite süsteemi, I versiooni. Algul leiame valemitest (11.43)–(11.45) baasperioodi ning aruandeperioodi keskmised hinnad ja tingliku hinna tuhandetes eurodes:

$$\bar{p}_1 = \frac{\sum p_1 q_1}{\sum q_1} = \frac{4336}{389} \approx 11{,}147,$$

$$\bar{p}_0 = \frac{\sum p_0 q_0}{\sum q_0} = \frac{3959}{365} \approx 10{,}847,$$

$$\bar{p}_{tingl} = \frac{\sum p_0 q_1}{\sum q_1} = \frac{4241{,}9}{389} \approx 10{,}905.$$

2.1. Keskmise hinna kogumuutus

Autode keskmise hinna kogumuutuse leiame muutuva struktuuri indeksi abil:

$$I^{\bar{p}}{}_{m.str} = \frac{\bar{p}_1}{\bar{p}_0} = \frac{11{,}147}{10{,}847} \approx 1{,}028.$$

Autode keskmise hinna absoluutne kogumuutus tuhandetes eurodes on

$$\Delta\bar{p}(p,q) = \bar{p}_1 - \bar{p}_0 = 11{,}147 - 10{,}847 = 0{,}3.$$

Järeldame, et sõiduautode X keskmise hinna kogukasv aruandeperioodil võrreldes baasperioodiga oli 2,8% ehk 0,3 tuhat eurot. See kasv oli tingitud nii hindade kui ka müügistruktuuri muutumisest.

2.2. Hindade muutusest tingitud osamuutus

Autode keskmise hinna osamuutuse, mis on tingitud ainult autode hindade muutumisest, saame püsiva struktuuri indeksi abil:

$$I^{\bar{p}}_{p.str} = \frac{\bar{p}_1}{\bar{p}_{tingl}} = \frac{11{,}147}{10{,}905} \approx 1{,}022.$$

Autode keskmise hinna absoluutne osamuutus, mis on tingitud autode hinna muutumisest, avaldub järgmiselt

$$\Delta\bar{p}(p) = \bar{p}_1 - \bar{p}_{tingl} = 11{,}147 - 10{,}905 = 0{,}242.$$

2.3. Müügistruktuuri muutusest tingitud osamuutus

Autode keskmise hinna osamuutuse, mis on tingitud autode müügistruktuuri muutumisest, saame struktuurinihete indeksi abil

$$I^{\bar{p}}_{str.n} = \frac{\bar{p}_{tingl}}{\bar{p}_0} = \frac{10{,}905}{10{,}847} \approx 1{,}005.$$

Autode keskmise hinna müügistruktuuri muutustest tingitud absoluutne osamuutus:

$$\Delta\bar{p}(q) = \bar{p}_{tingl} - \bar{p}_0 = 10{,}905 - 10{,}847 = 0{,}058.$$

Kokkuvõtvalt saame müüdud autode keskmise hinna osamuutuste kohta öelda, et hindade muutumise arvelt kasvas autode keskmine hind 2,2% ehk 0,242 tuhat eurot. Autode müügistruktuuri muutuste arvelt kasvas autode keskmine hind aga 0,5% ehk 0,058 tuhat eurot.

2.4. Kontroll

1. Keskmise hinna osamuutuseid näitavate indeksite korrutis peab võrduma keskmise hinna kogumuutust näitava indeksiga:

$$I^{\bar{p}}{}_{m.str} = I^{\bar{p}}_{p.str} I^{\bar{p}}_{str.n} = 1{,}022 \cdot 1{,}005 = 1{,}028.$$

2. Absoluutsete osamuutuste summa peab võrduma keskmise hinna absoluutse kogumuutusega:

$$\Delta\bar{p}(p,q) = \Delta\bar{p}(p) + \Delta\bar{p}(q) = 2{,}242 + 0{,}058 = 0{,}3.$$

11.10. Paasche ja Laspeyresi indeksid

Eelnevalt nägime, et liitkogumit iseloomustavate indeksite arvutamisel tuleb kasutada erinevatele komponentidele vastavaid kaalusid (vt näiteks valemid (11.6) ja (11.11)). Probleem on selles, millise perioodi kaalusid kasutada: kas baasperioodi või aruandeperioodi kaalusid.

Paasche indeks

Paasche indeksi korral käib kaalusüsteem perioodiga kaasas, kasutatakse aruandeperioodi t kaalusid f_t:

$$I_t^P = \frac{\sum f_t y_t}{\sum f_t y_0}. \tag{11.69}$$

Nimetus tuleb Saksa majandusteadlase Hermann Paasche (1851–1925) järgi. Paasche indeksi korral käivad kaalud kogu aeg kaasas ning kogumi struktuur vastab aruandeperioodi struktuurile. Indeksi muutus ajas kajastab ka struktuuri muutumist. Kui hinnaindeksina kasutatakse Paasche indeksit, siis ei näita see puhtalt hindade muutumist, vaid ka tarbimise struktuuri muutumist. Sellisena arvutatakse börsiindekseid, kus hindadeks on indeksi koosseisu kuuluvate väärtpaberite hinnad ja kaaludeks vastavate väärtpaberite arv börsil. Paasche indeks nõuab igal perioodil kaalude ümberarvutamist.

Teine võimalus on kasutada kogu aeg baasperioodi kaalusid ning ignoreerida muutusi kogumi struktuuris.

Laspeyresi indeksi

Laspeyresi indeksi korral kasutatakse aruandeperioodi t indeksi arvutamisel baasperioodi kaalusid f_0 ning ajas edasi liikudes on kogumi struktuur kogu aeg konstantne:

$$I_t^L = \frac{\sum f_0 y_t}{\sum f_0 y_0}. \tag{11.70}$$

Viimane indeks on nimetatud Saksa majandusteadlase Ernst Louis Étienne Laspeyresi (1834–1913) järgi. É. Laspeyres oli Tartu Ülikooli geograafia, etnograafia ja statistika professor aastatel 1869–1873. Laspeyresi indeksi arvutamine on lihtsam kui Paasche oma, sest kaalud on kogu aeg konstantsed ja ei pea pidevalt koguma andmeid uute kaalude arvutamiseks.

Hinnaindeksi puhul on valemites (11.69) ja (11.70) kaaluks f kogus ning y on hind. Koguseindeksi korral on vastupidi: kaaluks f on hind ja y on kogus. Näiteks mingi tööstusharu kogutoodangu indeksit võib

arvutada jooksevhindades (Paasche indeks) või püsivhindades (Laspeyresi indeks).

Jooksva aasta tarbijahinnaindeksi (THI) arvutamisel kasutab Eesti Statistikaamet eelneva aasta detsembrikuu baashindasid ja eratarbija kulutuste struktuuri. Iga aasta lõpus uuendatakse kulutuste struktuuri. See on vajalik, kuna inimeste tarbimisharjumused muutuvad kiiresti ning kui kasutada mitme aata vanust struktuuri, on see vananenud (vt näide 11.6). See tähendab, et aasta jooksul kasutatakse THI arvutamiseks Laspeyresi indeksi metoodikat ning ühe aasta jooksul kajastab THI ainult hindade muutumist. Kui aga võrrelda erinevate aastate tarbijahinnaindekseid, siis nende erinevus kajastab nii erinevust hindades kui ka tarbimise struktuuris. Järelikult võib aastate lõikes THI muutuda ka siis, kui hinnad ei muutu, aga inimesed muudavad oma tarbimisharjumusi. Teiselt poolt, kui mõne teenuse hind tõuseb oluliselt ning seetõttu inimesed loobuvad selle teenuse kasutamisest, siis THI ei pruugi muutuda.

THI muutuse tõlgendus

Tarbijahinnaindeksi muutus

- ühe kalendriaasta jooksul iseloomustab ainult hindade muutust;
- võrreldes eelmise aasta mingi kuuga, kirjeldab nii hindade kui ka tarbimisharjumuste muutumist.

Tabel 11.4. Tarbijahinnaindeks 2015. aasta veebruaris

THI muutus, võrreldes jaanuariga 2015	0,6%	Väljendab ainult hindade muutust
THI muutus, võrreldes veebruariga 2014	−0,8%	Väljendab nii hindade muutust kui ka tarbimise struktuuri muutust

Fisheri indeks

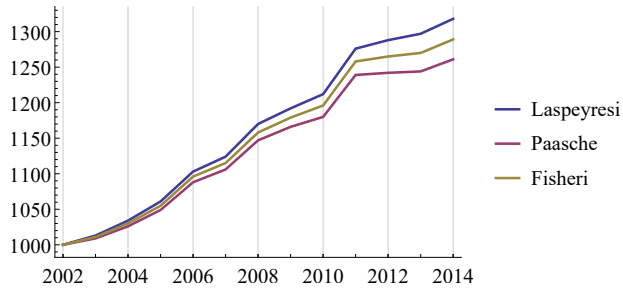
Mõlemal, nii Paasche kui ka Laspeyresi indeksil on oma puudused. Seepärast kasutatakse mõnikord **Fisheri indeksit**, mis on Laspeyresi ja Paasche indeksite geomeetriline keskmine ning seetõttu silub mõningaid moonutusi, mis on tingitud nende indeksite puhul kasutatavate kaalusüsteemide erinevusest:

$$I^F = \sqrt{I^P \cdot I^L}. \tag{11.71}$$

Nimetatud on see USA majandusteadlase Irwing Fisheri (1867–1947) järgi. Euroopa Parlamendi ja Nõukogu määruses „Euroopa Liidus kasutatava Euroopa rahvamajanduse ja regionaalse arvepidamise süsteemi

kohta“ öeldakse, et kuna riikide võrdluse korral on Laspeyresi ja Paasche valemi abil saadud tulemuste vahel sageli märgatav erinevus, on sel eesmärgil ainsana vastuvõetav Fisheri indeksi kasutamine.

Alates aastast 2002 on Uus-Meremaa Statistikaamet arvutanud tarbijahinnaindeksit paralleelselt kõigi kolme valemi järgi. Joonisel 11.5 on Laspeyresi, Paasche ja Fisheri valemi alusel leitud Uus-Meremaa tarbijahinnaindeksi dünaamika 2002–2014 (Statistics New Zealand, 2015). Baasaastal 2002 oli kõigi indeksite väärtus 1000. Graafikult on näha, et hinnatõusu korral näitab Laspeyresi indeks suuremat tõusu kui Paasche indeks. Osaliselt on see tingitud sellest, et kaupade kallinemisel tarbijad hakkavad ostma odavamaid asenduskaupu ning tarbimise struktuur muutub. Laspeyresi indeks seda ei arvesta. Fisheri indeksi väärtus on Laspeyresi ja Paasche indeksi vahel ning seepärast nimetatakse seda mõnikord ideaalseks hinnaindeksiks (*ideal price index*).



Joonis 11.5. Laspeyresi, Paasche ja Fisheri valemi alusel leitud Uus-Meremaa tarbijahinnaindeks 2002–2014

11.11. Börsiindeksid

Börsiindeksi konstrueerimisel tuleb otsustada:

- millised väärtpaberid kuuluvad indeksi koosseisu ja millistel juhtudel tuleb koosseisus teha muudatusi;
- milline on indeksi arvutamise eeskiri.

Börsiindeksite arvutuseeskirju on mitmeid. 1896. aastal Charles Dow lihtsalt liitis kokku 12 aktsia hinnad ning jagas need arvuga 12. Sellest ajast pärineb Dow Jonesi aktsiaindeks. Alates 1928. aastast kuulub indeksisse 30 New Yorgi börsil noteeritud suuraktsiat[1]. Indeksi arvutusvalem on

$$I_t = I_0 \frac{\sum_i p_{it}}{D_t}, \tag{11.72}$$

kus I_0 on indeksi väärtus esimesel päeval (harilikult 100 punkti), p_{it} on i-nda väärtpaberi hind ajahetkel t ja D_t on jagaja (*index divisor*),

[1] `https://www.djaverages.com/`

mis tagab indeksi pidevuse indeksi koosseisu ja turukapitalisatsiooni muutustel. Esimesel päeval võetakse see võrdseks indeksisse kuuluvate väärtpaberi hindade summaga. Turukapitalisatsiooni ja fondiemissiooni korral on

$$D_t^* = D_t \frac{\sum_i p_{it}^*}{\sum_i p_{it}}, \tag{11.73}$$

kus tärniga on tähistatud väärtused vahetult pärast muutust ja ilma tärnita vahetult enne muutust.

Miks on vaja jagajat D_t? Oletame, et börsil on kolm aktsiat hindadega 15, 20 ja 25 eurot. Nende hindade aritmeetiline keskmine on 20 eurot. Ühel päeval otsustab see ettevõte, kelle aktsia hind on 20 eurot, vähendada aktsia nimiväärtust kaks korda ning uueks hinnaks on 10 eurot. Seda nimetatakse **splittimiseks**. Aktsiakapital sellega ei muutu, sest suureneb aktsiate arv. Aga aktsiahindade aritmeetiline keskmine on nüüd 16,667 eurot ja indeks teeb hüppe.

Valemi (11.72) kasutamine koos valemiga (11.73) tasandab selle. Indeks peale splittimist on

$$I_t^* = I_0 \frac{\sum_i p_{it}^*}{D_t^*}.$$

Paneme sinna sisse D_t^* valemist (11.73):

$$I_t^* = I_0 \frac{\sum_i p_{it}^*}{D_t^*} = I_0 \sum_i p_{it}^* \frac{\sum_i p_{it}}{D_t \sum_i p_{it}^*} = I_0 \frac{\sum_i p_{it}}{D_t},$$

millest näeme, et splittimine indeksi väärtust ei muuda.

Enamike börsiindeksite konstrueerimisel soovitakse, et suurema aktsiate arvuga ettevõtete aktsiahinna muutus mõjutaks indeksit rohkem. Seepärast kasutatakse kaalutud indeksit, kus kaaludeks on aktsiate arv. Aktsiate arv vastab perioodile, mille jaoks indeksit arvutatakse, s.t kasutatakse Paasche indeksit:

$$I_t = \frac{\sum_i q_{it} p_{it}}{\sum_i q_{i0} p_{i0}}, \tag{11.74}$$

kus q_{it} on i-nda ettevõtte aktsiate arv ajahetkel t ja p_{it} aktsiate hind ajahetkel t. Korrutis $q_{it}p_{it}$ on ettevõtte turukapitalisatsioon ehk turuväärtus ajahetkel t. Summa üle kõigi indeksisse lülitatud aktsiate on aktsiate kogukapitalisatsioon ning indeks näitab kogukapitalisatsiooni muutumist.

Kui valemis (11.74) võtta indeksi algväärtuseks I_0 ning lisada ka indeksi pidevuse tagamiseks vajalik jagaja D_t, saame

$$I_t = I_0 \frac{\sum_i q_{it} p_{it}}{D_t}. \tag{11.75}$$

Nii arvutatakse mitmeid tuntuid indekseid, nagu S&P 500, Londoni börsi indeks FTSE 100, Nasdaq Balti indeksid, sealhulgas ka Nasdaq Tallinna börsi indeksit (kuni aastani 2005 TALSE).

Mingi ettevõtte kaalu leidmiseks tuleb selle ettevõte turukapitalisatsioon jagada aktsiate kogukapitalisatsiooniga. Näiteks S&P 500 kogukapitalisatsioon oli 2017. aasta veebruaris 21,4 triljonit dollarit. Ettevõtte Apple Inc. turukapitalisatsioon oli 735 miljardit dollarit. Apple'i aktsia kaal indeksis S&P 500 oli siis $735/21400 = 3{,}4\%$. Kui Apple'i aktsia hind suureneb 20% ja kõigi ülejäänud aktsiate hinnad jäävad samaks, siis indeksi S&P väärtus suureneb 0,69% ($3{,}4\% \cdot 20\%$).

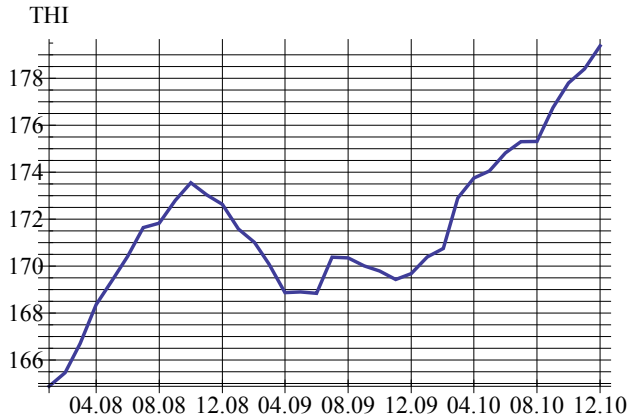
11.12. Ülesanded

Alus- ja ahelindeks

11.1. Tabelis on SKP elaniku kohta kahes Eesti maakonnas aastatel 2010–2013, tuhat eurot[2]. Leida alusindeks (baasaastaks võtta aasta 2010) ja ahelindeks. VASTUS lk 697.

	2010	2011	2012	2013
Harju maakond	15,7	17,6	19,0	20,2
Tartu maakond	9,7	10,5	11,4	12,3

11.2. Joonisel 11.6 on toodud tarbijahinnaindeks (THI) ajavahemikul jaanuar 2008 – detsember 2010. Baasaastaks on 1997 ja siis THI = 100.



Joonis 11.6. THI jaanuar 2008 – detsember 2010

1. Mitu protsenti olid hinnad kasvanud 2008. aasta augustis, võrreldes 1997. aastaga?

[2]Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel RAA0050: sisemajanduse koguprodukt maakonna järgi.

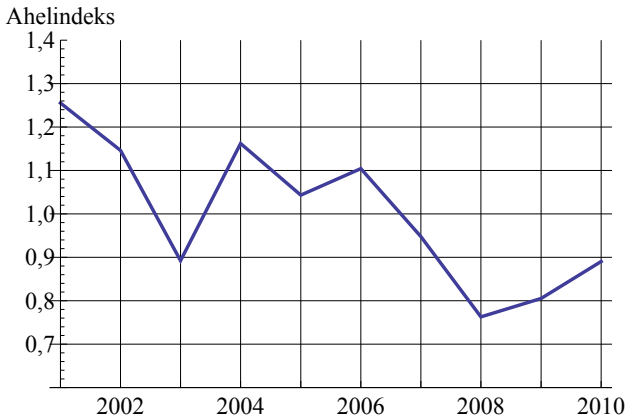
2. Mitu protsenti olid hinnad kasvanud 2008. aasta augustis, võrreldes sama aasta aprilliga?
3. Kuidas olid hinnad muutunud 2009. aasta aprillis, võrreldes 2008. aasta detsembriga?

VASTUS lk 697.

11.3. Joonisel 11.7 on toodud liiklusõnnetuste arvu ahelindeks aastatel 2001–2010[3].

1. Millistel aastatel vähenes liiklusõnnetuste arv, võrreldes eelneva aastaga?
2. Kuidas muutus liiklusõnnetuste arv 2003. aastal, võrreldes aastaga 2002?
3. Kuidas muutus liiklusõnnetuste arv 2004. aastal, võrreldes aastaga 2003?
4. Kuidas muutus liiklusõnnetuste arv 2006. aastal, võrreldes aastaga 2003?

VASTUS lk 697.



Joonis 11.7. Liiklusõnnetuste arvu ahelindeks 2001–2010

Individuaalindeksid

11.4. 2013. aastal oli kauba hind 43,5 eurot, 2014. aastal odavnes kaup 10%. 2013. aastal müüdi seda kaupa 2500 tükki, 2014. aastal suurenes müük 750 tüki võrra. Leida hinna, müügimahu ja käibe indeks. VASTUS lk 697.

11.5. Tabelis on kolme liiki mootorikütuse hinnad Eesti automaattanklates 2015. ja 2016. aasta veebruaris, eurot liitri kohta (*Hinnainfo* 2016). Leida hinnaindeksid. Milline kütus odavnes aasta jooksul kõige enam?

[3] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel TS094: inimkannatanutega liiklusõnnetused teedel maakonna järgi.

Kütus	veebr 2015	veebr 2016
Bensiin 95E	1,053	0,958
Bensiin 98	1,092	1,008
Diislikütus	1,109	0,935

Keskmised indeksid

11.6. Ühte kaubagruppi kuulub kolm kaupa: A, B ja C. Kauba A hinnaindeks oli 1,06, kauba B hinnaindeks 0,92 ja kauba C hinnaindeks 1,11. Kui suur oli selle kaubagrupi keskmine hinnaindeks, kui kauba A osakaal on 55%, kauba B osakaal 30% ja kaubal C 15%? VASTUS lk 697.

11.7. AS Tallinna Vesi otsesed tootmiskulud koosnevad neljast komponendist: vee erikasutusõiguse tasu (13,2%), kulu kemikaalidele (21,7%), elektrikulu (38,0%) ja saastetasu (27,1%). Sulgudes on nende kulukomponentide osatähtsus tootmiskuludes 2014. aastal 2015. aastal vee erikasutusõiguse tasu suurenes 4,2%, kulu kemikaalidele vähenes 11,9%, elektrikulu suurenes 0,1% ja saastetasu vähenes 53,7%[4]. Mitu protsenti muutusid otsesed tootmiskulud kokku? VASTUS lk 697.

11.8. Ehitusmaterjalidepoes on müügil kahte sorti vineeri. Sordi A hind tõusis 5%, sordi B hind jäi samaks. Mitu protsenti tõusis selles poes vineeri hind, kui sort B moodustab 60% kogu müüdavast vineerist? VASTUS lk 697.

11.9. Väikeettevõte toodab kahte toodet: A ja B. Toote A omahinna indeks oli 1,6 ja toote B omahinna indeks 1,3. Keskmine omahinna indeks oli 1,39. Kui suur oli kummagi toote osakaal kogutoodangu hulgas? VASTUS lk 697.

Teguriindeksid

11.10. Toodangu maksumuse indeksanalüüsil arvutati välja kolm indeksit: maksumuse koondindeks $I_Q = 1{,}024$ ning kaks teguriindeksit: koguseindeks $I_q = 1{,}068$ ja hinnaindeks $I_p = 0{,}959$. Tõlgendada neid indekseid. VASTUS lk 697.

11.11. Bussitranspordi sõitjateveo iseloomustamisel kasutatakse sõitjakäivet, mida mõõdetakse sõitjakilomeetrites. Üks sõitjakilomeeter vastab ühe sõitja vedamisele ühe kilomeetri kaugusele. 2014. aastal veeti Eestis kohalikel ja kaugliinidel kokku 148 miljonit sõitjat ning sõitjakäive oli 2393 miljonit sõitjakilomeetrit. 2015. aastal olid vastavad arvud 154 miljonit ja 3146 miljonit[5].

1. Mitu protsenti suurenes sõitjakäive 2015. aastal?
2. Kui suur oli kummalgil aastal keskmine sõidukilomeetrite arv ühe sõitja kohta?

[4] AS Tallinna Vesi konsolideeritud majandusaasta aruanne 31. detsembril lõppenud majandusaasta kohta. `https://www.tallinnavesi.ee/`

[5] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel TS541: sõitjatevedu bussiga kohalikel ja kaugliinidel.

3. Mitu protsenti suurenes sõitjakäive tänu sõitjate arvu suurenemisele?
4. Mitu protsenti suurenes sõitjakäive seetõttu, et suurenes sõidukilomeetrite arv ühe sõitja kohta?

VASTUS lk 698.

Struktuuriindeksid

11.12. Toodangu keskmise omahinna indeksanalüüsil leiti kolm indeksit: muutuva struktuuri indeks $I_{m.str} = 0{,}936$, püsiva struktuuri indeks $I_{p.str} = 1{,}04$ ja struktuurinihete indeks $I_{str.n} = 0{,}9$. Tõlgendada leitud indekseid. VASTUS lk 698.

11.13. Põllumajandusettevõttel on taliteravilja külvipinda kokku 1600 hektarit. 2015. aastal kasvatati tuhandel hektaril talinisu, mille saagikus oli 5,3 tonni hektari kohta. Ülejäänud külvipinnal kasvatati taliotra, mille saagikus oli 4,6 t/ha. 2016. aastal vähendati taliodra külvipinda poole võrra ning selle võrra suurenes talinisu külvipind. 2016. aasta oli teraviljakasvatajale halb aasta: talinisu saagikus oli 2,9 t/ha ja taliodral 2,7 t/ha.

1. Kui suur oli taliteravilja kogutoodang 2015. ja 2016. aastal?
2. Kui suur oli taliteravilja keskmine saagikus 2015. ja 2016. aastal?
3. Mitu protsenti muutus taliteravilja keskmine saagikus?
4. Mitu protsenti muutus keskmine saagikus tänu sellele, et talinisu ja taliodra saagikus oli 2016. aastal madalam?
5. Mitu protsenti muutus keskmine saagikus tänu sellele, et muudeti taliteraviljade külvipinda?

VASTUS lk 698.

Paasche, Laspeyresi, Fisheri indeks

11.14. Arvestame piimatoodete hinnaindeksi arvutamisel ainult kahte toodet: piim ja juust. Tabelis on piima ja juustu hind ning tarbimine ühe elaniku kohta aastas aastatel 2000 ja 2012 (*20 aastat Eesti piimaturul 1995–2013* 2013).

	Jaehind, €/kg		Tarbimine, kg	
	2000	2012	2000	2012
Piim	0,54	0,83	129,0	106,0
Juust	4,19	7,57	13,2	21,1

Leida 2012. aasta piimatoodete hinnaindeks, kui baasaastaks võtta aasta 2000. Hinnaindeksi arvutamiseks kasutada nii Paasche, Laspeyresi kui ka Fisheri valemit. Milline neist kolmest näitab kõige suuremat inflatsiooni? VASTUS lk 698.

Börsiindeksid

11.15. Nasdaq Tallinna börsi põhinimekirjas oli 2017. aasta veebruaris 14 ettevõtet kogukapitalisatsiooniga 2354 miljonit eurot. Harju Elektri turukapitalisatsioon oli 59,4 miljonit eurot[6].

[6] `http://www.nasdaqbaltic.com`

1. Leida Harju Elektri aktsia kaal Tallinna börsil.
2. Kui Harju Elektri aktsia hind tõuseb 40% ja ülejäänud aktsiate hind jääb samaks, kui palju muutub Tallinna börsi indeks?

VASTUS lk 698.

ÜL11Indeksid

Järgmiste ülesannete andmed on failis ÜL11Indeksid

Keskmine indeks

A.11.1. Tabelis on toodud ehitusettevõtte töötajate arv ametite kaupa ja vastava ameti palgatõus protsentides. Leida, mitu protsenti kasvas palk keskmiselt selles ehitusettevõttes. VASTUS lk 698.

Teguriindeksid

A.11.2. Kohalikul omavalitsusel on oluline roll sotsiaalse kaitse korraldamisel, mida tehakse mitmesuguste sotsiaaltoetuste abil. Tabelis on toodud kolme erinevat tüüpi rahaliste toetuste suurus ja toetuse saajate arv ühes omavalitsuses kahel järjestikusel aastal. Leida

a) kulude muutumist iseloomustav indeks;
b) indeks, mis iseloomustab toetuste suuruse muutusest põhjustatud kulude muutust;
c) indeks, mis iseloomustab toetuse saajate arvu muutumisest põhjustatud kulude muutust;
d) mitu eurot muutusid kulud toetustele kokku;
e) kulude absoluutne muutus, mis oli põhjustatud toetuste suuruse muutusest;
f) kulude absoluutne muutus, mis oli põhjustatud toetuse saajate arvu muutumisest.

VASTUS lk 698.

Teguri- ja struktuuri-indeksid

A.11.3. Tööstustoodete statistika kogumisel kasutab Eesti Statistikaamet tööstustoodete loetelu TTL, mis on ühine kõikides Euroopa Liidu riikides. Tabelis on andmed trikookangast rõivaste toodangu kohta aastatel 2011 ja 2012: müüdud toodang ja müüdud toodangu maksumus[7]. Leida

a) maksumuse kogumuutus protsentides;
b) maksumuse osamuutused protsentides, mis on põhjustatud
   - toodete hindade muutumisest;
   - koguste muutumisest;
c) trikookangast rõivaste hinna kogumuutus protsentides;
d) hinna osamuutused protsentides, mis on põhjustatud
   - toodete hindade muutumisest;
   - koguste muutumisest.

[7] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel TO75: tööstustoodang tööstustoodete loetelu (TTL) järgi, 2011; TO74: tööstustoodang tööstustoodete loetelu (TTL) järgi, 2012.

Vastus lk 698.

A.11.4. Baltika Grupp arendab rõivabrände Monton, Mosaic, Baltman, Bastion ja Ivo Nikkolo ning tegeleb rahvusvahelise rõivakaubandusega. Tabelis on Baltika Grupi jaemüük (tuhat eurot) ning müügipind (m^2) turgude lõikes aastatel 2014 ja 2015[8]. Üheks jaemüüki iseloomustavaks suuruseks on müügiefektiivsus, mis näitab jaemüüki ruutmeetri kohta. Leida

a) keskmine müügiefektiivsus aastatel 2014 ja 2015;
b) keskmise müügiefektiivsuse kogumuutus protsentides;
c) jaemüügi muutumisest tingitud keskmise müügiefektiivsuse osamuutus protsentides;
d) müügipinna muutumisest tingitud keskmise müügiefektiivsuse osamuutus protsentides.

Vastus lk 698.

A.11.5. Tabelis on piimatoodete hind ja keskmine ostukogus leibkonnaliikme kohta kuus 2006. ja 2007. aastal[9].

1. Leida, kui palju muutus keskmine piimatoodete hind protsentuaalselt ja absoluutselt.
2. Kui suur osa muutusest oli tingitud üksikute piimatoodete hindade muutusest (protsentides ja absoluutselt)?
3. Kui suur osa keskmise hinna muutusest oli tingitud ostukoguste muutusest (protsentides ja absoluutselt)?

Vastus lk 698.

A.11.6. Ettevõte, kus on alla kümne töötaja, on mikroettevõte. Tabelis on kolme mikroettevõtte realiseerimise netokäive ühe töötaja kohta ja töötajate arv kahel järjestikusel perioodil. Analüüsida keskmist netokäivet ühe töötaja kohta: leida muutuva struktuuri, püsiva struktuuri ja struktuurinihete indeksid ning tõlgendada neid. Vastus lk 698.

A.11.7. Kõigi tegevusalade keskmine brutopalk leitakse kaalutud keskmisena tegevusalade keskmistest brutopalkadest, kaaluks on töötajate arv. Seetõttu võib keskmise brutopalga muutus olla põhjustatud nii tegevusalade keskmiste palkade muutustest kui ka töötajate arvu muutusest. Näiteks kui tegevusalade keskmised palgad jäävad samaks, aga suureneb finantsvahenduses töötavate töötajate arv, suureneb ka keskmine palk, sest finantsvahenduses on palk kõige kõrgem.

Tabelites on toodud keskmine brutokuupalk (eurot) Eestis tegevusalade kaupa ning keskmine töötajate arv igal tegevusalal aasta-

[8] Baltika Grupp IV kvartal ja 12 kuud 2014 tulemused, IV kvartal ja 12 kuud 2015 tulemused. `http://www.baltikagroup.com/`

[9] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel LE31: toidukaupade keskmine ostukogus ja hind leibkonnaliikme kohta kuus.

tel 2008–2015[10]. Teha kindlaks, kui suur osa iga-aastasest keskmise palga muutusest on põhjustatud tegevusalade palkade muutumisest ja kui suur osa töötajate arvu muutusest (struktuurimuutused). Selleks viia läbi keskmise palga indeksanalüüs: leida muutuva struktuuri, püsiva struktuuri ja struktuurinihete indeksid aastatel 2009–2015. Indeksid leida nii, et igal aastal võtta baasaastaks eelnev aasta. Sellisel juhul on muutuva struktuuri indeks Eesti keskmise brutopalga (kõik tegevusalad kokku) ahelindeks. Analüüsi tulemuste kokkuvõtvaks esitamiseks sobib diagramm, kus on toodud kõigi kolme indeksi (muutuva struktuuri, püsiva struktuuri ja struktuurinihete indeksi) dünaamika. Vastus lk 698.

[10] Allikas: Eesti Statistikaamet [e-andmebaas] `http://pub.stat.ee/`. Tabel PA5211: keskmine bruto- ja netokuupalk põhitegevusala (EMTAK 2008) järgi, tabel PA5216: töötajate keskmine arv, taandatud täistööajale, ja selle jaotus põhitegevusala (EMTAK 2008) järgi.

Ülesannete vastused

1. ptk. Sissejuhatus

1.1 a) Sugu nimiskaalas; b) perekonnaseis nimiskaalas; c) laste arv intervallskaalas; d) haridus järjestusskaalas; e) vanus intervallskaalas; f) amet nimiskaalas; g) kategooria järjestusskaalas; h) telefoni nr nimiskaalas. **1.2** a) Nimiskaala; b) intervallskaala; c) nimiskaala; d) järjestusskaala. **1.3** 1. Järjestusskaala. 2. Järjestusskaala. 3. Intervallskaala. 4. Nimiskaala. 5. Nimiskaala. 6. Intervallskaala. 7. Nimiskaala. **1.4** 1. Ei või suhet leida. 2. Võib. 3. Ei või. 4. Võib. **A.1.1** Tabel ÜV.1 ja joonis ÜV.1.

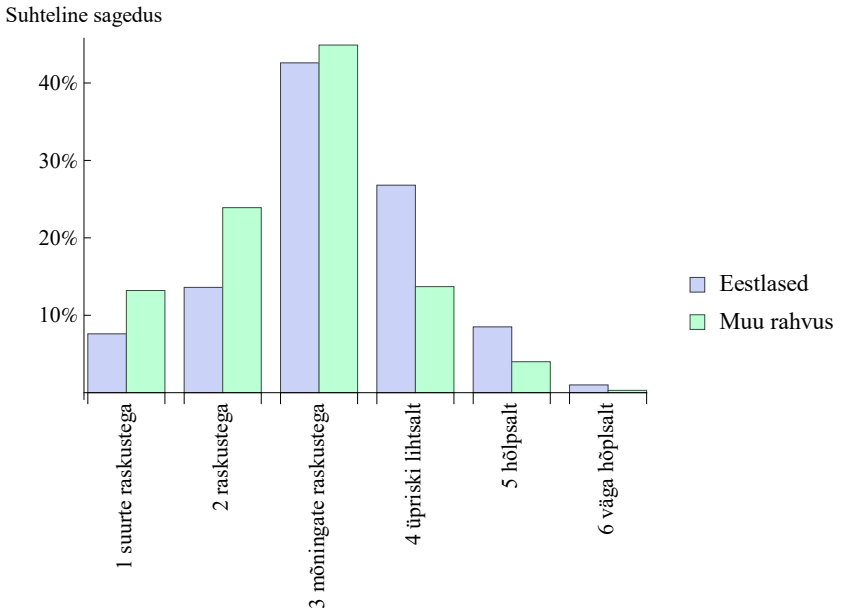
Tabel ÜV.1. Ülesande A.1.1 vastus

Vastusevariant	Eestlased	Muu rahvus
1 suurte raskustega	7,6%	13,2%
2 raskustega	13,6%	23,9%
3 mõningaste raskustega	42,6%	44,9%
4 üpriski hõlpsalt	26,8%	13,7%
5 hõlpsalt	8,5%	4,0%
6 väga hõlpsalt	1,0%	0,3%

A.1.2 1. Vt tabel ÜV.2. 2. 1920. 3. 680. 4. Vt tabel ÜV.3 ja joonis ÜV.2. 5. Vt tabel ÜV.4 ja joonis ÜV.3.

A.1.3 Klasside arv 7,977; klassi laius 0,0301; klassi laius ümardatult 0,03. Sagedustabel:

Ülemine piir	Sagedus
2,65	5
2,68	2
2,71	13
2,74	18
2,77	21
2,80	27
2,83	24
2,86	16
KOKKU	126



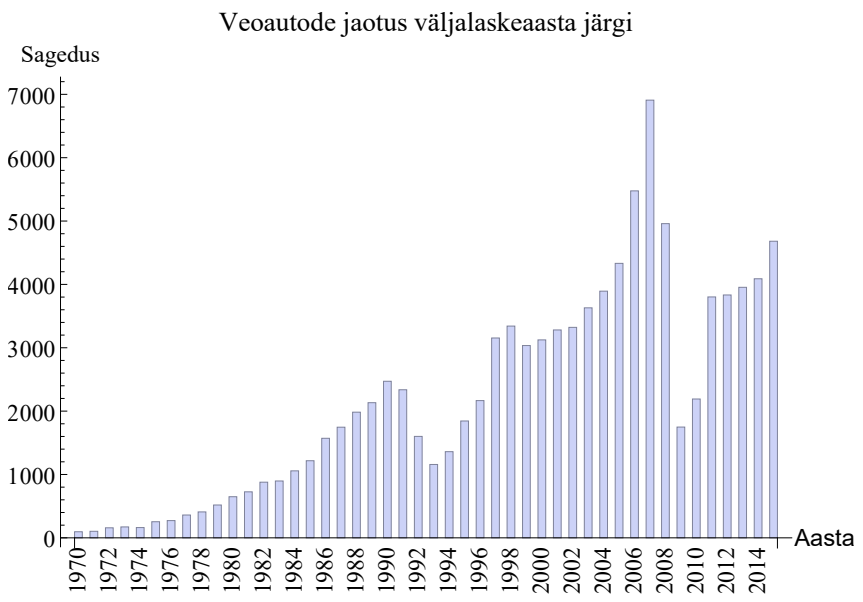
Joonis ÜV.1. Ülesande A.1.1 vastus

Tabel ÜV.2. Ülesande A.1.2 1. osa vastus

Keretüüp	Sagedus	Keretüüp	Sagedus
KAUBIK	50 825	PAAK	913
SADUL	10 809	METSAVEOK	732
MADEL	9 735	KRAANA	511
PIKAP	8 821	KONTEINERIVEOK	80
KALLUR	6 970	ELAMU	51
SIHTOTSTARBELINE	4 983	RUNG	35
FURGOON	3 940	VÕISTLUSAUTO	23
VAHETUSKERE	1 688	LAHTINE	5
KÜLMIK	1 609	Määramata	2

Tabel ÜV.3. Ülesande A.1.2 4. osa vastus

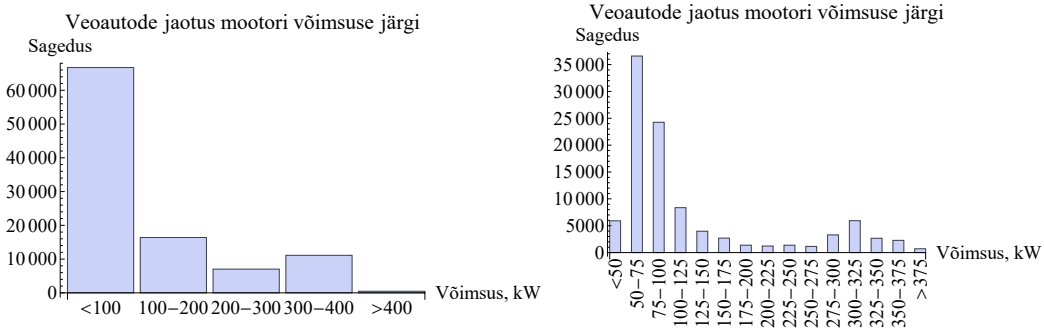
Väljalaske-aasta	Sagedus	Väljalaske-aasta	Sagedus	Väljalaske-aasta	Sagedus
1970	96	1986	1 571	2002	3 322
1971	103	1987	1 746	2003	3 630
1972	158	1988	1 983	2004	3 893
1973	172	1989	2 133	2005	4 332
1974	162	1990	2 471	2006	5 476
1975	254	1991	2 337	2007	6 908
1976	272	1992	1 601	2008	4 959
1977	360	1993	1 158	2009	1 748
1978	409	1994	1 360	2010	2 191
1979	518	1995	1 844	2011	3 802
1980	649	1996	2 166	2012	3 833
1981	726	1997	3 154	2013	3 954
1982	878	1998	3 342	2014	4 089
1983	897	1999	3 035	2015	4 681
1984	1 057	2000	3 124		
1985	1 217	2001	3 281		



Joonis ÜV.2. Ülesande A.1.2 4. osa vastus

Tabel ÜV.4. Ülesande A.1.2 5. osa vastus

Klasside arv 5		Klasside arv 15	
Klassi ülemine piir	Sagedus	Klassi ülemine piir	Sagedus
100	66 738	50	5 905
200	16 394	75	36 581
300	7 023	100	24 252
400	11 114	125	8 360
>400	463	150	3 972
KOKKU	101 732	175	2 689
		200	1 373
		225	1 213
		250	1 369
		275	1 151
		300	3 290
		325	5 939
		350	2 659
		375	2 273
		>375	706
		KOKKU	101 732



Joonis ÜV.3. Ülesande A.1.2 5. osa vastus

2. ptk. Statistilised keskmised

2.1 47 minutit. **2.2** 18. **2.3** 12. **2.4** 3,7 eurot. **2.5** 35. **2.6** a) Suurenes 8%; b) vähenes 10%. **2.7** a) 12; b) 26; c) suur. **2.8** Enne aritmeetiline keskmine 41,2 aastat, mediaan 37 aastat. Pärast aritmeetiline keskmine 36,4 aastat, mediaan 37 aastat. **2.9** Kuna aritmeetiline keskmine oli mediaanist paremal, siis oli üksikuid üliõpilasi, kel testi tegemine võttis kaua aega. Enamik sai sellega kiiresti hakkama. Järelikult oli test pigem kerge. **2.10** b) ja c). **2.11** Ettevõte asus oma näitajaga kolmandas neljandikus. **2.12** a) 375; b) 375; c) 750. **2.13** Meestel 178 g, naistel 240 g. **2.14** Mood. **2.15** Harjumaal 1 kord, Võrumaal 12 korda. **2.16** Mood „1“, mediaan „2“. **2.17** „Nimi“ on nimiskaalas. Ainukeseks kasutatavaks statistiliseks keskmiseks on mood. Kuna kõik nimed esinevad üks kord, siis mood puudub. „Silmade värv“ on nimiskaalas. Keskmine silmade värv on kõige sagedamini esinev silmade värv ehk mood ja see on pruun. „Käitumishinne“ on järjestusskaalas. Sobib kasutada nii mediaani kui moodi. Mediaan on „rahuldav“ ja ka mood on „rahuldav“. „Pikkus“ on intervallskaalas. Sobib kasutada aritmeetilist keskmist ja mediaani. Aritmeetiline keskmine on 1,62 cm, mediaan 1,61 cm. Mood ei sobi, kuna kõik väärtused esinevad vaid üks kord. **2.18** 2013. aasta I poolaastal 241 822 eurot, 2014. aasta I poolaastal 311 346 eurot. **2.19** 800, 800, 1100, 1500 ja 2200 eurot. **2.20** 2,35 tundi. **2.21** Päevakäivete diagrammilt on näha, et aritmeetilise keskmise on tõstnud kõrgele üks päev, mil käive oli erakordselt kõrge. Trahvisumma aluseks tuleks võtta päevakäivete mediaan. **2.22** 85,6 km/h. **2.23** 14,63 €/tk. **2.24** 4,16 t/ha. **2.25** 1,56 €/kg. **2.26** 1,56 €/kg. **2.27** 26,1 toodet töötaja kohta. **2.28** 26,1 toodet töötaja kohta. **2.29** 1,0023. **2.30** 11,2%. **2.31** 2,04%. **2.32** Õigus on ettevõttel. Keskmine hinnatõus aastas oli 9,85%, sest keskmine kasvutempo (geomeetriline keskmine) oli 1,0985. **2.33** 6,6%.

A.2.1 Jõgeva maakonnas 3,2 loodud ettevõtet valla kohta, Põlva maakonnas 2,9 loodud ettevõtet valla kohta. **A.2.2** 526,29 eurot. **A.2.3** Elanike arvu aritmeetiline keskmine on kõige suurem 1. Pärnu maakonnas; 2. Tartu maakonnas.

	Pärnu maakond	Põlva maakond	Tartu maakond
		Ilma maakonnakeskuseta	
Aritmeetiline keskmine	2452,9	1879,7	2410,9
Mediaan	2483	1464	1883
		Koos maakonnakeskusega	
Aritmeetiline keskmine	4461,4	2191,2	6772,5
Mediaan	2507,5	1465	1932
		Protsentuaalne muutus	
Aritmeetiline keskmine	81,9%	16,6%	180,9%
Mediaan	1,0%	0,1%	2,6%

A.2.4 Ärindus ja haldus 2,29 kuud, sotsiaalteadused 2,06, tervis ja heaolu 5,04 kuud. **A.2.5** Mediaan 7,6%. Sellest suurem oli töötuse määr Ida-Viru, Lääne-Viru, Põlva, Pärnu, Rapla, Saare ja Valga maakonnas. **A.2.6** 2002. aastal esimeses neljandikus: Eesti 48, 1. kvartiil 53. 2013. aastal teises neljandikus: Eesti 73, 1. kvartiil 64, mediaan 82. **A.2.7** Kvartiilid on

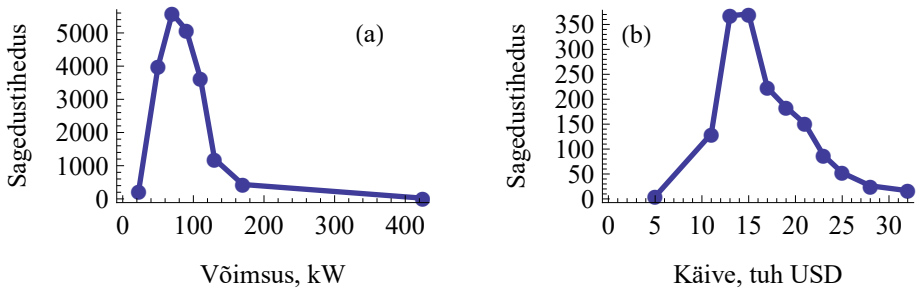
Kvartiili järk	Kvartiil
1	1,93
2	9,61
3	32,26

Eesti oli kolmandas neljandikus.

A.2.8 a) 4 053 831,02 eurot; b) 110 752 283,54 eurot; c) vastused on tabelis; d) 48.

Protsentiili järk	Protsentiil (ümardatud)
0,1	1 529
0,5	6 216
0,95	84 104
0,99	279 816

A.2.9 Madise. **A.2.10** Keskmine hind on mood 3,00 €/kg. **A.2.11** 1. Mbare 0,29 $, Renkini Bus Terminus 0,34 $. 2. Mbare 73,5%, Renkini Bus Terminus 69,6%. 3. Mbare. 4. Mbare 52,9%, Renkini Bus Terminus 21,7%. **A.2.12** 1. Aritmeetiline keskmine 94,0 kW, mediaan 83,4 kW, mood 75,2 kW. 2. Vt joonis ÜV.4 (a). **A.2.13** 1. Aritmeetiline keskmine 16,8 tuhat $, mediaan 15,7 tuhat $, mood 14,0 tuhat $. 2. Vt joonis ÜV.4 (b).



Joonis ÜV.4. Graafik (a) ülesande A.2.12 2. osa vastus. Graafik (b) ülesande A.2.13 2. osa vastus

A.2.14 1,386. **A.2.15** 3,06%. **A.2.16** 88,42 €. **A.2.17** 1,767 €. **A.2.18** Keskmised kasvutempod kuus:

2009	2010	2011	2012	2013
1,019	1,027	1,004	0,982	0,993

Kõige väiksem kasvutempo oli 2012. aastal.

3. ptk. Variatsioonnäitarvud ja jaotuse kuju näitarvud

3.1 a) 0; b) 2,8; c) $\approx 3{,}286$. **3.2** 5. **3.3** Kogumil B. **3.4** Kogumi B standardhälve on 0, see on kõige väiksem. Kogumi C standardhälve on kõige suurem, sest üksikud väärtused on aritmeetilisest keskmisest 10 mõlemal pool kaugemal kui kogumis A. **3.5** a) Aritmeetiline keskmine suureneb arvu a võrra, standardhälve jääb samaks. b) Nii aritmeetiline keskmine kui ka standardhälve suurenevad b korda. **3.6** Rohkem varieerus toodang, mille variatsioonikordaja on 7,4%. Kulude variatsioonikordaja on 6,4%. **3.7** Pinnaühiku hinna varieerumine oli suurem väiksematel korteritel, sest variatsioonikordaja 40,5% on suurem kui suuremate korterite pinnaühiku hinna variatsioonikordaja 39,3%. **3.8** Standardhälve 26 tuhat eurot, variatsioonikordaja ei muutu. **3.9** a) 0,44; b) 0,56; c) 0,04; d) 0,96. **3.10** Pille tulemused: kitsas matemaatika $z \approx 0{,}735$, eesti keel $z \approx 0{,}938$. Eesti keeles oli tulemus silmapaistvam. **3.11** 1 B, 2 A, 3 D, 4 C. **3.12** a) $\sum x_i/n$; b) $\sum x_i^3/n$; c) $\sum (x_i - \bar{x})^4/n$. **3.13** Šugu 0,499, rahvus 0,407. **3.14** $V = \sqrt{(1-p)/p}$. **3.15** Kvartiilhaare on toodud tabelis. Töötlevas tööstuses palkade hajuvus suurenes, hulgi- ja jaekaubanduses jäi samaks ning veonduse ja laonduse tegevusalal vähenes.

Tegevusala	2008	2013
Töötlev tööstus	0,44	0,47
Hulgi- ja jaekaubandus	0,52	0,52
Veondus ja laondus	0,40	0,39

A.3.1 Vastused on tabelis. Kõige rohkem varieerub elektrienergia toodang, sest sellel on variatsioonikordaja kõige suurem.

	Tootmiskulud, tuh $	Elektrienergia toodang aastas, mld kWh	Tunnitasu, $/h
Aritmeetiline keskmine	765	47,32	1,978
Standardhälve	506,8	41,0	0,194
Variatsioonikordaja	66,2%	86,6%	9,8%

A.3.2 Vastused on tabelis. Kõige rohkem varieerusid AEG kohvimasinate hinnad, kõige vähem Philipsi omad.

	Terve kogum	AEG	Moulinex	Philips	Severin
Aritmeetiline keskmine	53,4	64,6	50,0	39,7	38,7
Standardhälve	29,2	38,5	17,2	8,1	12,5
Variatsioonikordaja	54,7%	59,5%	34,4%	20,4%	32,3%

A.3.3 1. Valim 1: $\bar{x} = 0{,}1671\,\text{cm}$, $\sigma = 0{,}002\,21\,\text{cm}$; valim 2: $\bar{x} = 0{,}1664\,\text{cm}$, $\sigma = 0{,}003\,28\,\text{cm}$. 2. Valim 1: $V = 0{,}0132$, valim 2: $V = 0{,}0197$. Teise tootmisliini toodangu mõõtmed varieeruvad rohkem. 3. 0,1667 cm. 4. Tegelik keskmine 0,1694 cm,

tegelik standardhälve 0,003 28 cm ja variatsioonikordaja 0,0193. 5. Ei mõjutanud. **A.3.4** a) 0,053; b) 0,11. **A.3.5** Töötuse määr $z = 0{,}592$, mediaanvanus $z = 0{,}0494$. Eesti elanike mediaanvanus on Euroopa keskmisele lähemal. **A.3.6** Aastal 1961 Montrose veini hinna $z = -1{,}04$, aastal 1962 $z = -0{,}88$. Järelikult on 1961. aastal erinevus suurem. **A.3.7** a) Kõige sümmeetrilisem on SKP kasvumäär aastas, $A = 0{,}089$. b) Kõige suurema positiivse asümmeetriaga on imikute suremus, $A = 1{,}22$. c) Kõige suurema negatiivse asümmeetriaga on põllumajanduses toodetud lisandväärtuse kasvumäär aastas, $A = -1{,}278$. **A.3.8** Vastused on tabelis.

	FI	DK1	DK2	EE
Aritmeetiline keskmine	1,11%	1,44%	1,48%	0,95%
Miinimum	−55,2%	−41,2%	−47,1%	−31,6%
Maksimum	115,3%	127,4%	128,0%	49,6%
Asümmeetriakordaja	2,35	2,51	2,41	0,65
Püstakuse kordaja	12,06	13,39	13,12	0,991

Nii Soome kui ka mõlemas Taani hinnapiirkonnas on positiivne asümmeetria ja suur püstakus. See tähendab, et aritmeetilisest keskmisest suuremaid muutusi on rohkem kui väiksemaid (positiivne asümmeetria), aga ekstreemsed muutused on harva esinevad (suur püstakus). On näha, et hinnamuutuste maksimumväärtused on väga ekstreemsed.

Eesti hinnapiirkonnas on muutuste jaotus rohkem sümmeetrilisem, kuigi esineb ka nõrk positiivne asümmeetria. Jaotus on oluliselt lamedam kui ülejäänud kolmes piirkonnas: erineva suurusega muutuste esinemine on ühtlasema sagedusega.

A.3.9 1. Keskmine eluiga on nõrgalt vasakpoolse asümmeetriaga, $A = -0{,}49$, ülejäänud kaks tugeva parempoolse asümmeetriaga: „inimesi ühe televiisori kohta“ $A = 3{,}41$ ja „inimesi ühe arsti kohta“ $A = 4{,}65$. 2. Tunnusel „inimesi ühe arsti kohta“ on jaotuse „saba“ lamedam ja kaugemale ulatuv, asümmeetriakordaja on suurem ja püstakus on suurem ($E = 24{,}5$). 3. Kõige rohkem varieerub tunnus „inimesi ühe arsti kohta“, sest variatsioonikordaja on kõige suurem, $V = 2{,}508$. **A.3.10** 1. 4. 2. 10,8 tuhat \$. 3. Rohkem varieerub hind, variatsioonikordaja 51%. 4. Silindrite arv, $E = -0{,}2$. 5. Mujal toodetud on võimsamad, mootori keskmine võimsus 147,5 hj. USA autodel on keskmine võimsus 139,9 hj. **A.3.11** 1. 1200 W. 2. 41,85 kuni 65,81 eurot. 3. 18,53 eurot. **A.3.12** Toit ja mittealkohoolsed joogid: detsiilhaare 1221,34 eurot, suhteline detsiilhaare 1,44. Side: detsiilhaare 282,25 eurot, suhteline detsiilhaare 1,61. Kulutused sidele varieeruvad rohkem, sest suhteline detsiilhaare on suurem. **A.3.13** 1.–3. Vastused on tabelis ÜV.5. 4. Kõige riskantsem on Nordea Pensionifond A, kõige vähem riskantsem on LHV Pensionifond S.

A.3.14 a) Sturgesi valemist klasside arv 10; klassi laius 52 435, klassi laius ümardatult 55 000. Sagedustabel sõltub sellest, mis võtta esimese klassi ülemiseks piiriks. Üks võimalikke sagedustabeleid (viimase klassi ülemine piir peab olema suurem kui maksimumväärtus) on esitatud tabelis ÜV.6. Selline histogramm pole ilmselt sobiv, sest enamik väärtusi on koondunud ühte klassi. b) Kvartiilhaare on 7831,95, klassi laius Freedmani valemist 1957. Ümardatult võib võtta näiteks 2000. Kui esimese

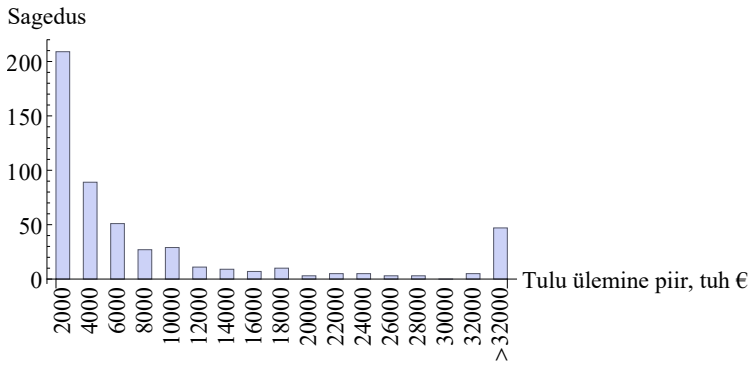
Tabel ÜV.5. Ülesande A.3.13 vastus

	LHV S	Nordea A	SEB	Swedbank K2
Variatsioonamplituud	9,13	24,73	13,62	11,95
Astak	4	1	2	3
Kvartiilhälve	5,04	9,2	4,75	5,59
Astak	3	1	4	2
Standardhälve	2,68	6,03	3,32	3,49
Astak	4	1	3	2
Variatsioonikordaja	0,804	0,617	0,925	0,777
Astak	2	4	1	3
Astakute summa	13	7	10	10

Tabel ÜV.6. Ülesande A.3.14 a) osa vastus

Ülemine piir	Sagedus	Ülemine piir	Sagedus	Ülemine piir	Sagedus
100	12	220 100	5	440 100	0
55 100	476	275 100	1	525 000	1
110 100	14	330 100	1		
165 100	3	385 100	0		

klassi ülemiseks piiriks võtta 2000, siis tuleb 34 klassi. Kuid alates klassist ülemise piiriga 20 000 on klassides väga vähe liikmeid. Seepärast võib klassi laiust 2000 kasutada näiteks klassini 32 000 ja kõik need, mille väärtus on suurem kui 32 000, koondada viimasesse klassi. Selliseid väärtusi tuleb 47. Sellise sagedustabeli abil saadud histogramm on joonisel ÜV.5.



Joonis ÜV.5. Ülesande A.3.14 b) osa vastuse juurde

4. ptk. Tõenäosusteooria elemente

4.1 1. 52. 2. Teoreetiline. 3. 1/52. 4. 1/13. 5. 2/13. 6. 0. **4.2** 1/3. **4.3**
1. Katsetulemusi on 16 ja nende hulk S_1:

$$S_1 = \left\{ \begin{array}{cccc} (1,1) & (1,2) & (1,3) & (1,4) \\ (2,1) & (2,2) & (2,3) & (2,4) \\ (3,1) & (3,2) & (3,3) & (3,4) \\ (4,1) & (4,2) & (4,3) & (4,4) \end{array} \right\}$$

2. Katsetulemusi on 12 ja nende hulk S_2:

$$S_2 = \left\{ \begin{array}{ccc} (1,2) & (1,3) & (1,4) \\ (2,1) & (2,3) & (2,4) \\ (3,1) & (3,2) & (3,4) \\ (4,1) & (4,2) & (4,3) \end{array} \right\}$$

4.4 a) $1/13 \approx 0{,}0769$; b) $4/51 \approx 0{,}0784$; c) $3/51 \approx 0{,}0588$. **4.5** a) 0,026; b) 0,076. **4.6** 0,75. **4.7** 0,045. **4.8** a) $\approx 0{,}937$; b) $\approx 0{,}00161$. **4.9** a) 0,32; b) 0,6. **4.10** a) 0,5; b) 0,5; c) 0,1. **4.11** Ligikaudu kahel nädalal. **4.12** a) 0,25; b) 0,5. **4.13** a) 0,633; b) 0,367; c) 0,117; d) 0,283; e) 0,40; f) 0,60; g) 0,529. **4.14** 0,5. **4.15** a) 0,25; b) 0,25; c) 0,5. **4.16** 15. **4.17** 1. 0,16. 2. 0,512. **4.18** a) 0,729; b) 0,009; c) 0,0009. **4.19** a) 0,65; b) 0,319; c) 0,102. **4.20** 0,38. **4.21** 0,63. **4.22** 1. 0,1925. 2. 0,7075. 3. 0,2925. **4.23** a) 0,04; b) 0,54; c) 0,42. **4.24** 0,0351. **4.25** 0,375. **4.26** a) 140; b) 180; c) 390. **4.27** 1. 0,2 2. 0,6. 3. Ei ole sõltumatud, sest $P(B|S) \neq P(B)$, kus B on sündmus „bakalaureusetöö hinne 5“ ja S tähistab sündmust „statistika hinne 5“. **4.28** a) 0,065; b) 0,084; c) 0,162. **4.29** 0,786. **4.30** 1. 0,25. 2. 0,3. **4.31** 0,4. **4.32** 1. 0,12. 2. 0,625. **4.33** 0,601. **4.34** Jah, sest tõenäosus, et see on rämpskiri, on 0,94. **4.35** 0,208, jah. **4.36** Valem, mida tuntakse ka Bayesi valemina:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})}. \qquad \text{(ÜV.1)}$$

5. ptk. Juhusliku suuruse jaotusseadused

5.1 1. $0{,}1 + 0{,}2 + 0{,}4 + 0{,}2 + 0{,}1 = 1$. 2. $P(X = 10) + P(X = 11) = 0{,}3$. 3. $P(X > 12) = 0{,}3$. 4. $P(X > 14) = 0$. **5.2** 1. Jaotusseadus on tabeli kujul. 3. $0{,}15 + 0{,}25 + 0{,}40 + 0{,}20 = 1$.

Kasutatud operatsiooniruumide arv	1	2	3	4
Tõenäosus	0,15	0,25	0,40	0,20

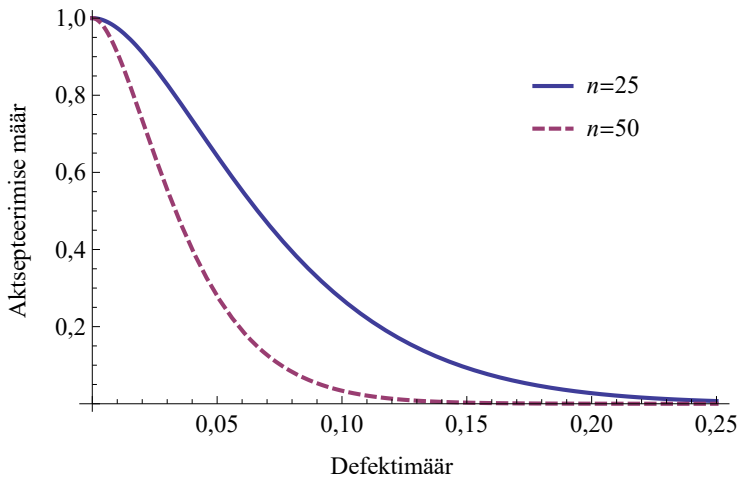
5.3 1. Jaotusseadus on tabelis. 2. 0,211. 3. 0,43.

x_i	1	2	3	4	5
p_i	0,098	0,272	0,211	0,227	0,193

5.4 a) Vastused on tabelis; b) 92%; c) 8%.

Töötajate arv	Jaotusfunktsiooni väärtus
0	47%
1	86%
2	92%
3	94%
4	96%
5 või rohkem	100%

5.5 Projekti A oodatav kasum on 5600 eurot ja projekti B oodatav kasum 4350 eurot. Valida tuleb projekt A, sest selle oodatav kasum on suurem. **5.6** Väikese bussi korral on oodatav kulu 190 tuhat eurot aastas, suure bussi korral 200 tuhat eurot aastas. Valida tuleb väike buss, sest oodatav kulu on väiksem. **5.7** 1. Oodatavad väärtused: suur kogus −16 tuhat eurot, väike kogus 8 tuhat eurot, valida tuleb väikese koguse tellimine. 2. Suur kogus tuleb tellida siis, kui päikeselise suve tõenäosus on suurem kui $5/6 \approx 0{,}83$. **5.8** Oodatav kasum on 31,7 tuhat eurot kuus. **5.9** Oodatav kasum on 23 eurot päevas. **5.10** Oodatav kasum on 9,1 eurot. **5.11** 36,25 päeva. **5.12** 1. 200 eurot. 2. Neli broneeringut, s.t üks ülebroneering. Sellisel juhul on oodatav tulu 240 eurot. Viie korral on 235 ja kuue korral 200 eurot. **5.13** 1. Jaotustihedus lõigus $[20, 60]$ $f(x) = 0{,}025$. Jaotusfunktsioon $F(x) = 0{,}025x - 0{,}5$. 2. a) 0,125; b) 0,25; c) 0,625; d) 0,375. **5.14** Keskväärtus ja mediaan on 275 eurot, standardhälve 14,43 eurot. Tõenäosus, et hind on suurem kui 280 eurot, on 0,4. **5.15** 914. **5.17** 0,156. **5.18** a) $4{,}81 \cdot 10^{-5}$; b) 0,00985; c) 0,173; d) 0,706; e) 0,959. **5.19** a) 0,262; b) 0,393; c) 0,246; d) 0,0989. **5.20** a) 0,349; b) 0,302. **5.21** 7,5%. **5.22** Vt joonis ÜV.6



Joonis ÜV.6. Ülesande 5.22 vastus

5.23 1. 0,87. 2. Keskväärtus 1,6, standardhälve 0,98. 3. 160 tuhat dollarit. 4. Kahjum tekib siis, kui õnnestunud puuraukude arv on 0 või 1. Selle tõenäosus on 0,475.

Järelikult on kasumi saamise tõenäosus suurem, 0,525. **5.24** 1. a) 0,642; b) $9{,}54 \cdot 10^{-27}$. 2. 0,736. 3. 54,5 eurot. 4. 54,5 eurot. **5.25** 1. 15. 2. Üheksa serverit. Seda saab leida kahel moel. Üks võimalus on leida tõenäosus $P(X \leq m)$ erineva konkureerivate kasutajate arvu m korral. Näiteks seitsme serveri korral $P(X \leq 350) \approx 23{,}95\%$, kaheksa serveri korral $P(X \leq 400) \approx 99{,}87\%$. Teine võimalus on kasutada tabelarvutusfunktsiooni BINOM.INV(720;0,5;0,9999)=410. Järelikult peab servereid olema ühe võrra rohkem kui kaheksa. 3. $5{,}97 \cdot 10^{-12}$. **5.26** 1. 5,7%. 2. 1,7. 3. 24,5%. **5.27** a) 9; b) 3. **5.28** Kolme broneeringu korral 288,9 eurot, nelja korral 256,2 ja viie korral 209,18 eurot. **5.29** 1. $Y_{max} = 4$, sest kui broneeringute arv $n = 102$, siis $P(m > 98) = 0{,}0067$ ja kui $Y = 5$ ning $n = 103$, siis $P(m > 98) = 0{,}0192$, kus m on kinnitatud broneeringute arv. 2. Oodatav kasum on maksimaalne 11 ülebroneeringu korral. Mõningad oodatava kasumi väärtused on toodud tabelis ÜV.7. Näeme, et ülebroneerimiste arvu suurenemisel oodatav kasum kasvab kuni ülebroneerimiste arv saavutab väärtuse 11 ja seejärel hakkab kahanema.

Tabel ÜV.7. Ülesande 5.29 vastuse juurde

Ülebroneerimiste arv	Oodatav kasum, €
1	200,0
2	399,8
3	599,1
...	...
10	1612,4
11	1618,5
12	1578,7
13	1497,4

5.30 $V = \sqrt{\frac{1-p}{np}}$. Katsete arvu n suurenedes variatsioonikordaja väheneb. Täpsemalt: V on pöördvõrdeline ruutjuurega katsete arvust. **5.31** a) 0,69; b) 0,006. **5.32** 1. Vastused on tabelis. 2. 0,00285. 3. 0,105.

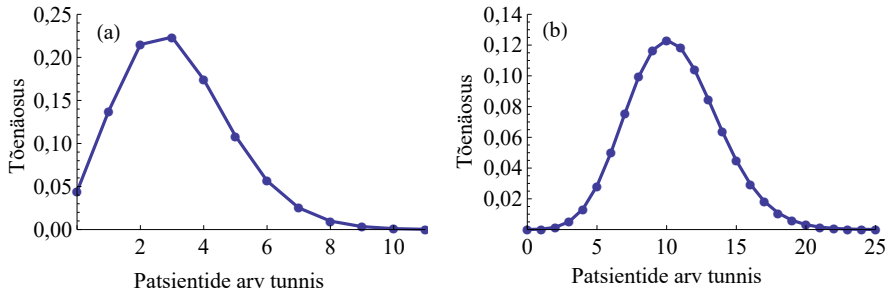
m	0	1	2	3	4	5
$P(X = m)$	0,255	0,348	0,238	0,108	0,037	0,010

5.33 a) 0,29; b) 0,446. **5.34** 1. 0,195. 2. 0,105. 3. 0,091. **5.35** 1. 0,184. 2. Kahel kuul aastas. 3. 13. **5.36** 1. Joonis ÜV.7 (a). 2. $9{,}93 \cdot 10^{-5}$. 3. 10,558. 4. Joonis ÜV.7 (b).

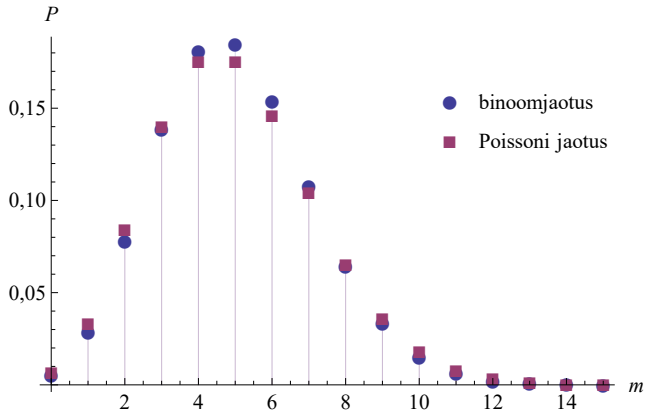
5.37 Vastused on tabelis.

Päevas unustatud kaartide arv	0 kuni 23	24 kuni 26	27 kuni 29	30 kuni 32	33 ja rohkem
Oodatav päevade arv	6,96	6,74	7,35	5,89	6,07

5.38 Vt joonis ÜV.8.



Joonis ÜV.7. Ülesande 5.36 1. osa vastus (a) ja 4. osa vastus (b)



Joonis ÜV.8. Ülesande 5.38 vastus

5.40 1. $f(x) = 0{,}0833e^{-0{,}0833x}$, $F(x) = 1-e^{-0{,}833x}$. 2. 0,565. 3. 0,082. **5.41** a) 1,7 tundi; b) 0,05. **5.42** Tõenäosus hommikupoole 0,0872, keskpäeval 0,194. **5.43** 2,12%. **5.44** 1,5 aastat. **5.45** a) 0,218; b) 0,0476; c) 0,218. **5.46** Kui eeldada, et klientide saabumise vaheline aeg on eksponentjaotusele alluv juhuslik suurus, siis tõenäosus, et järgmist klienti tuleb oodata vähemalt viis minutit, ei sõltu sellest, kui palju teda on juba oodatud. Järelikult, osakonnajuhataja põhjendus ei ole loogiline. **5.47** 1. $e^{-\lambda}$. 2. $e^{-\lambda}$. **5.48** a) 0,149; b) 0,182; c) 0,669. **5.49** 0,447. **5.50** 2. Esimesel tänaval on selliste mootorrataste osakaal suurem. Esimesel tänaval $P(X > 40) \approx 0{,}088$, teisel tänaval $P(X > 40) \approx 0{,}080$. **5.51** 1. 25,2%. 2. 28%. 3. 0,98%. 4. 86,1%. **5.52** 1. 76%. 2. 5,92 ehk *ca* 6 minutit. **5.53** a) $f(\mu) = \frac{1}{\sqrt{2\pi}} \approx 0{,}4$; b) $f(\mu) = \frac{1}{4\sqrt{2\pi}} \approx 0{,}1$. **5.54** 1,67 ml. **5.55** 1095. **5.56** $0{,}65 \pm 0{,}12$. **5.57** 91,26 eurot. **5.58** 9. kümnendikus, sest $P(X < 1{,}667) \approx 0{,}811$. **5.59** a) 0,135%, 1350 tk/mln; b) 0,00034%, 3,4 tk/mln; c) $9{,}9 \cdot 10^{-10}$, 0,000 99 tk/mln ehk *ca* 1 tk/mld. **5.60** 1. $\mu \approx 5{,}56$, $\sigma \approx 0{,}934$. a) 0,562; b) 0,184. 2. 0,385. 3. 0,421. Nüüd on suurem. **5.61** Mitte ühtegi avariid 19 nädalal, 3 avariid kolmel nädalal. **5.62** a) 0,143; b) 0,020. **5.63** a) 0,0498; b) 0,147; c) 0,353; d) 0,000912. **5.64** 1. 0,00248. 2. 0,22. **5.65** 0,12; kuuel nädalal. **5.66** Kahe vastusevariandi korral 0,412, kolme vastusevariandi korral 0,038. **5.67** 0,223. **5.68** 0,175.

5.69 1. 0,189. 2. 0,274. 3. 0,922. **5.70** 1. 0,535. 2. 0,188. **A.5.1** 1. 0,307% 2. 7,54 tuhat eurot; 3. 11,6 tuhat eurot; 4. 26,1 tuhat eurot. **A.5.2** 1. Keskväärtus on 24,4 katkestust aastas ja dispersioon 26. 2. Kuna keskväärtus ja dispersioon eriti palju ei erine, siis võib eeldada, et katkestuste arv aastas allub Poissoni jaotusele. 3. Kahel aastal. **A.5.3** 4083 liitrit. **A.5.4** Alumine barjäär 1,127; ülemine barjäär 1,145. **A.5.5** 1. Aritmeetiline keskmine 190,05 g ja standardhälve 0,7046 g. 2. 15,7%.

6. ptk. Valikuuringud

6.1 1469 ± 38 AUD usaldatavusega 0,9; 1469 ± 45 AUD usaldatavusega 0,95. **6.2** 14,97 ± 2,55 tundi aastas. **6.3** 300. **6.4** 12,1. **6.5** 0,99. **6.6** 16,6. **6.7** 25 korda. **6.8** 29,3%, 10,6%. **6.9** Usaldatavuse $\beta = 0{,}95$ korral $z_{0{,}025} \approx 1{,}96$. Kui $\beta = 0{,}99$, siis $z_{0{,}005} \approx 2{,}58$. Kui $\beta = 0{,}999$, siis $z_{0{,}0005} \approx 3{,}29$. **6.10** 0,97. **6.11** Keskmine ülesande sooritamiseks kuluv aeg usaldatavusega 0,95 a) 13,0 ± 1,6 minutit; b) 13,0 ± 1,4 minutit. **6.12** Vastused on tabelis.

Tegevusala	Parandusliige
Jäätmetöötlus ja -kõrvaldus	0,559
Ehitus	0,928
Jaemüük posti või Interneti teel	0,966

6.13 a) 0,022±0,008 mg/kg; b) 0,142±0,12 mg/kg. **6.14** 0,9. **6.15** 1515. **6.16** Vajalik valimi maht tuleb 819. Lisaks on vaja küsitleda veel 619 isikut. **6.17** 12,0% ± 1,6%. **6.18** a) 7,93% ± 0,60%; b) 420,9 ± 43,9 €. **6.19** Vastused on tabelis.

Ettevõtte suurusgrupp	Valimi maht n
5–9	152
10–49	453
50–249	207
250+	44

6.20 1. 2,8%. 2. 18% ± 2,2%. **6.21** Kasutada tuleb korrigeeritud usalduspiire: 1,34% ± 0,90% usaldatavusega 0,9. **6.22** 9601. **6.23** Enda teenitud palk 48,0% ± 3,6%, perekonnalt/partnerilt saadud raha 33,0%±3,4%, riigi toetused 15,0%±2,6%. Märkus: nende usalduspiiride leidmisel võib kasutada lihtsamat ligikaudset valemit (6.43). Kuna variandi „muu“ valinute arv on 49 (< 100), siis selle usaldusvahemiku leidmiseks tuleks kasutada täpset, kuid keerulisemat valemit. **6.24** Kuna variantide „rahuldav“ ja „mitterahuldav“ valijate arv oli väga väike, tuleb kasutada valemeid (6.45)–(6.47). „Hea“ 78,1%–95,3%, „rahuldav“ 2,4%–16,9%, „mitterahuldav“ 1,1%–13,3%. **6.25** a) Alumiseks piiriks on 228. elemendi väärtus ja ülemiseks piiriks 273. elemendi väärtus; b) alumiseks piiriks on 221. elemendi väärtus ja ülemiseks piiriks 280. elemendi väärtus.

A.6.1 a) Valimi keskmine 180,8 g, valimi dispersioon 88,7 g^2, valimi standardhälve 9,42 g; b) 1,33 g; c) 180,8 ± 2,6 g. **A.6.2** 1. Usalduspiirid ja suhtelised vead

on tabelis. 2. Kattuvad inspektor ja tehnik, spetsialist ja kooliinspektor, direktor ja sõltumatu konsultant. 3. Kõige täpsemini on määratud mänedžeride keskmine palk. Täpsus tuleb kõige suurem sellepärast, et valimi maht on oluliselt suurem kui teiste ametite korral.

Amet	Δx	Alumine piir	Ülemine piir	Suhteline viga
Inspektor	915	25 183	27 013	3,5%
Tehnik	586	26 798	27 970	2,1%
Koordinaator	1947	34 972	38 866	5,3%
Spetsialist	1365	40 745	43 475	3,2%
Kooliinspektor	1302	41 397	44 001	3,0%
Insener	965	45 851	47 781	2,1%
Mänedžer	1039	54 037	56 115	1,9%
Konsultant	2499	58 102	63 100	4,1%
Direktor	1682	65 657	69 021	2,5%
Sõltumatu konsultant	4125	65 230	73 480	5,9%
Asepresident	3924	89 323	97 171	4,2%

A.6.3 1245 ± 46 eurot aastas. Üheliikmelisel leibkonnal kulub toidule rohkem. **A.6.4** 1. Põhja-Eestis 8390 ± 312 eurot, Kirde-Eestis 6178 ± 316 eurot. 2. Valimi standardhälve on Põhja-Eestis 7178,5 ja Kirde-Eestis 3939,7 eurot. Põhja-Eestis varieeruvad sissetulekud rohkem kui Kirde-Eestis. Usaldusvahemiku poollaius on Põhja-Eestis 312 ja Kirde-Eestis 316 eurot. Kuigi Põhja-Eestis on varieerumine oluliselt suurem, on usaldusvahemikud ligikaudu ühesugused, sest Põhja-Eestis on ka valimi maht suurem. 3. Põhja-Eestis $699{,}2 \pm 26{,}0$ eurot, Kirde-Eestis $514{,}9 \pm 26{,}3$ eurot. **A.6.5** Finantsvahenduses 1771 ± 2540 eurot ja kinnisvara valdkonnas 3209 ± 3727 eurot ühe töötaja kohta aastas. Ei saa väita. **A.6.6** 185 ± 71 mln \$ usaldatavusega 0,75; 185 ± 107 mln \$ usaldatavusega 0,9; 185 ± 134 mln \$ usaldatavusega 0,95. **A.6.7** Teadmusmahukad $43{,}4\% \pm 14{,}4\%$, traditsioonilised $30{,}9\% \pm 9{,}2\%$. **A.6.8** Normaaljaotust kasutades $31{,}52 \pm 4{,}19$ min, t-jaotust kasutades $31{,}52 \pm 4{,}42$ min. Tegemist on väikese valimiga ja normaaljaotust kasutades alahindame usaldusvahemiku laiust. **A.6.10** $36{,}32\% \pm 2{,}25\%$. **A.6.11** $24{,}6\% \pm 2{,}7\%$. **A.6.12** Usalduspiirid on tabelis. Muutusi ei toimunud, kõikide vastusevariantide oktoobri ja detsembri usaldusvahemikud kattuvad osaliselt.

	Okt		Dets	
Vastusevariant	Alumine piir	Ülemine piir	Alumine piir	Ülemine piir
tunduvalt rohkem	2,0%	5,4%	3,4%	7,6%
mõnevõrra rohkem	8,4%	14,3%	8,8%	14,7%
umbes samad	30,1%	38,9%	32,2%	41,2%
mõnevõrra vähem	12,5%	19,3%	11,4%	18,0%
tunduvalt vähem	29,4%	38,1%	25,5%	34,0%
ei oska öelda	1,1%	3,9%	1,7%	4,9%

A.6.13 Vastused on tabelis. Kuna 3. variandi valinute arv $37 < 100$, siis selle osakaalu usalduspiiride leidmisel tuleb kasutada valemeid (6.45)–(6.47).

Vastusevariant	Osakaal p	Δp	Alumine piir	Ülemine piir
1. Jah	46,6%	3,8%	42,9%	50,4%
2. Võiks olla valikaine	49,7%	3,8%	45,9%	53,5%
3. Ei	3,7%		2,5%	5,4%

A.6.14 (1094,4, 1179,5) eurot aastas. 80 eurot kuus on 960 eurot aastas, mis on väiksem kui mediaani alumine usalduspiir. Jah, sellised kulutused jäävad väiksema 50% hulka tõenäosusega 0,95. 95 eurot kuus on 1140 eurot aastas. See jääb mediaani usaldusvahemiku sisse ning ei saa otsustada, kas see on väiksem või suurem kui üldkogumi mediaan. **A.6.15** (36 581, 47 092) eurot aastas ühe töötaja kohta.

7. ptk. Hüpoteeside statistiline kontrollimine

7.1 H_0: $\mu = 4{,}3$, H_1: $\mu \neq 4{,}3$. **7.2** H_0: $\mu \geq 330\,\mathrm{g}$, H_1: $\mu < 330\,\mathrm{g}$. **7.3** H_0: $\mu = \mu_0$, H_1: $\mu \neq \mu_0$, kus μ_0 on linnaelanike keskmine vanus. **7.4** 1. H_0. 2. H_1. 3. H_1. 4. H_0. 5. H_1. 6. H_0. 7. H_1. **7.5** H_0: $p \leq 25\%$, H_1: $p > 25\%$. I liiki viga on, kui tegelikult ei hakka uut toodet tarbima üle 25% tarbijatest, kuid valimisse sattusid näiteks eriti uuendusmeelsed tarbijad ja seetõttu valimi põhjal otsustame, et tasub ära. Tehakse kulutusi uue toote juurutamiseks, kuid need kulutused tulu ei too. II liiki viga on, kui tegelikult hakkaks seda toodet tarbima rohkem kui 25% tarbijatest, kuid valim sattus selline, et otsustame: ei tasu uut toodet juurutada.

7.6 H_0: $\mu \geq 40$, H_1: $\mu < 40$. I liiki viga on, kui väljapakutud võtted tegelikult ei vähenda detaili valmistamiseks kulunud aega, kuid juhuvalim tuleb selline, et otsustame: vähendavad. Hakatakse kasutama uusi võtteid, mis midagi ei muuda. II liiki viga on, kui ettepanekud tegelikult vähendavad keskmist aega, kuid valimi põhjal otsustame, et ei vähenda. Kasulikke võtteid kasutama ei hakata.

7.7 Kahepoolne hüpotees: H_0: $\mu = 20$ aastat, H_1: $\mu \neq 20$ aastat. Teststatistik $z = -3{,}54$, parempoolne kriitiline väärtus nivool 0,05 on 1,96. Kuna $|-3{,}54| > 1{,}96$, on H_0 ümber lükatud ja keskmine tööstaaž ei ole 20 aastat. Voldikus olev väide ei vasta tõele ja seda tuleb muuta.

7.8 Ühepoolne hüpotees: H_0: $\mu \leq 2$, H_1: $\mu > 2$. Teststatistiku empiiriline väärtus on 8,82 ja kriitiline väärtus 1,64. Kuna $8{,}82 > 1{,}64$, võtta vastu H_1: keskmine tarnijate arv on suurem kui kaks.

7.9 Kahepoolne hüpotees: H_0: $\mu = 1$ toll, H_1: $\mu \neq 1$ toll. Teststatistiku empiiriline väärtus 3,54, kriitiline 2,58. Kuna $3{,}54 > 2{,}58$, võtta vastu H_1. Järeldus: tööpink vajab seadistamist.

7.10 Ühepoolne hüpotees: H_0: $\mu \geq 2{,}38$ päeva, H_1: $\mu < 2{,}38$ päeva. Teststatistiku empiiriline väärtus $-1{,}99$. a) Olulisuse nivool 5% on kriitiline väärtus $-1{,}645$, $-1{,}99 < -1{,}645$, võtta vastu H_1, uus süsteem on parem. b) Olulisuse nivool 1% on kriitiline väärtus $-2{,}33$. Kuna $-1{,}99 > -2{,}33$, võtta vastu H_0: uus süsteem ei ole parem.

7.11 Ühepoolne hüpotees: H_0: $\mu \leq 1{,}5$ m, H_1: $\mu > 1{,}5$ m. Teststatistiku empiiriline väärtus 3,81, olulisuse nivool 5% on kriitiline väärtus 1,645. Kuna $3{,}81 > 1{,}645$, võtta vastu H_1: kaubad on kõrgemal kui 1,5 m.

7.12 Kriitiline väärtus on 1,96. a) $z = 1{,}43 < 1{,}96$, võtta vastu H_0. b) $z = 2{,}02 > 1{,}96$, võtta vastu H_1. Valimi mahu suurenedes teststatistik suureneb ning suureneb võimalus nullhüpoteesi ümberlükkamiseks.

7.13 Kriitiline väärtus on 1,96. a) $z = 2{,}04 > 1{,}96$, võtta vastu H_1. b) $z = 1{,}36 < 1{,}96$, võtta vastu H_0. Mida suurem on hajuvus (standardhälve), seda väiksem on teststatistik ning seda väiksem on võimalus nullhüpotees ümber lükata.

7.14 c, d ja f.

7.15 b, c ja f.

7.16 Kuna tegemist on suurte valimitega, võib kriitilise väärtuse leidmiseks kasutada standardiseeritud normaaljaotust. Kahepoolse hüpoteesi korral olulisuse nivool 0,05 on kriitiline 1,96 ja olulisuse nivoo 0,01 korral 2,58. Nivool 0,05 on oluline erinevus omakapitali tootlusel, sest $|-2{,}093| > 1{,}96$. Nivool 0,01 on oluline erinevus varade tootlusel, sest $|-3{,}084| > 2{,}58$.

7.17 Kahepoolne hüpotees: H_0: $\sigma_1^2 = \sigma_2^2$, H_1: $\sigma_1^2 \neq \sigma_2^2$. Jagades suurema dispersiooni väiksemaga, saame F-statistiku väärtuseks 1,98. Parempoolne kriitiline väärtus olulisuse nivool 0,05 on 1,55. Kuna $1{,}98 > 1{,}55$, võtta vastu H_1. Vaimse tervise varieerumine töötute ja hõivatute hulgas on erinev.

7.18 Dispersioonide testimisel F-statistik 2,263, parempoolne kriitiline väärtus 1,8, võtta vastu H_1, dispersioonid on erinevad ja kasutada tuleb t-testi erinevate dispersioonide korral. Kui noormehed on valim 1 ja neiud valim 2, siis t-testi H_0: $\mu_1 \leq \mu_2$ ja H_1: $\mu_1 > \mu_2$. t-statistik 1,777, vabadusastmete arv 97, kriitiline väärtus nivool 0,05 on 1,66: võtta vastu H_1. Neiud mängivad *online* mänge vähem kui noormehed.

7.19 Olgu p rahulolematute klientide osakaal. Hüpoteesipaar on H_0: $p \leq 0{,}1$ ja H_1: $p > 0{,}1$. Teststatistiku empiiriline väärtus 5,01, kriitiline väärtus nivool 0,05 on 1,645. Kuna $5{,}01 > 1{,}64$, võtta vastu H_1. Järeldus: rahulolematus on oluliselt suurem kui 10%, klienditeeninduses tuleb teha muudatusi.

7.20 Tuludeklaratsiooni täitmiseks on kaks võimalust: kas paberil või e-maksuametis. Olgu paberil deklaratsiooni täitnute osakaal p. Paberil täideti rohkem siis, kui $p > 0{,}5$. Järelikult hüpoteesipaar: H_0: $p \leq 0{,}5$, H_1: $p > 0{,}5$. Valimi maht $n = 164 + 82 = 246$ ja paberil täitnute osakaal valimis $\hat{p} = 0{,}667$. Teststatistik $z = 5{,}23$. Kriitiline väärtus olulisuse nivool 0,01 on normaaljaotuse täiendkvantiil $z_{0{,}01} = 2{,}33$. Kuna $9{,}52 > 2{,}33$, võtta vastu H_1: 1999. aasta tuludeklaratsiooni täitis paberil rohkem inimesi kui e-maksuametis.

7.21 Püstitatud hüpoteesipaar: H_0: $p_1 \leq p_2$ ja H_1: $p_1 > p_2$, kus p_1 on töötute osakaal meeste hulgas ning p_2 töötute osakaal naiste hulgas. Vastavad osakaalud valimis $\hat{p}_1 = 12{,}2\%$ ja $\hat{p}_2 = 8{,}7\%$. Teststatistik $z = 4{,}8$. Kriitiline väärtus standardiseeritud normaaljaotusest $z_{0{,}01} = 2{,}33$. Kuna $4{,}8 > 2{,}33$, võtta vastu H_1. Töötute osakaal meeste hulgas on suurem kui naiste hulgas.

7.22 Kui p_1 on Reformierakonna toetajate osakaal esimese küsitluse ajal ja p_2 toetajate osakaal teise küsitluse ajal, siis H_0: $p_1 \geq p_2$ ja H_1: $p_1 < p_2$. Teststatistik $z = -1{,}561$, kriitiline väärtus ühepoolse hüpoteesi korral olulisuse nivool 0,05 on $-1{,}645$. Kuna $-1{,}561 > -1{,}645$, võtta vastu H_0: erakonna toetuse suurenemine ei ole tõestatud.

7.23 Suurem kui 26,17%.

7.24 Kõikidel juhtudel H_0: $N^+ \leq N^-$ ja H_1: $N^+ > N^-$, kus N^+ on ainult nooremale vastanud ettevõtete arv ja N^- ainult vanemale vastanud ettevõtete arv. Testimise tulemused on tabelis.

	Ainult nooremale	Parempoolne kriitiline n_{krp}	Vastu võetud hüpotees
Kogu valim	34	27	H_1
Müügiagent	12	11	H_1
Abitööline restoranis	22	18	H_1
Täistööaeg	25	21	H_1
Osaline tööaeg	9	9	H_0
Alaline töökoht	15	14	H_1
Ajutine töökoht	19	16	H_1

7.25 Olgu N^+ nende arv, kes peavad veini Cabernet Sauvignon paremaks ja N^- nende arv, kes peavad seda veini halvemaks. Siis H_0: $N^+ = N^-$ ja H_1: $N^+ \neq N^-$. Korrigeeritud valimi maht $n = 9$, esimest veini hindas paremaks 4 eksperti. Kriitilised väärtused olulisuse nivool 0,05 on 2 ja 7. Kuna $2 < 4 < 7$, tuleb jääda nullhüpoteesi juurde: hinnang veinidele on ühesugune.

7.26 1. H_0: müügitulu kasv ja arendustegevus ei ole seotud, H_1: müügitulu kasv ja arendustegevus on seotud. 2. Oodatavad sagedused on tabelis. 3. $\chi^2 = 4{,}558$. 4. Kriitiline 3,84. Kuna $4{,}558 > 3{,}84$, võtta vastu H_1. Müügitulu kasv on arendustegevusega seotud.

	Müügitulu kasv	
Toote arendus	Ei	Jah
Ei	3,12	13,88
Jah	7,88	35,12

7.27 a) Üldine 119, rühmadevaheline 2, rühmasisene 117; b) üldine 149, rühmadevaheline 5, rühmasisene 145.

7.28 Vastused on tabelis.

Hajuvuse allikas	*SS*	*df*	*MS*	*F*
Rühmadevaheline	385,8	2	192,9	0,761
Rühmasisene	6847,3	27	253,60	
Üldine	7233,1	29		

7.29 1. Belgia, Taani, Prantsusmaa, Saksamaa, Holland. 2. Belgia, Soome, Saksamaa.

7.30 Olgu perefirmad valim 1 ja mitte-perefirmad valim 2. 1. Kahe kogumi osakaalude võrdlemine suurte valimite korral, H_0: $p_1 \leq p_2$ ja H_1: $p_1 > p_2$. Teststatistik $z = 10{,}869$, kriitiline nivool 0,05 on 1,64: võtta vastu H_1. Perefirmad on sagedamini

esikümne hulgas kui mitte-perefirmad. 2. Keskväärtuste võrdlemine t-testiga sõltumatute valimite korral, t-testi valikuks tuleb eelnevalt testida dispersioone. F-test näitab, et dispersioonid on ühesugused: F-statistik on 0,534 ja vasakpoolne kriitiline väärtus 0,705 (kahepoolne hüpotees). Järelikult tuleb kasutada sõltumatute valimite t-testi võrdsete dispersioonide korral. H_0: $\mu_1 \geq \mu_2$ ja H_1: $\mu_1 < \mu_2$. Teststatistik $t = -1{,}718$. Kuna valimid on suured, võib kriitilise väärtuse võtta normaaljaotusest ja vasakpoolse hüpoteesi korral on see $-1{,}64$. Kuna $-1{,}718 < -1{,}64$, tuleb vastu võtta sisukas hüpotees: perefirmades on keskmine käive ühe töötaja kohta väiksem.

A.7.1 H_0: $\mu = 190$, H_1: $\mu \neq 190$. Kriitiline väärtus olulisuse nivool 0,05 on 1,96. Tootmisliini L621 korral $z = 0{,}631 < 1{,}96$, võtta vastu H_0, pudelite mass vastab nominaalsele. Tootmisliini L635 korral $z = 9{,}69 > 1{,}96$, võtta vastu H_1, pudelite mass ei vasta nominaalsele.

A.7.2 Statistiku empiiriline väärtus 1,92. a) H_0: $\mu = 1500$, H_1: $\mu \neq 1500$, kriitiline väärtus olulisuse nivool 5% on 1,96, võtta vastu H_0, tootmisliin ei vaja reguleerimist; b) H_0: $\mu \leq 1500$, H_1: $\mu > 1500$, kriitiline väärtus olulisuse nivool 5% on 1,645, võtta vastu H_1, tootmisliin vajab reguleerimist.

A.7.3 H_0: $\mu \geq 15{,}4$, H_1: $\mu < 15{,}4$. Statistiku empiiriline väärtus $-1{,}901$. Kriitiline väärtus olulisuse nivool 5% on $-2{,}132$ (ühepoolne hüpotees). Kuna $-1{,}901 > -2{,}132$, tuleb vastu võtta nullhüpotees. Kõrvaklappide mürasummutus vastab kasutusjuhendis olevale väärtusele.

A.7.4 H_0: $\mu \leq 0$, H_1: $\mu > 0$. Statistiku empiiriline väärtus 2,5. Kriiline väärtus olulisuse nivool 0,05 on 1,89, võtta vastu sisukas hüpotees ja analüütikul on õigus. Olulisuse nivool 0,01 on kriitiline väärtus 3,00 ja võtta vastu nullhüpotees, analüütiku väide pole tõestatud. Järeldus on, et väide väga tugevalt tõestatud pole. Kui investor võib lubada endale ka kahjusid, siis võib Appeli reeglit järgida (kasutab olulisuse nivood 0,05).

A.7.5 Mõlemal juhul H_0: $\mu_1 = \mu_2$ ja H_1: $\mu_1 \neq \mu_2$. a) Statistik $t = 1{,}718$ on suurem kui kriitiline väärtus 1,692, võtta vastu H_1. Olulisuse nivool 0,1 on arenduskulud lennuki- ja masinatööstuse ettevõtetes erinevad. b) Statistik $t = -0{,}414$ jääb kriitiliste väärtuste $-1{,}686$ ja 1,686 vahele, ei lange kriitilisse piirkonda ning tuleb vastu võtta nullhüpotees. Olulisuse nivool 0,1 ei ole arenduskulud lennuki- ja keemiatööstuse ettevõtetes erinevad.

A.7.6 Kõikidel juhtudel H_0: $\mu_1 = \mu_2$ ja H_1: $\mu_1 \neq \mu_2$. Kuna valimite mahud on suured, leiame kriitilise väärtuse standardiseeritud normaaljaotusest ja see on 1,96. Tabelist on näha, et erinevad on finantsvõimenduse kordaja, võlakordaja ja dividendide maksmise tase.

Suhtarv	t-statistik	Võtta vastu H_1
Omakapitali tootlus (ROE)	$-1{,}66$	
Maksueelse kasumi osatähtsus netokäibes	$-1{,}35$	
Finantsvõimenduse kordaja	$-2{,}94$	H_1
Varade käibekordaja	1,66	
Võlakordaja	$-3{,}21$	H_1
Dividendide maksmise tase, % (1995. a.)	$-3{,}02$	H_1

A.7.7 Kõikidel juhtudel H_0: $\mu_1 = \mu_2$ ja H_1: $\mu_1 \neq \mu_2$. Kuna valimite mahud on suured, leiame kriitilise väärtuse standardiseeritud normaaljaotusest ja see on 1,96. Tabelist on näha, et erinevad on töötajate arv, ambulatoorsete visiitide arv ja residentide arv.

Näitaja	t-statistik	Võtta vastu H_1
Teenuste arv	−1,64	
Töötajate arv	2,86	H_1
Kulud	1,33	
Haigusjuhtude arv	0,29	
Ambulatoorsete visiitide arv	5,42	H_1
Residentide arv	4,55	H_1

A.7.8 Kui Põhja-Eesti elanikud on valim 1 ja Kirde-Eesti elanikud valim 2, siis H_0: $\mu_1 \geq \mu_2$ ja H_1: $\mu_1 < \mu_2$. Olulisuse tõenäosus 0,0329 < 0,05, järelikult H_0 on ümber lükatud. Kui võrrelda t-statistikut ja vastavat kriitilist väärtust, siis $-1{,}846 < -1{,}654$, nullhüpotees on ümber lükatud. Olulisuse nivool 0,05 on tõestatud, et Kirde-Eesti töötute hulgas on töötuse kestvus keskmiselt suurem kui Põhja-Eesti töötute hulgas.

A.7.9 Kui D on keskmine hinnapakkumiste erinevus, siis H_0: $D \leq 0$ ja H_1: $D > 0$. 1. Erinevuste valimi keskmine on 32,37 $ ja standardhälve 78,68 $. Teststatistik $t = 2{,}25$, kriitiline väärtus nivool 0,05 on 1,7. Kuna 2,25 > 1,7, on H_0 ümber lükatud ja tõestatud, et informeeritumale ostjale pakuti keskmiselt madalamat hinda. 2. Lisaks teststatistiku võrdlemisele kriitilisega võib otsuse vastuvõtmisel lähtuda olulisuse tõenäosusest $p = 0{,}016$, mis on väiksem kui olulisuse nivoo 0,05 ning H_0 on ümber lükatud.

A.7.10 Kui D on keskmine müügimahtude erinevus enne ja peale reklaami, siis H_0: $D \geq 0$ ja H_1: $D < 0$. Olulisuse tõenäosus tuleb 0,02 < 0,05, järelikult tuleb vastu võtta sisukas hüpotees: peale reklaami toodete müügimaht suurenes. Võib ka võrrelda t-statistikut −2,36 ja vastavat vasakpoolset kriitilist väärtust −1,81.

A.7.11 Kui mehed on valim 1 ja naised valim 2, siis H_0: $\sigma_1^2 \leq \sigma_2^2$, H_1: $\sigma_1^2 > \sigma_2^2$. Statistiku empiiriline väärtus on 1,35 ja kriitiline väärtus 1,18. Kuna 1,35 > 1,18, võtta vastu H_1: meestel varieeruvad kulutused vabale ajale rohkem kui naistel. Statistiku empiirilise ja kriitilise väärtuse asemel võime vaadata ka olulisuse tõenäosust: $p = 0{,}00149 < 0{,}05$, võtta vastu H_1.

A.7.12 Olgu neljapäevased tulumäärad valim 1 ja ülejäänud päevade tulumäärad valim 2. Otsustamaks, kumba sõltumatute valimite t-testi kasutada, testime dispersioone kahepoolse F-testiga, kus H_0: $\sigma_1^2 = \sigma_2^2$ ja H_1: $\sigma_1^2 \neq \sigma_2^2$. F-testi olulisuse tõenäosus on $p = 0{,}996 > 0{,}05$, võtta vastu H_0 ja järelikult dispersioonid ei ole erinevad. Keskmise tulumäära võrdlemiseks kasutame ühepoolset t-testi võrdsete dispersioonide korral, H_0: $\mu_1 \geq \mu_2$, H_1: $\mu_1 < \mu_2$. Olulisuse tõenäosus ühepoolse hüpoteesi korral $p = 0{,}037 < 0{,}05$, võtta vastu H_1. Neljapäeviti on tulumäär väiksem kui ülejäänud nädalapäevadel.

A.7.13 Olgu äriklass valim 1 ja turistiklass valim 2. Otsustamaks, kumba sõltumatute valimite t-testi kasutada, testime dispersioone kahepoolse F-testiga, kus

H_0: $\sigma_1^2 = \sigma_2^2$, H_1: $\sigma_1^2 \neq \sigma_2^2$. F-testi olulisuse tõenäosus $p = 0{,}012 < 0{,}05$, võtta vastu H_1 ja järelikult dispersioonid on erinevad. Järgnevalt kahepoolne t-test erinevate dispersioonide korral, H_0: $\mu_1 = \mu_2$ ja H_1: $\mu_1 \neq \mu_2$. Olulisuse tõenäosus kahepoolse hüpoteesi jaoks $0{,}236 > 0{,}05$, võtta vastu H_0. Lennule mitteilmumine ei ole äriklassis ja turistiklassis erinev.

A.7.14 I taseme haridustasemega isikute hulgas (98 isikut) on suhtelisest vaesuspiirist allpool 30 isikut, osakaal $\hat{p}_1 = 0{,}306$. II ja III taseme haridustasemega isikute (402) hulgas on suhtelisest vaesuspiirist allpool 54 isikut, osakaal $\hat{p}_2 = 0{,}134$. H_0: $p_1 \leq p_2$, H_1: $p_1 > p_2$. Statistik $z = 3{,}466$. Kriitiline väärtus nivool 0,01 on 2,33 (ühepoolne hüpotees). Kuna $3{,}466 > 2{,}33$, võtta vastu H_1. I taseme haridusega inimeste hulgas on rohkem suhtelisest vaesusest allpool elavaid isikuid, kui II ja III haridustasemega isikute hulgas.

A.7.15 Ühepoolne hüpotees: H_0: $Me \leq 37$, H_1: $Me > 37$. Neid, kelle vanus on suurem kui mediaan, on 9. Kuna ühegi töötaja vanus ei võrdu mediaaniga, siis valimit korrigeerida pole vaja ning valimi maht on 15. Olulisuse nivoole 10% vastav vasakpoolne kriitiline väärtus on ühepoolse hüpoteesi korral 5. Tegemist on parempoolse hüpoteesiga ja kasutada tuleb parempoolset kriitilist väärtust $15 - 5 = 10$. Kuna $9 < 10$, võtta vastu H_0. Vanuseline diskrimineerimine ei ole tõestatud ja valitud töötajad võib koondada kohtuprotsessi kartmata.

A.7.16 Ühepoolne hüpotees: H_0: $n^+ \leq n^-$, H_1: $n^+ > n^-$. Korrigeeritud valimi maht on 12. Neid elanikke, kelle hinnang 2013. aastal oli parem kui 2011. aastal, on 10. Vasakpoolne kriitiline väärtus olulisuse nivool 0,05 on 3. Kuna tegemist on parempoolse hüpoteesiga, kasutame parempoolset kriitilist väärtust $12-3 = 9$. Võtta vastu H_1, sest $10 > 9$. Olulisuse nivool 0,05 on tõestatud, et hinnang paranes.

A.7.17 a) Märgitest: H_0: $N^+ \leq N^-$ ja H_1: $N^+ > N^-$, kus N^+ on hinnatõusude ja N^- hinnalanguste arv. Valimis on hinnatõusude arv $n^+ = 129$, valimi maht $n = 250$. Vasakpoolne kriitiline väärtus binoomjaotusest $n_{krv} = 112$ ja parempoolne $n_{krp} = 138$. Tegemist on parempoolse hüpoteesiga ja kasutame parempoolset kriitilist väärtust. Kuna $129 < 138$, tuleb vastu võtta nullhüpotees: hinnatõususid ei esine sagedamini kui hinnalangusi. b) Osakaalu testimine: H_0: $p \leq 0{,}5$ ja H_1: $p > 0{,}5$, kus p on hinnatõusude osakaal hinnamuutuste hulgas. Valimis on hinnatõusude osakaal 0,516. Teststatistik $z = 0{,}506$ ja kriitiline väärtus standardiseeritud normaaljaotusest on 1,64. Kuna $0{,}506 < 1{,}64$, tuleb jääda H_0 juurde: hinnatõususid ei esine sagedamini kui hinnalangusi.

A.7.18 H_0: jaotus allub Poissoni jaotusele, H_1: jaotus ei allu Poissoni jaotusele. Vastava Poissoni jaotuse keskväärtus $\lambda = 4{,}98$. Nullhüpotees on, et jaotus allub Poissoni jaotusele, sisukas hüpotees on, et ei allu. Teststatistik $\chi^2 = 21{,}04$ ja kriitiline väärtus $\chi^2_{kr} = 22{,}36$. Kuna $21{,}04 < 22{,}36$, võtta vastu H_0. Valeühenduste arv päevas allub Poissoni jaotusele.

A.7.19 H_0: külastatavus allub ühtlasele jaotusele, H_1: külastatavus ei allu ühtlasele jaotusele. a) Kõigi nädalapäevade lõikes $\chi^2 = 14{,}9$, kriitiline väärtus 12,59. Kuna $14{,}9 > 12{,}59$, on H_0 ümber lükatud ja külastatavus ei ole kõigi nädalapäevade lõikes ühesugune. b) Tööpäevade lõikes $\chi^2 = 9{,}09$, kriitiline väärtus 9,49. Kuna $9{,}09 < 9{,}49$, tuleb jääda H_0 juurde: tööpäevade lõikes on külastatavus ühesugune.

A.7.20 H_0: mitteilmumiste arvu jaotus allub binoomjaotusele, H_1: jaotus ei allu binoomjaotusele. Äriklass: binoomjaotuse parameeter $p = 0{,}05952$, statistik $\chi^2 = 10{,}61$, kriitiline väärtus 16,92. Kuna $10{,}61 < 16{,}92$, võtta vastu H_0. Äriklassis allub mitteilmumiste arv binoomjaotusele. Turistiklass: binoomjaotuse parameeter $p = 0{,}06349$, statistik $\chi^2 = 40{,}07$, kriitiline väärtus 16,92. Kuna $40{,}07 > 16{,}92$, on H_0 ümber lükatud. Turistiklassis ei allu mitteilmumiste arv binoomjaotusele. Põhjuseks võib olla see, et turistiklassis esineb rohkesti mitmekesi reisimist ning binoomjaotuse eeldus pole täidetud.

A.7.21 H_0: aktsia tulumäära jaotus allub normaaljaotusele, H_1: jaotus ei allu normaaljaotusele. Vastava normaaljaotuse parameetrid: $\mu = 0{,}1285\%$, $\sigma = 1{,}464\%$. Teststatistiku leidmiseks tuleb väärtused jagada sagedusklassidesse. Sturgesi valemi järgi sobib klassi laiuseks 1,2%. Kui esimese klassi ülemine piir on $-2{,}56\%$, siis viimase klassi ülemine piir on 4,64% ja klasside arv on 7. Teststatistik $\chi^2 = 1{,}83$, kriitiline väärtus 9,49. Kuna $1{,}83 < 9{,}49$, tuleb vastu võtta H_0. Aktsia tulumäär allub normaaljaotusele.

A.7.22 H_0: vastus küsimusele ei sõltu vastaja majanduslikust olukorrast, H_1: vastus küsimusele sõltub vastaja majanduslikust olukorrast. Olulisuse tõenäosus $p = 0{,}65 > 0{,}05$ ja tuleb vastu võtta H_0: vastus küsimusele ei sõltu vastaja majanduslikust olukorrast. Kui leida χ^2-statistik ja võrrelda seda kriitilise väärtusega, siis $\chi^2 = 1{,}66$ ja olulisuse nivool 0,05 $\chi^2_{kr} = 7{,}81$. Kuna $1{,}66 < 7{,}81$, saame sama tulemuse: võtta vastu H_0.

A.7.23 H_0: kuulumine sissetuleku kvintiili ei sõltu laste arvust, H_1: kuulumine sissetuleku kvintiili sõltub laste arvust. Olulisuse tõenäosus $p = 1{,}3 \cdot 10^{-128}$ näitab, et sisukas hüpotees on tõestatud väga kindlalt. Kuulumine madalamasse või kõrgemasse sissetuleku kvintiili sõltub laste arvust leibkonnas.

A.7.24 H_0: reklaami mõju käibele ei sõltu reklaamikanalist, H_1: reklaami mõju käibele sõltub reklaamikanalist. Olulisuse tõenäosus $0{,}023 < 0{,}05$, võtta vastu H_1: reklaami mõju käibele sõltub reklaamikanalist.

A.7.25 1. H_0: liha valik ei mõjuta kalorisisaldust, $\mu_1 = \mu_2 = \mu_3$. H_1: liha valik mõjutab kalorisisaldust, vähemalt üks μ_i on teistest erinev. Olulisuse tõenäosus $3{,}86 \cdot 10^{-6} < 0{,}05$, võtta vastu H_1. Liha valik mõjutab kalorisisaldust. 2. Tuleb läbi viia kolm t-testi, kus H_0: $\mu_i = \mu_k$, H_0: $\mu_i \neq \mu_k$. Teststatistiku kriitiline väärtus 2,01. Sea- ja loomaliha korral t-testi teststatistik $-0{,}24$. Kuna $|-0{,}24| < 2{,}01$, võtta vastu H_0, oluline erinevus puudub. Sea- ja kanaliha korral t-testi teststatistik $4{,}92 > 2{,}01$, võtta vastu H_1. Kalorisisaldus on oluliselt erinev. Looma- ja kanaliha korral t-testi teststatistik $4{,}96 > 2{,}01$, võtta vastu H_1. Kalorisisaldus on oluliselt erinev. Järelikult, sea- ja loomalihast tehtud hot dog'ide kalorisisaldus ei ole erinev, aga kanalihast tehtud hot dog'ide kalorisisaldus on neist oluliselt erinev.

A.7.26 Kasutada tuleb χ^2-testi, H_0: hilinemine ei sõltu sihtkohast, H_1: hilinemine ja sihtkoht on omavahel seotud. χ^2-testi olulisuse tõenäosus $p = 0{,}347$ on suurem kui olulisuse nivoo 0,05 ja võtta vastu H_0. Kirjade hilinemine ei sõltu sihtkohast.

A.7.27 Sõltuvate valimite t-test. H_0: $D \leq 0$ ja H_1: $D > 0$, kus D on 2000. ja 2001. aasta kasvuprognooside vahe. Olulisuse tõenäosus $1{,}2 \cdot 10^{-9} < 0{,}01$, võtta vastu

H_1. Analüütikute hinnang oli 2000. aasta SKP kasvule optimistlikum kui 2001. aasta kasvule.

A.7.28 Tegemist on sõltumatute valimitega ning t-testiga tuleb võrrelda keskväärtusi. Eelnevalt tuleb läbi viia F-test dispersioonide võrdlemiseks: H_0: $\sigma_1^2 = \sigma_2^2$, H_1: $\sigma_1^2 \neq \sigma_2^2$. F-testi olulisuse tõenäosus on 0,835, järelikult võtta vastu nullhüpotees ja dispersioonid ei ole erinevad. Keskväärtuste võrdlemisel on hüpoteesipaar H_0: $\mu_1 = \mu_2$, H_1: $\mu_1 \neq \mu_2$. t-test võrdsete dispersioonide korral annab kahepoolse hüpoteesi puhul olulisuse tõenäosuseks 0,798. Kuna 0,798 on suurem kui olulisuse nivoo 0,05, võtta vastu H_0. Spordipoes sooritatud ostude arv ei sõltu ilmast.

A.7.29 Kasutada tuleb dispersioonanalüüsi. a) H_0: keskmine sõnade arv reklaamis ei ole ajakirjades erinev, H_1: keskmine sõnade arv reklaamis on ajakirjades erinev. Olulisuse tõenäosus $0{,}00071 < 0{,}05$, võtta vastu H_1, keskmine sõnade arv reklaamis on erinevates ajakirjades erinev. b) H_0: keskmine lausete arv reklaamis ei ole ajakirjades erinev, H_1: keskmine lausete arv reklaamis on ajakirjades erinev. Olulisuse tõenäosus $0{,}14 > 0{,}05$, võtta vastu H_0, keskmine lausete arv reklaamis ei ole ajakirjades erinev. Kuna sõnade arv on erinev ja lausete arv ei ole erinev, siis järelikult keskmine lausepikkus on erinev. Sõnade arvu ANOVA analüüsi kokkuvõtvast tabelist on näha, et kõrgema haridustasemega lugejale mõeldud ajakirjades oli sõnade arv reklaamis keskmiselt kõrgem. Järelikult on nendes ajakirjades avaldatud reklaamides pikemad laused.

A.7.30 Kasutada tuleb χ^2-testi, H_0: reklaamibänneris esinevate sõnade arv ei ole erinevate riikide korral erinev ja H_1: sõnade arv on erinev. Olulisuse tõenäosus $p = 5{,}36 \cdot 10^{-35} < 0{,}05$, sisukas hüpotees on kindlalt tõestatud. Sõnade arv reklaamibänneris on erinevates riikides erinev.

A.7.31 Kasutada tuleks märgitesti, sest tegemist on erinevate toodete hindadega ning seetõttu ei ole sõltuvate valimitega t-testi kasutamine õige: ei tohiks leida keskmist erinevate toodete hinnaerinevustest. Kasutame ühepoolset hüpoteesi. Kui n^+ on nende kaupade arv, mis Tallinnas on odavamad, ja n^- nende kaupade arv, mis Tallinnas on kallimad, siis H_0: $n^+ \leq n^-$ ja H_1: $n^+ > n^-$. Tallinnas on madalam hind 21 kaubal. Korrigeeritud valimi maht on 40, sest kurgi hind on ühesugune. Olulisuse nivool 0,05 on vasakpoolne kriitiline väärtus 15, parempoolne $40 - 15 = 25$. Kuna $21 < 25$, tuleb jääda H_0 juurde. Ei saa väita, et Tallinnas on hinnad keskmiselt madalamad kui Riias. Kui aga kasutada siiski t-testi, tuleb samuti võtta vastu H_0: olulisuse tõenäosus $0{,}332 > 0{,}05$.

A.7.32 Kasutada tuleb ühepoolset F-testi, H_0: $\sigma_{DV}^2 \leq \sigma_{KV}^2$ ja H_1: $\sigma_{DV}^2 > \sigma_{KV}^2$. Aastatel 2004–2007 oli fondi LHV DV tootluse dispersioon suurem kui fondil LHV KV. Aastal 2008 aga tuleb olulisuse nivool 0,01 võtta vastu nullhüpotees, LHV DV dispersioon ei olnud suurem.

Aasta	2004	2005	2006	2007	2008
F-empiiriline	4,265	1,884	1,809	7,445	1,225
F-kriitiline	1,338	1,339	1,340	1,340	1,341
Kumb hüpotees vastu võtta	H_1	H_1	H_1	H_1	H_0

A.7.33 Mõlemal juhul tuleb kasutada sõltuvate valimite t-testi keskväärtuste võrdlemiseks. 1. H_0: $D \leq 0$ ja H_1: $D > 0$, kus D on 2007. ja 2008. aasta prognooside erinevuste keskmine. t-statistik $-3{,}157$ ja kriitiline väärtus $-1{,}76$. Teststatistik langeb kriitilisse piirkonda: $-3{,}157 < -1{,}76$ (olulisuse tõenäosus $0{,}0035 < 0{,}05$), võtta vastu H_1. Olulisuse nivool 0,05 on tõestatud, et 2008. aastaks planeeritud kulud olid suuremad kui 2007. aastaks planeeritud kulud. 2. H_0: $D \geq 0$ ja H_1: $D < 0$, kus D on 2008. ja 2009. aasta prognooside erinevuste keskmine. t-statistik 1,539 ja vastav kriitiline väärtus 1,76. Teststatistik ei lange kriitilisse piirkonda, $1{,}539 < 1{,}761$ (olulisuse tõenäosus $0{,}0731 > 0{,}05$), võtta vastu H_0. Olulisuse nivool 0,05 ei ole tõestatud, et 2009. aastaks planeeritud kulud olid väiksemad kui 2008. aastaks planeeritud kulud.

A.7.34 Kasutada tuleb osakaalude testimist. H_0: $p_1 = p_2$, H_1: $p_1 \neq p_2$. Teststatistik $z = 0{,}787$. Kriitilised väärtused olulisuse nivoo 0,05 korral on $-1{,}96$ ja 1,96. Kuna $-1{,}96 < 0{,}787 < 1{,}96$, võtta vastu H_0. Säästusid omavate leibkondade osakaal ei olnud muutunud.

A.7.35 Kasutada tuleb dispersioonanalüüsi, H_0: variandid olid ühesuguse raskusastmega ja H_1: variandid ei olnud ühesuguse raskusastmega. Olulisuse tõenäosus $0{,}524 > 0{,}05$, võtta vastu H_0. Variandid olid ühesuguse raskusastmega.

A.7.36 Kasutada tuleb sõltumatute valimite t-testi keskväärtuste testimiseks. Eelnevalt on vaja F-testi abil kindlaks teha, kas dispersioonid on võrdsed või mitte. F-testi olulisuse tõenäosus $p = 0{,}292 > 0{,}05$ ja võtta vastu H_0: dispersioonid on võrdsed. Olgu kogum 1 need poed, mille logo esineb vastavates portaalides. t-testi korral nullhüpotees $\mu_1 \leq \mu_2$ ja sisukas hüpotees $\mu_1 > \mu_2$. Testi olulisuse tõenäosus $p = 0{,}019 < 0{,}05$, võtta vastu H_1. Olulisuse nivool 0,05 on tõestatud, et poodides, mille logo esineb portaalides shopper.com või pricegrabber.com, on vastava mälu hind kõrgem.

A.7.37 Kasutada tuleb χ^2-testi. H_0: külastajate arv allub normaaljaotusele, H_1: külastajate arv ei allu normaaljaotusele. Testimiseks on vaja variatsioonrida intervallida. Sturgesi valemi kasutamisel saadakse klasside arvuks 7,5 ning klassi laiuse ligikaudseks väärtuseks 353. Kui klassi laiuseks võtta 350 ning esimese klassi ülemine piir võtta 5500, siis saadakse allpool olev tabel.

Ülemine piir u	n_{emp}	F	P	n_o	$\frac{(n_{emp}-n_o)^2}{n_o}$
5500	3	0,083	0,083	7,53	2,72
5850	18	0,206	0,123	11,22	4,09
6200	25	0,400	0,194	17,65	3,06
6550	12	0,623	0,223	20,30	3,39
6900	14	0,811	0,188	17,07	0,55
7250	12	0,926	0,115	10,50	0,21
7600	4	0,978	0,052	4,73	0,11
7950	3	0,995	0,022	2,00	0,50

F on vastava normaaljaotuse jaotusfunktsiooni väärtus ülemise piiri u jaoks, P normaaljaotusest leitud klassi langemise tõenäosus ning n_o oodatav sagedus. Teststatistik χ^2 tuleb 14,64. Kriitiline väärtus olulisuse nivool 0,05 on 11,07. Kuna

14,64 > 11,07, on H_0 ümber lükatud. Poe külastajate arv päevas ei allu normaaljaotusele.

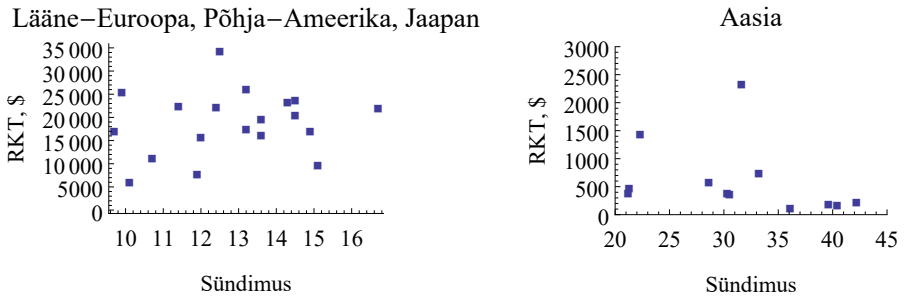
A.7.38 Kasutada tuleb märgitesti, ühepoolne hüpotees. Olgu n^+ nende analüütikute arv, kes arvavad, et Jaapani Keskpanga vastavat otsust on oodata 2015. aasta teises pooles või hiljem, ja n^- nende arv, kes pakuvad varasemat aega. H_0: $n^+ \leq n^-$ ja H_1: $n^+ > n^-$. 2015. aasta teist poolt või hilisemat aega pakkus 18 analüütikut, varasemat aega 10. Korrigeeritud valimi maht on 28. Vasakpoolne kriitiline väärtus olulisuse nivool 0,05 on 10, parempoolne kriitiline väärtus $28 - 10 = 18$. Kuna arv 18 ei ületa kriitilist väärtust, tuleb jääda H_0 juurde: ei saa väita, et enamik majandusanalüütikuid prognoosivad seda 2015. aasta teises pooles või hiljem.

A.7.39 Kasutada tuleb dispersioonide võrdlemist F-testiga, ühepoolne hüpotees. H_0: $\sigma^2_{L635} \leq \sigma^2_{L621}$, H_1: $\sigma^2_{L635} > \sigma^2_{L621}$. Olulisuse tõenäosus $p = 0{,}056 > 0{,}05$, järelikult tuleb jääda H_0 juurde: liinilt L635 tuleva toodangu massi varieeruvus ei ole suurem.

8. ptk. Korrelatsioonanalüüs

8.1 4. **8.2** 7,49%. **8.3** $w_A = \frac{\sigma_B^2 - cov_{AB}}{\sigma_A^2 + \sigma_B^2 - 2cov_{AB}}$. **8.4** Osakaal 84,1%, portfelli standardhälve 6,6%. **8.5** Kõige tugevam on seos d) ja kõige nõrgem seos b). **8.6** 0,009. **8.7** 1. Nivool 0,05 on oluline varad, nivool 0,01 on oluline juhatuse liikmete maksehäired, nivool 0,001 on olulised logaritmitud varad ja ettevõtte vanus aastates. Logaritmitud varade muutus oluline ei ole. 2. Jah, sest ettevõtte vanuse ja juhatuse suuruse vahel on statistiliselt oluline positiivne korrelatsioon. 3. Ei, sest maksehäirete ja juhatuse suuruse vahel on statistiliselt oluline negatiivne korrelatsioon: suurema juhatuse korral on juhatuse liikmetel maksehäireid vähem. **8.8** 0,857.

A.8.1 1. RKT ja SÜND −0,629; SURM −0,303; IM.SURM −0,602; IGA M 0,643; IGA N 0,650. 2. Positiivselt meeste ja naiste keskmine eluiga, negatiivselt ülejäänud. 3. Sünnid kõige tugevamini, surmade arv kõige nõrgemini. 4. 3. grupis 0,184 ja 5. grupis −0,701. Arenenud riikides (Lääne-Euroopa, Põhja-Ameerika, Jaapan) on RKT ja sündimuse vahel nõrk positiivne korrelatsioon, Aasia riikides negatiivne korrelatsioon. 5. Vt joonis ÜV.9.



Joonis ÜV.9. Ülesande A.8.1 5. osa vastus

A.8.2 1. Korrelatsioonimaatriks on allpool. 2. S&P 500 ja DJIA. 3. S&P 500 ja FTSE. 4. FTSE 100 ja NASDAQ 100 ning FTSE 100 ja Nikkei 225. 5. FTSE 100. 6. DJIA.

	S&P 500	DJIA	Nasdaq 100	Nikkei 225	FTSE 100	OMX
S&P 500	1					
DJIA	0,971	1				
Nasdaq 100	0,951	0,903	1			
FTSE 100	0,032	0,064	−0,143	1		
OMXS	0,861	0,899	0,763	0,263	1	
Nikkei 225	0,784	0,838	0,846	−0,143	0,716	1

A.8.3 Korrelatsioonimaatriks koos Harjumaaga:

	Töötuse määr	Keskmine bruto-kuupalk	Sündinud ette-võtted	Sekundaar-sektor	Tertsiaar-sektor
Töötuse määr	1				
Keskmine brutokuupalk	−0,389	1			
Sündinud ettevõtted	−0,170	0,911	1		
Sekundaarsektor	0,208	−0,363	−0,362	1	
Tertsiaarsektor	−0,112	0,613	0,619	−0,880	1

A.8.4 Arco Vara −0,237, Harju Elekter −0,382, Olympic Entertainment Group −0,0757. Kõige juhuslikum on Olympic Entertainment Groupi aktsia tulumäära muutumine.

A.8.5 Korrelatsioonikordaja 0,857; parameetri empiiriline väärtus 3,32. Olulisuse nivool 5% on kriitiline väärtus 2,78. Kuna $3{,}32 > 2{,}78$, võtta vastu H_1, seos on oluline. Olulisuse nivool 1% on kriitiline väärtus 4,6. Kuna $3{,}32 < 4{,}6$, võtta vastu H_0, seos ei ole oluline.

A.8.6 Korrelatsioonikordaja väärtus 0,828, statistiku empiiriline väärtus 2,56. Kasutada tuleb ühepoolset hüpoteesi, H_0: $r \leq 0$ ja H_1: $r > 0$. Kriitiline väärtus olulisuse nivool 5% on 2,35. Kuna $2{,}56 > 2{,}35$, on H_0 ümber lükatud. Eksperiment kinnitab hüpoteesi, et mitme tüki hinna reklaamimisel müügimaht suureneb.

A.8.7 t-jaotuse täiendkvantiil $t_{0{,}025}(37) = 2{,}03$ ja korrelatsioonikordaja kriitiline väärtus 0,316. Olulised seosed on müügitulu ja puhaskasumi vahel ($r = 0{,}962$), müügitulu ja investeeringute vahel ($r = 0{,}903$), puhaskasumi ja investeeringute vahel ($r = 0{,}909$) ning tööjõukulude ja tootlikkuse vahel ($r = 0{,}470$).

A.8.8 Eesti ja Läti vahel 0,4 ning Eesti ja Soome vahel 0. Töötajate vajaduste järjestus langeb rohkem kokku Eesti ja Läti väikeettevõtetes kui Eesti ja Soome omades.

A.8.9 1. $r = 0{,}712$; seos on oluline, vastav kriitiline korrelatsioonikordaja nivool 5% on 0,514. 2. $r_s = 0{,}284$. 3. Hajumisdiagrammilt selgub, et on üks erind: Harjumaa. 4. 0,279, seos ei ole oluline, vastav kriitiline korrelatsioonikordaja nivool 5% on 0,532.

9. ptk. Regressioonanalüüs

9.1 a) Maja hind Y ja maja pindala X; b) reklaamiminuti hind Y ja telesaadete vaadatavus X; c) autode arv 1000 elaniku kohta Y ja SKP elaniku kohta X; d) SKP Y ja tööealise elanikkonna suurus X.

9.2 110,4 cm. Iga aastaga lisandub keskmiselt 6 sentimeetrit.

9.3 Ajakirjal, mille tellijaskonna suurus on tuhande võrra suurem, on reklaami hind 5,28 dollarit suurem. Ajakirjal, mille tellijaskonna hulgas on naiste osakaal 1% võrra suurem, on reklaami hind 11 dollarit väiksem. Ajakirjal, mille lugejate mediaansissetulek on ühe dollari võrra suurem, on reklaami hind 1,22 dollarit suurem.

9.4 Kui müügimaht aastas kasvab tuhat tonni, siis müügitulu aastas suureneb 6,7 tuhande dollari võrra. Kui kulud reklaamile suurenevad tuhat dollarit aastas, siis müügitulu kasvab 2,7 tuhande dollari võrra aastas. Kui teiste maiustuste reklaamikulud suurenevad tuhat dollarit aastas, siis köhakommide müügitulu väheneb 0,01 tuhande dollari võrra. Müügitulu aastas 218,5 tuhat dollarit.

9.5 Vanuse suurenedes ühe aasta võrra kasvab tööga rahulolu a_1 palli. Kui hariduse omandamise aeg on ühe aasta võrra pikem, on tööga rahulolu a_2 palli võrra suurem. Kui töökogemus on ühe kuu võrra pikem, on tööga rahulolu a_3 palli võrra suurem. Kui müügikogemus on ühe kuu võrra pikem, on tööga rahulolu a_4 palli võrra suurem.

9.6 1. 440 grammi. 2. Kuna tuletis $C'(30) = 0{,}628$ on positiivne, siis 30-aastaselt vanuse kasvades kohvi tarbimine suureneb. 3. Kuna tuletis $C'(65) = -0{,}142$ on negatiivne, siis 65-aastaselt vanuse kasvades kohvi tarbimine väheneb. 4. 28–89-aastased. 5. 58,5-aastased.

9.7 1. Töötuse määra tõus ühe protsendipunkti võrra vähendab majanduskasvu 0,325 protsendipunkti võrra. 2. Ekspordi osatähtsuse suurenemine ühe protsendipunkti võrra suurendab majanduskasvu 0,0939 protsendipunkti võrra. 3. Kuna Gini indeksi ruutliikme kordaja on negatiivne, on tegemist allapoole avatud parabooliga. Suure ebavõrdsuse korral on mõju majanduskasvule negatiivne: Gini indeksi suurenedes majanduskasv väheneb. 4. Tuleb leida Gini indeksi mõju kirjeldava parabooli maksimumkoht. Tuletis Gini indeksi järgi $\mathrm{SKP}' = 2{,}45 - 2 \cdot 0{,}0386\,\mathrm{GIN}$. Tuletise nullkoht on 31,7. Kui Gini indeks on sellest väiksem, siis Gini indeksi suurenemine suurendab majanduskasvu.

9.8 Parim on teine mudel, kus on kaks argumenttunnust. Otsustamiseks tuleb leida korrigeeritud determinatsioonikordajad R_a^2, mille väärtus esimese mudeli korral on 0,419, teise mudeli korral 0,475 ja kolmanda mudeli korral 0,466.

9.9 Mudelis oli neli argumenttunnust.

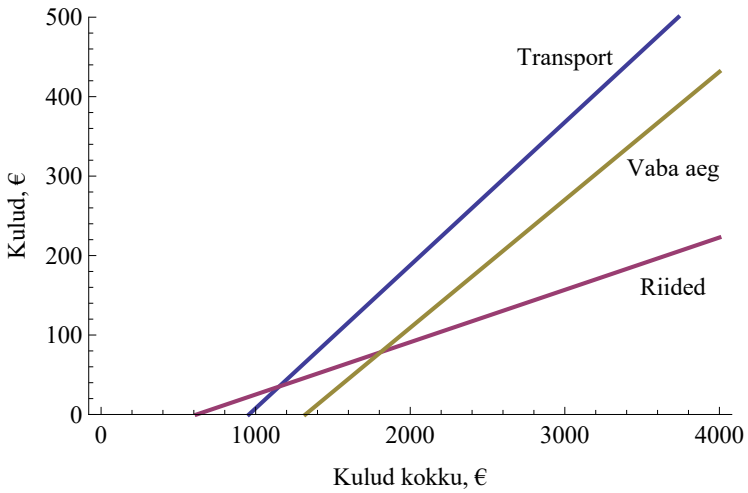
9.11 1. Mudel 1: t-statistiku kriitiline väärtus on 2,06, mis leitakse tabelarvutuses funktsiooniga T.INV(0,05; 25). Tunnuse x_1 korral $t = 2{,}58 > 2{,}06$, võtta vastu H_1. Mudel 2: t-kriitiline on 2,074. t-statistikud: x_1 korral 3,32, x_2 korral 6,47, x_3 korral $-4{,}77$ ja x_4 korral 2,43. Kõikide tunnuste korral $|t| > 2{,}074$, võtta vastu H_1. Mudel 3: t-kriitiline on 2,08. t-statistikud: x_1 korral 3,21, x_2 korral 5,24, x_3 korral $-4{,}00$, x_4 korral 2,28 ja x_5 korral 0,14. Tunnuse x_5 korral $|t| < 2{,}08$, võtta vastu H_0. Ülejäänud tunnuste korral võtta vastu H_1. 2. Kuna keskmine temperatuur kasvu

ajal (x_2), sademed augustis ja septembris (x_3), sademed oktoobrist märtsini (x_4) on statistiliselt olulised, siis on tõestatud, et klimaatilised tingimused enne viinamarjade kasvu ja kasvamise ajal mõjutavad veini hinda. 3. Kuna keskmine temperatuur septembris (x_5) ei ole mudelis 3 statistiliselt oluline, siis pole tõestatud, et see mõjutab veini hinda. 4. Kuna mudelites on argumenttunnuste arv erinev, tuleb mudelite võrdlemiseks kasutada korrigeeritud determinatsioonikordajat. Artikli autorid ei ole seda tabelisse lisanud, järelikult tuleb see arvutada. Mudeli 1 korrigeeritud determinatsioonikordaja on 0,180, mudeli 2 korral 0,797 ja mudeli 3 korral 0,787. Järelikult, kõige parem on mudel 2.

9.12 1. $\hat{\tilde{y}} = 0{,}0141 + 0{,}111\tilde{x}_1 - 0{,}146\tilde{x}_2 + 1{,}655\tilde{x}_3$. 2. Kui ühe kilogrammi juustu hind ELi siseturul kasvab ühe euro võrra, siis piima kokkuostuhind kasvab 0,111 euro võrra. Kui ühe kilogrammi või hind ELi siseturul kasvab ühe euro võrra, siis piima kokkuostuhind kahaneb 0,146 euro võrra. Kui ühe kilogrammi odra hind Eesti siseturul kasvab ühe euro võrra, siis piima kokkuostuhind kasvab 1,655 euro võrra.

A.9.1 Mudel $\hat{y} = 0{,}02534x + 502{,}1$, kus x on SKP elaniku kohta (eurot) ja y keskmine brutokuupalk eurodes. Brutopalga varieeruvusest on 81,7% põhjustatud elaniku kohta tuleva SKP erinevustest.

A.9.2 1. Riided: $\hat{y}_{riided} = -40{,}7 + 0{,}0659x$, transport: $\hat{y}_{transp} = -172 + 0{,}18x$; vaba aeg: $\hat{y}_{v.aeg} = -213 + 0{,}161x$ kus x on tarbimiskulud kokku. 2. Tarbimiskulude suurenedes 1 euro võrra suurenevad kulud riietele ja jalanõudele 0,0659 eurot, kulud transpordile 0,18 eurot ja kulud vabale ajale 0,161 eurot. Kõige kiiremini kasvavad kulud transpordile ning kõige aeglasemalt kulud riietele. 3. Kõige suurema kirjeldusvõimega on vaba aja kulutuste tarbimismudel, $R^2 = 0{,}544$. 4. Joonis ÜV.10. 5. Kulud riietele tekivad, kui kogukulud pereliikme kohta on 618 eurot aastas. Kulud transpordile tekivad, kui kogukulud pereliikme kohta on 957 eurot aastas. Kulud vabale ajale tekivad, kui kogukulud pereliikme kohta on 1322 eurot aastas.



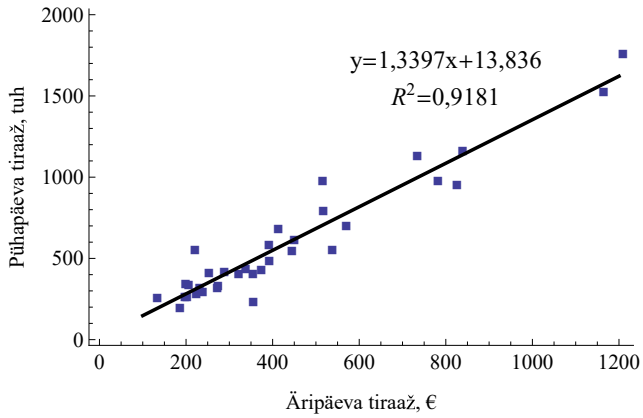
Joonis ÜV.10. Ülesande A.9.2 4. osa diagramm

A.9.3 1. Mazda hinna sõltuvust auto vanusest kirjeldab mudel $\hat{y} = -1848x + 20001$, kus x on auto vanus aastates ja y hind eurodes. 2. Auto, mille vanus on ühe aasta võrra suurem, maksab 1848 eurot vähem. 3. Uus Mazda maksaks *ca* 20 tuhat eurot. 4. Mazda, mille vanus on 6 aastat, maksaks 8911 eurot.

A.9.4 a) $\hat{C} = 5193 + 0{,}0205q$, kus C on kulud miljonites eurodes ning q toodetud autode arv aastas; b) püsikulu usaldusvahemik on (797, 9589) miljonit eurot; c) piirkulu usaldusvahemik (0,017, 0,024) miljonit eurot auto kohta ehk 17 kuni 24 tuhat eurot auto kohta; d) piirkulu langeb kokku, püsikulu punkthinnang on BMW-l suurem, kuid usaldusvahemikud kattuvad.

A.9.5 1. $\widehat{\Delta\pi}_t = 9{,}08 - 0{,}469U_t$, kus $\Delta\pi_t$ on THI muutus ning U_t on töötuse määr protsentides. 2. 19,4%.

A.9.6 1. Hajumisdiagramm koos lisatud regressioonjoonega on joonisel ÜV.11. 2. Mudel on $\hat{y} = 13{,}8 + 1{,}34x$, kus y on pühapäevane ning x argipäevane tiraaž, mõlemad tuhandetes. 3. Vabaliige $-59{,}1$ kuni 86,8; lineaarliikme kordaja 1,20 kuni 1,48. 4. 91,8%. 5. 684 ± 39 tuhat. 6. 684 ± 226 tuhat. Individuaalväärtuse prognoosi usaldusvahemiku poollaius on *ca* 5,7 korda suurem kui keskväärtuse prognoosil. 7. 2693 ± 320. Usaldusvahemiku poollaius on suurem, kuna tiraaž 2 miljonit asub keskmisest $\bar{x} = 431$ tuhat oluliselt kaugemal kui eelmises osas väärtus 500 tuhat.

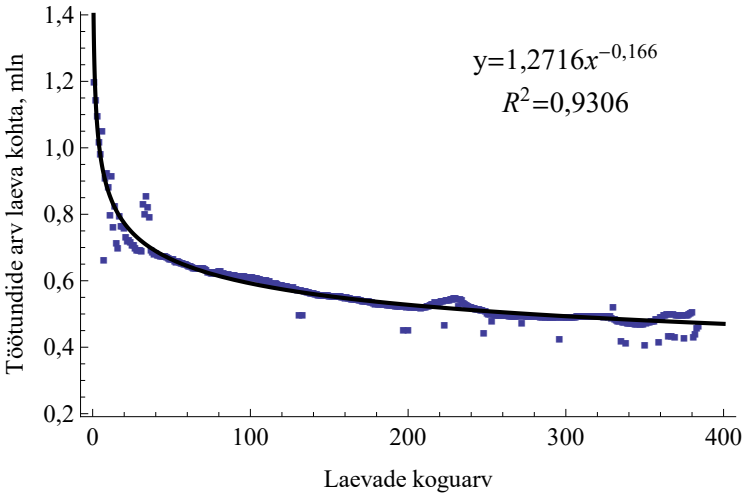


Joonis ÜV.11. Ülesande A.9.6 vastuse juurde

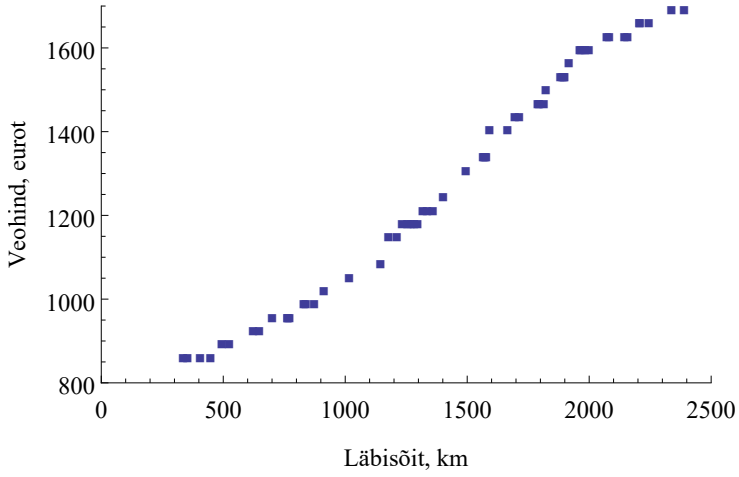
A.9.7 1. ja 2. osa vt joonis ÜV.12 3. $r = 11\%$.

A.9.8 1. Vt joonis ÜV.13. 2. Parim mudel on ruutpolünoom $\hat{y} = 0{,}0002x^2 + 0{,}0385x + 822{,}05$, kus y on veohind eurodes ning x läbisõit kilomeetrites. Selle mudeli determinatsioonikordaja on kõige suurem, $R^2 = 0{,}9942$. Vt ka joonis ÜV.14. 3. Parim mudel on ruutpolünoom $\hat{y} = -0{,}0004x^2 + 1{,}8327x - 657{,}81$, mille determinatsioonikordaja $R^2 = 0{,}9825$ on kõige suurem. 4. a) 0,585 €/km; b) 0,945 €/km; c) 0,153 €/km.

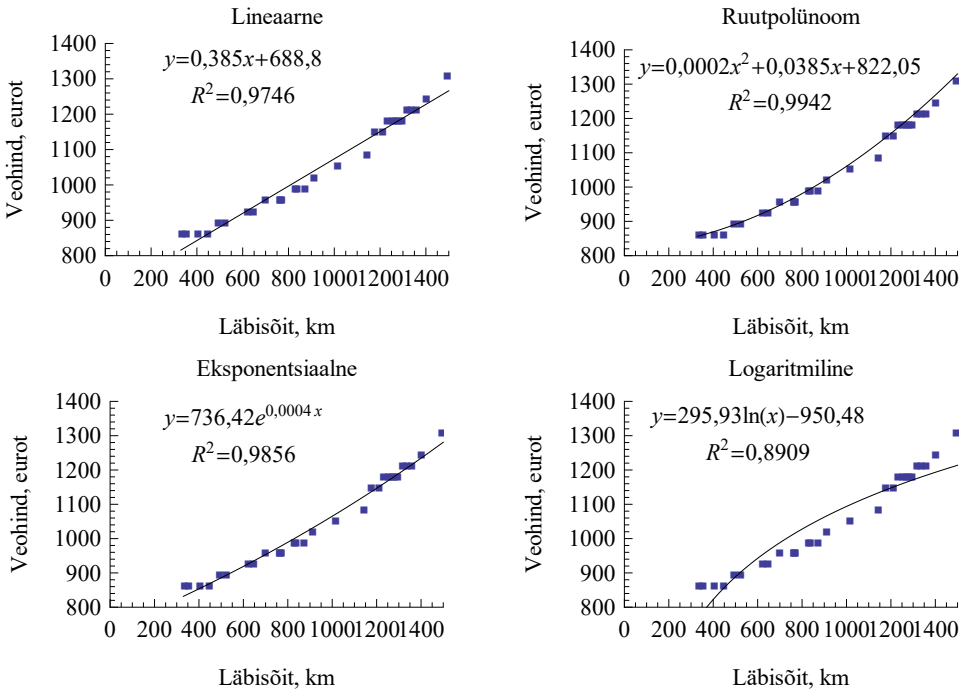
A.9.9 1. ja 2. osa vt joonis ÜV.15. Mudeliks on $\hat{y} = -0{,}0069x^2 + 0{,}1757x + 0{,}0812$, kus y on kasvumäär ning x veetemperatuur °C. 3. Temperatuur 12,7 °C, kasvumäär on siis 1,20%.



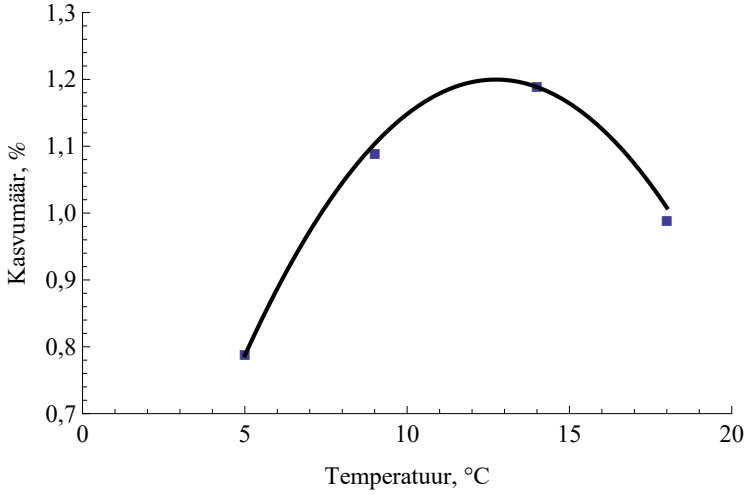
Joonis ÜV.12. Ülesande A.9.7 vastuse juurde



Joonis ÜV.13. Ülesande A.9.8 1. osa vastus



Joonis ÜV.14. Ülesande A.9.8 2. osa vastuse juurde. Joonisel on neli erineva kujuga regressioonjoont, kõige parem on ruutpolünoom, mille R^2 on kõige suurem



Joonis ÜV.15. Ülesande A.9.9 vastuse juurde

A.9.10 1. Vt tabelit. Käibe suurenemine sõltub nii sellest, mitmes linnas reklaami näidati kui ka sellest, kui tihti näidati, sest mõlemad tunnused on statistiliselt olulised. 2. 0,00077. 3. 0,0013.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	1,125	20,78	0,0541	0,959
Linnade arv	5,28125	0,727	7,261	0,00077
Reklaamide sagedus nädalas	9,4375	1,455	6,487	0,00130

A.9.11 Ei paranda. Läbisõit on statistiliselt mitteoluline ($p = 0{,}353$) ning korrigeeritud determinatsioonikordaja väheneb ($R_a^2 = 0{,}9213 \to R_a^2 = 0{,}9209$).

A.9.12 Excelis tehtud regressioonanalüüs annab järgmise tulemuse:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,945
R Square	0,892
Adjusted R Square	0,885
Standard Error	133,5
Observations	32

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	4283063,0	2141531	120,2	9,22E-15
Residual	29	516726,5	17818,16		
Total	31	4799789,5			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	−1339	173,8	−7,70	1,71E-08	−1694	−983
x1	12,74	0,905	14,08	1,69E-14	10,89	14,59
x2	85,95	8,73	9,85	9,34E-11	68,10	103,80

Mudel on statistiliselt oluline, F-testi olulisuse tõenäosus ANOVA tabelis (*Significance F*) on $9{,}22 \cdot 10^{-15} < 0{,}05$.

Parameetrite statistilist olulisust näitavaid olulisuse tõenäosusi (veerg *P-value*) võrdleme olulisuse nivooga 0,05:

vanus x_1: $p = 1{,}69 \cdot 10^{-14} < 0{,}05$, võtta vastu H_1;
osalejate arv x_2: $p = 9{,}34 \cdot 10^{-11} < 0{,}05$, võtta vastu H_1.

Mõlema tunnuse kordajad on statistiliselt olulised, järelikult kollektsionääri hüpoteesid on tõestatud olulisuse nivool 0,05. Kella hinda y võib modelleerida mudeliga

$$\hat{y} = -1339 + 12{,}741x_1 + 85{,}95x_2,$$

kus x_1 on kella vanus aastates ja x_2 oksjonil osalejate arv. See mudel seletab ära 89,2% kella hinna varieerumisest (determinatsioonikordaja $R^2 = 0{,}892$). Mudeli põhjal võib väita, et oksjonil osalejate arvu suurenemine ühe võrra suurendab antiikse kella oksjonihinda 85,95 dollari võrra.

A.9.13 1. Mudel on

$$\widehat{\text{MÜÜK}} = -2{,}79 + 0{,}0438\,\text{HIND}.$$

Hinnaindeksi tõustes ühe palli võrra suureneb autode müük 0,0438 mln auto võrra aastas. See ei ole loogiline. 2. Mudel on

$$\widehat{\text{MÜÜK}} = 2{,}081 - 0{,}0282\,\text{HIND} + 0{,}0787\ \text{TULU} - 0{,}2097\ \text{ARV}.$$

Nüüd mõjutab hinnatõus autode müüki negatiivselt, hinnaindeksi ühepalline tõus vähendab müüki 0,0282 miljoni auto võrra aastas. Elanike kogusissetuleku suurenemine aastas 1 miljardi dollari võrra suurendab autode müüki 0,0787 miljoni võrra. Autode koguarvu tõustes ühe miljoni võrra väheneb autode müük aastas 0,2097 miljoni võrra.

A.9.14 1. Mudel on

$$\widehat{\text{TM}} = 893{,}5 + 10{,}76\,\text{RM} - 7{,}92\,\text{RT} - 6{,}12\,\text{ÜM}.$$

2. Nivool 0,01 on statistiliselt olulised rahvaarvu muutus ($p = 2{,}11 \cdot 10^{-6}$) ja ülalpeetavate määr ($p = 8{,}88 \cdot 10^{-13}$), nivool 0,05 on statistiliselt oluline registreeritud töötus ($p = 0{,}0169$). 3. Determinatsioonikordaja $R^2 = 0{,}344$. 4. Kui rahvaarvu suurenemine on 1%, siis füüsilise isiku tulumaks ühe elaniku kohta suureneb 10,76 eurot aastas. Kui registreeritud töötus suureneb 1%, siis füüsilise isiku tulumaks ühe elaniku kohta väheneb 7,92 eurot aastas. Kui ülalpeetavate määr suureneb 1%, siis füüsilise isiku tulumaks ühe elaniku kohta väheneb 6,12 eurot aastas. 5. Ebaharilikud omavalitsused ja erindid on tabelis.

Omavalitsus	Standardiseeritud jääk	
Aegviidu vald	2,31	
Emmaste vald	2,14	
Harku vald	2,15	
Kasepää vald	−2,30	
Kihnu vald	2,58	
Piirissaare vald	−2,56	
Salme vald	2,90	
Saue linn	3,07	erind
Viimsi vald	2,37	
Vormsi vald	4,37	erind

A.9.15 1. Seletavate tunnuste järjestus tähtsuse järjekorras, sulgudes on korrelatsioonikordaja sõltuva tunnusega HIND: VANNIT ($r = 0{,}725$), KRUNT ($r = 0{,}607$),

GARAAZ ($r = 0{,}487$), VANUS ($r = -0{,}409$), KAMIN ($r = 0{,}316$). Kuna viie tunnusega mudelis kaminate arv ei ole statistiliselt oluline ($p = 0{,}324 > 0{,}05$), on lõplikus mudelis neli tunnust:

$$\widehat{\text{HIND}} = 14{,}6 + 12{,}6\,\text{VANNIT} + 0{,}905\,\text{KRUNT} + 2{,}85\,\text{GARAAZ} - 0{,}102\,\text{VANUS}.$$

2. Maja, milles vannitubade arv on ühe võrra suurem, maksab 12,6 tuhat dollarit rohkem. Tuhande ruutjala võrra suurem krunt tõstab maja hinda 0,905 tuhande dollari võrra. Maja, mille juures on garaažikohtade arv ühe võrra suurem, maksab 2,85 tuhat dollarit rohkem. Maja, mille vanus on ühe aasta võrra suurem, maksab 0,102 tuhat dollarit vähem. 3. Hinna sõltuvus kinnisvaramaksust: $\widehat{\text{HIND}} = 17{,}83 + 2{,}573\,\text{MAKS}$. Selle mudeli korrigeeritud determinatsioonikordaja $R_a^2 = 0{,}652$. Punktis 1 leitud mudeli korral $R_a^2 = 0{,}739$, mis on suurem. Järelikult mudel, kus seletavateks tunnusteks on kinnisvara iseloomustavad suurused, on müügihinna prognoosimiseks parem.

A.9.16 1. Majanduskasvu mudel on

$$\widehat{\text{Growth}} = 0{,}1134 - 0{,}01363\,\text{GDP60} + 0{,}0458\,\text{Mining} - 0{,}0340\,\text{PrimExp70} + \\ + 0{,}000735\,\text{Invest} + 0{,}04434\,\text{YrsOpen} - 0{,}01232\,\text{RevCoup}.$$

2. Majanduskasvule mõjuvad positiivselt kaevandamise osatähtsus SKP-s (Mining), kodumaiste investeeringute suhe SKP-sse (Invest) ning avatud majanduse kestvus (YrsOpen). Negatiivselt mõjub SKP elaniku kohta (GDP60), primaarkaupade ekspordi suhe SKP-sse (PrimExp70) ning revolutsioonide ja riigipöörete arv (RevCoup). SKP elaniku kohta negatiivne mõju on seletatav sellega, et madala elatustasemega riikides on kasv kiirem, elatustaseme kasvades kasv aeglustub.

A.9.17 $\hat{y} = -0{,}616 + 0{,}204x_1 + 0{,}00580x_2$, kus y on piima tarbimine elaniku kohta aastas (gallonit), x_1 kulud reklaamile (senti elaniku kohta aastas) ja x_2 sissetulek elaniku kohta aastas (dollarit). Selles mudelis on kõik tunnused statistiliselt olulised nivool 0,05 ning korrigeeritud determinatsioonikordaja kõige suurem ($R_a^2 = 0{,}816$). Mudel seletab ära 86,2% piima tarbimise varieerumisest ($R^2 = 0{,}862$).

A.9.18 Mudelite determinatsioonikordajad ja korrigeeritud determinatsioonikordajad: mudel (9.126) $R^2 = 0{,}816$, $R_a^2 = 0{,}795$; mudel (9.127) $R^2 = 0{,}935$, $R_a^2 = 0{,}928$; mudel (9.128) $R^2 = 0{,}952$, $R_a^2 = 0{,}940$. Kuigi kahe seletava tunnusega mudeli (9.128) korrigeeritud determinatsioonikordaja R_a^2 on suurem, kui ühe seletava tunnusega mudelil (9.127), pole SKP selles mudelis statistiliselt oluline ($p = 0{,}13 > 0{,}05$). Lisaks on SKP kordaja märk negatiivne, mis pole loogiline. Põhjuseks on multikollineaarsus: SKP ja põhivarade vaheline korrelatsioonikordaja 0,968 on suurem kui kaubaveo ja SKP (0,903) ning kaubaveo ja põhivara investeeringute vahel (0,967). Sobiv mudel on see, kus seletavaks tunnuseks on ainult põhivara investeeringud: $\hat{y} = 363{,}7 + 0{,}8731x_2$.

A.9.19 1. Tunnuste järjestus korrelatsioonikordajate järgi, sulgudes on korrelatsioonikordaja sõltuva tunnusega KWH: PINDALA ($r = 0{,}879$), TOAD ($r = 0{,}685$), VANUS ($r = -0{,}099$), TÄITUVUS ($r = 0{,}031$). Mudel on

$$\widehat{\text{KWH}} = -5473 + 0{,}168\,\text{PINDALA} + 73{,}7\,\text{TÄITUVUS}.$$

2. Esineb multikollineaarsus. Tunnuste PINDALA ja TOAD omavaheline korrelatsioonikordaja 0,853 on suurem kui TOAD ja KWH vaheline korrelatsioonikordaja 0,685. 3. Hotellis, mille pindala on ühe ruutmeetri võrra suurem, on energiakulu 0,168 tuhande kWh võrra suurem. Hotellis, mille keskmine täituvus on 1% suurem, on energiakulu 73,7 tuhat kWh suurem. 4. 82,1%.

A.9.20 1. Mudel on $\hat{y} = 0{,}628x + 0{,}116$, kus y on toiduainete müügikäive aastas (mld \$) ja x rahvaarv miljonites. Vabaliige on statistiliselt mitteoluline, $p = 0{,}252$. 2. $\hat{y} = 0{,}652x$. Ühe miljoni elaniku kohta on toidukaupade käive keskmiselt 652 miljonit dollarit aastas.

A.9.21 1. $\widehat{\mathrm{CAL}} = -1{,}482 + 8{,}842\,\mathrm{RASV} + 4{,}070\,\mathrm{VALK} + 3{,}977\,\mathrm{SV}$. Vabaliige on statistiliselt mitteoluline, olulisuse tõenäosus $0{,}183 > 0{,}05$. 2. Vabaliikmeta mudel: $\widehat{\mathrm{CAL}} = 8{,}826\,\mathrm{RASV} + 4{,}023\,\mathrm{VALK} + 3{,}955\,\mathrm{SV}$. 3. Süsivesikute SV kordaja 3,978 Johnsoni mudelis langeb hinnatud mudeli kordaja usaldusvahemikku $(3{,}912, 3{,}998)$. Ülejäänud kahe tunnuse kordajad ei lange. 4. Kõige rohkem rasvad, kõige vähem süsivesikud. Arvestades aga valkude ja süsivesikute kordajate usalduspiire, võivad need anda ühepalju energiat, sest usaldusvahemikud osaliselt kattuvad. 5. 244 kcal.

A.9.22 Riided: $\hat{y}_{riided} = -111 + 0{,}0952x - 6{,}66 \cdot 10^{-5}x^2$. Ruutliikme olulisuse tõenäosus $p = 2{,}69 \cdot 10^{-7} < 0{,}05$, on statistiliselt oluline. Mudel paranes, sest korrigeeritud determinatsioonikordaja suurenes: lineaarse mudeli korral $R_a^2 = 0{,}134$, ruutliikmega mudeli korral $R_a^2 = 0{,}177$. Transport: $\hat{y}_{transp} = -321 + 0{,}242x - 1{,}41 \cdot 10^{-4}x^2$. Ruutliikme olulisuse tõenäosus $p = 2{,}11 \cdot 10^{-12} < 0{,}05$, on statistiliselt oluline. Mudel paranes, sest korrigeeritud determinatsioonikordaja suurenes: lineaarse mudeli korral $R_a^2 = 0{,}323$, ruutliikmega mudeli korral $R_a^2 = 0{,}386$. Vaba aeg: $\hat{y}_{v.aeg} = -11{,}4 + 0{,}077x + 1{,}91 \cdot 10^{-4}x^2$. Ruutliikme olulisuse tõenäosus $p = 3{,}33 \cdot 10^{-86} < 0{,}05$, on statistiliselt oluline. Mudel paranes, sest korrigeeritud determinatsioonikordaja suurenes: lineaarse mudeli korral $R_a^2 = 0{,}543$, ruutliikmega mudeli korral $R_a^2 = 0{,}790$. Kõikides mudelites on x kulud kokku. Riiete ja transpordi korral on ruutliikme kordaja negatiivne, järelikult suuremate kogukulude korral kasv aeglustub. Vaba aja kulude korral on ruutliikme kordaja positiivne, järelikult suuremate kogukulude korral kasvavad kulud vabale ajale kiiremini kui väikeste kogukulude korral.

A.9.23 2. $\ln \widehat{CPI} = 1{,}713 + 0{,}228 \ln GDP$. 3. Jah, F-testi olulisuse tõenäosus $1{,}79 \cdot 10^{-30} < 0{,}05$. 4. $\mu = 0{,}228 \pm 0{,}032$, usaldusvahemik kattub osaliselt Shao jt usaldusvahemikuga. 5. Riigid, kus standardiseeritud jääk on väiksem kui -3: Turkmenistan, Iraak, Sudaan. 6. $\ln \widehat{CPI} = 1{,}089 + 0{,}297 \ln GDP$. 7. Mudel on statistiliselt oluline, sest F-testi olulisuse tõenäosus $6{,}38 \cdot 10^{-9} < 0{,}05$. 8. $\mu = 0{,}297 \pm 0{,}075$. 9. Eesti CPI mudelväärtus 56,1 on väiksem kui empiiriline väärtus 69, s.t Eestis on korruptsiooni tajumine väiksem kui elatustaseme põhjal võiks prognoosida.

A.9.24 1. Lineariseeritud mudel: $\ln T_{ij} = \ln k + p \ln(T_i T_j) - q \ln D$. 2. Lineaarse mudeli hinnang $\ln \hat{T}_{ij} = -15{,}9 + 1{,}165 \ln(T_i T_j) - 0{,}41 \ln D$. 3. Jah, kauguse D naturaallogaritm on statistiliselt oluline, $p = 1{,}79 \cdot 10^{-8}$. 4. $\hat{T}_{ij} = \frac{1{,}2 \cdot 10^{-7} (T_i T_j)^{1{,}165}}{D^{0{,}41}}$.

A.9.25 1. $\hat{y} = 1{,}168 + 132(1/D)$. 2. Ei ole, $p = 0{,}12 > 0{,}05$. 3. $\hat{y} = \frac{174{,}3}{D}$. 4. Jah, mingist maakonnast pärit patsientide arv 1000 elaniku kohta on pöörvõrdeline vastava maakonna kaugusega haiglast.

A.9.26 1. $\hat{y} = 10{,}24 + 0{,}47x$, kus y on kümnes minutis kulunud kilokalorite kogus ning x isiku kehakaal kilogrammides. Kaalu olulisuse tõenäosus $1{,}97 \cdot 10^{-15} < 0{,}05$, kaal on statistiliselt oluline ja mõjutab energiakulu: iga lisakilogramm suurendab jalutamisel energiakulu 0,47 kcal võrra kümnes minutis. 2. Vanuse olulisuse tõenäosus $0{,}83 > 0{,}05$, vanus ei mõjuta energiakulu. 3. Soo olulisuse tõenäosus $0{,}29 > 0{,}05$, sugu ei mõjuta energiakulu.

A.9.27 1. $72{,}61\% \pm 0{,}77\%$. 2. $25{,}09 \pm 0{,}31$. 3. $27{,}6\% \pm 2{,}0\%$. 4. Mudel on

$$\widehat{\mathrm{LAEN}} = 35{,}3 + 0{,}35432\,\mathrm{HIND} + 1{,}112\,\mathrm{HSUHE} - 13{,}27\,\mathrm{OMAND} + 12{,}39\,\mathrm{PEREK}.$$

5. Kõik seletavad tunnused on olulised nivool 0,01. 6. a) Kui kinnisvara hind on tuhande dollari võrra suurem, siis laen on ligikaudu 0,35 tuhande dollari võrra suurem. b) Kaasomandis oleva kinnisvara korral võetakse laene ligikaudu 13 tuhande dollari võrra vähem. c) Kui laenuvõtja on abielus, siis võetakse laenu *ca* 12 tuhande dollari võrra rohkem.

A.9.28 1. Mudel on

$$\begin{aligned}\ln \widehat{\mathrm{TTASU}} = {} & 4{,}588 + 0{,}257\,\ln \mathrm{KÄIVE} + 0{,}0112\,\mathrm{ROE} + 0{,}158\,D_1 + \\ & + 0{,}181\,D_2 - 0{,}283\,D_3.\end{aligned}$$

2. D_3 on oluline nivool 0,01 ($p = 0{,}0048$), D_2 on oluline nivool 0,05 ($p = 0{,}034$) ja D_1 on oluline nivool 0,1 ($p = 0{,}077$). 3. Kuna tegevusalale vastavad fiktiivsed tunnused on statistiliselt olulised, siis tegevusala mõju töötasule on tõestatud. 4. 1312,9 tuhat dollarit aastas. 5. 1537,5 tuhat dollarit aastas. 6. Kõige suurem tarbekaupade tootmisel, kõige väiksem transpordis. 7. 0,158. 8. Valemist (A.81) saame, et $\frac{\Delta\hat{y}}{\hat{y}} = e^{0{,}158} - 1 = 0{,}171$, järelikult 17,1% suurem. 9. 24,6% väiksem.

A.9.29 Konstruktsiooni tüübi lisamisel mudelisse saame

$$\begin{aligned}\widehat{\mathrm{HIND}} = {} & 13{,}6 + 11{,}6\,\mathrm{VANNIT} + 1{,}10\,\mathrm{KRUNT} + 3{,}02\,\mathrm{GARAAZ} - \\ & - 0{,}115\,\mathrm{VANUS} + 0{,}73D_2 + 2{,}59D_3 + 2{,}72D_4.\end{aligned}$$

Fiktiivne tunnus D_2 vastab konstruktsiooni 2. tüübile puitsõrestik ja telliskivi, tunnus D_3 vastab 3. tüübile alumiiniumsõrestik ning D_4 4. tüübile puitsõrestik. 1. tüüp telliskivi on baaskategooria. Kuna kõik fiktiivsed tunnused on statistiliselt mitteolulised ($p_{D2} = 0{,}686$, $p_{D3} = 0{,}189$, $p_{D4} = 0{,}136$), siis konstruktsiooni tüübi mõju maja hinnale ei ole tõestatud.

A.9.30 1. Standardiseeritud kordajad: HIND $-0{,}537$, TULU 2,592, ARV $-1{,}175$. 2. Kõige rohkem mõjutab müüdud autode arvu elanike sissetulek: kui sissetulek suureneb ühe standardhälbe võrra, siis müüdud autode arv suureneb ligikaudu 2,6 standardhälbe võrra. Kõige vähem mõjutab autode müüki hind: kui hind suureneb ühe standardhälbe võrra, väheneb autode müük ligikaudu 0,54 standardhälbe võrra.

A.9.31 Lineariseeritud mudel on

$$\ln \hat{y} = 2{,}039 + 0{,}5811 \ln K + 0{,}2655 \ln L + 0{,}0945 \ln T, \quad R^2 = 0{,}925,$$

kus y on käive (tuhat eurot), K varad (tuhat eurot), L töötajate arv, ning T teadus- ja arenduskulud (tuhat eurot). Cobbi-Douglase kuju:

$$\hat{y} = 7{,}684 K^{0{,}5811} L^{0{,}2655} T^{0{,}0945}.$$

Kõige rohkem mõjutab käivet varade suurus: varade suurenemisel 1% võrra suureneb käive 0,581%. Töötajate arvu suurenedes 1% võrra suureneb käive 0,266% ning teadus- ja arenduskulude suurendamine 1% võrra suurendab käivet 0,0945%.

A.9.32 1. $\ln \hat{Q} = -0{,}177 + 0{,}233 \ln K + 0{,}807 \ln L$. 2. 1,04. 3. $\ln A = 0{,}0145$, $\alpha = 0{,}254$. 4. $\hat{Q} = 1{,}01 K^{0{,}254} L^{0{,}746}$. 5. Vigade piires ühtivad.

A.9.33 1. Kulufunktsioonid on järgmised:

Kululiik	Kulufunktsioon	Seletavad tunnused
Kulud kütusele	$\hat{C}_1 = -406{,}7 + 361{,}3 x_6 + 1{,}165 x_7$	kütuse hind x_6, kütuse kogus x_7
Tööjõukulu	$\hat{C}_2 = 731{,}6$	
Kulud siseliinidele	$\hat{C}_3 = -13{,}26 + 0{,}140 x_5$	kohtmiilid siselendudel x_5
Lennukite liisimine	$\hat{C}_4 = 169{,}8 + 0{,}144 x_2$	liisitud lennukite arv x_2
Lennujaamamaksud	$\hat{C}_5 = 29{,}04 + 0{,}00989 x_3$	reisijate arv x_3
Turustuskulud	$\hat{C}_6 = 169{,}5$	
Lennukite hoolduskulud	$\hat{C}_7 = 80{,}7 + 0{,}0022 x_4$	kohtmiilid x_4
Amortisatsioonikulu	$\hat{C}_8 = 105{,}6$	
Reisijate teenindus	$\hat{C}_9 = 46{,}7 + 0{,}00274 x_3$	reisijate arv x_3
Muud kulud	$\hat{C}_{10} = 145{,}6 + 0{,}059 R$	kogutulu R

2. a)–c) Kululiikide, kogukulude ja kasumi prognoosid 2009. aasta kvartalite kaupa on järgnevas tabelis, kõik suurused on miljonites dollarites.

Kululiik	Q1	Q2	Q3	Q4
Kulud kütusele	720,4	842,1	742,2	866,7
Tööjõukulu	731,6	731,6	731,6	731,6
Kulud siseliinidele	404,0	414,3	426,3	408,4
Lennukite liisimine	227,0	226,4	224,4	224,2
Lennujaamamaksud	171,6	190,7	195,2	180,0
Turustuskulud	169,5	169,5	169,5	169,5
Lennukite hoolduskulud	138,7	142,4	144,4	138,6
Amortisatsioonikulu	105,6	105,6	105,6	105,6
Reisijate teenindus	86,2	91,5	92,8	88,5
Muud kulud	320,5	308,9	319,6	290,9
KULUD KOKKU	3075,0	3223,1	3151,5	3204,1
KASUM	−113,0	−456,1	−204,5	−742,1

d) 2009. aasta prognoositud kasum on −1515,7 miljonit dollarit.

10. ptk. Aegread

10.1 5 senti. **10.2** Vastused on tabelis.

Aasta	Absoluutne ahel-juurdekasv	Absoluutne alus-juurdekasv	Ahel-juurdekasvu-tempo	Tagasivaatav aheljuurde-kasvutempo
2010	35,9	35,9	0,008	0,008
2011	557,4	593,3	0,120	0,107

Aasta	Ahel-kasvutempo ehk ahelindeks	Alus-kasvutempo ehk alusindeks
2010	1,008	1,008
2011	1,120	1,129

10.3 1. 1,1329. 2. 16 258,2 mln eurot. 3. Tegelik väärtus oli väiksem 23,1 mln eurot ehk 0,14%. **10.4** 0,669 mln eurot. **10.5** I kv 190 tuhat eurot, II kv 547 tuhat eurot, aasta keskmine 368 tuhat eurot. **10.6** 1. $30\frac{5}{12} \approx 30{,}4167$. 2. Juuni 1,0139, juuli ja august 0,9812. 3. Juuni 819,22 kWh, juuli 794,76 kWh, august 796,72 kWh. Korrigeeritud tarbimine oli kõige suurem juunis ja kõige väiksem juulis. **10.7** Juulis 37,6 tuhat eurot, augustis 42,5 tuhat eurot ja septembris 38,9 tuhat eurot. Kõige suurem korrigeeritud käive oli augustis. **10.8** 9,66 eurot. **10.9** a) 27,8 tuhat eurot; b) 27,4 tuhat eurot. **10.10** 1. 88,2%. 2. 88,5%. **10.11** 0,35. **10.12** 1. Aastatel 1970–1990 suurenes Eesti rahvaarv keskmiselt 10,08 tuhande elaniku võrra aastas. See on iive. 2. $\hat{N} = 1369 + 0{,}84t$. **10.13** 1. Kasv 1923 tuhat tonni ehk 53,7%. 2. Kasv 549 tuh tonni ehk 16,6%. **10.14** 1. 47,2 USD senti. 2. Ligikaudu 27%. 3. 2018. **10.15** Parabooli miinimumkoht on $6{,}73 \approx 7$, järelikult juulis. **10.16** MSE eurot^2, $RMSE$ eurot, ME eurot, MAD eurot, MPE ühikuta või protsentides ja $MAPE$ ühikuta või protsentides. **10.17** Alahindab. **10.18** Suurema läbipaistvuse korral on prognoosid täpsemad, läbipaistvuse indeksi suurenemine 1 võrra vähendab prognoosi suhtelist absoluutviga 0,01 võrra ehk 1 protsendipunkti võrra.

A.10.1 a) Väärtused 30.04.2014: absoluutne aheljuurdekasv 0,00240, kasvutempo 1,00174 ja juurdekasvutempo 0,174%; b) kuu keskmine kurss 1,381, keskmine absoluutne aheljuurdekasv 0,000316, keskmine kasvutempo 1,000229 ja keskmine juurdekasvutempo 0,0229%.

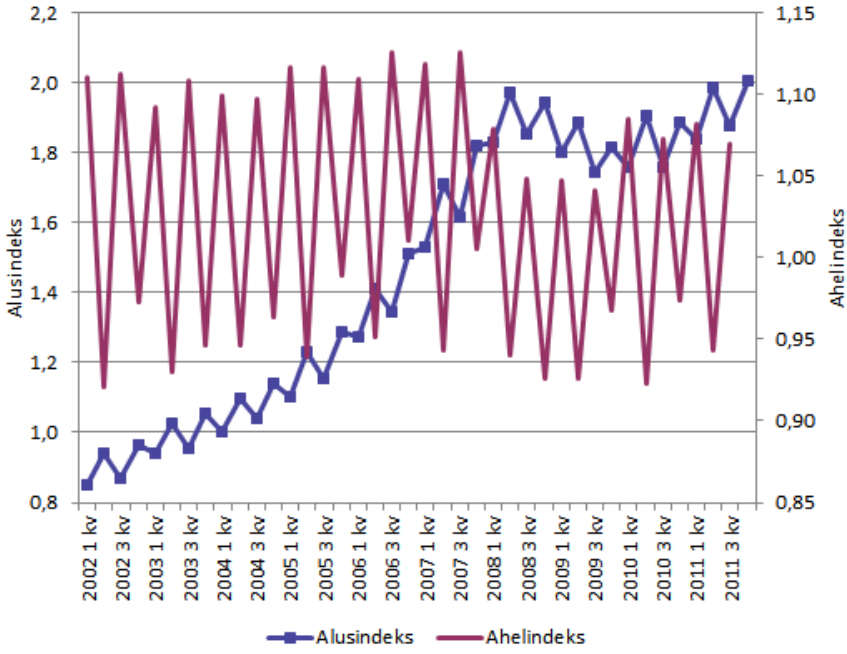
A.10.2 Väärtused 2011. aasta 4. kvartalis: alusindeks 2,0061 ja ahelindeks 1,0696. Diagramm on joonisel ÜV.16.

A.10.3 Kvartalite keskmised miljonites eurodes:

I 6874,0,
II 6860,7,
III 6877,5,
IV 6902,5.

Aasta keskmine 6878,7 miljonit eurot.

A.10.4 19,22 °C.



Joonis ÜV.16. Ülesande A.10.2 vastuse juurde

A.10.5 Kõige suurem koorimuusika kontserdil (1,22) ja kõige väiksem lastekontserdil (1,053).

A.10.6 a) 2012. aasta, kuu keskmine kasvutempo 0,982; b) märts, 133,08 tuhat tonni.

A.10.7 1. Keskmine absoluutne aheljuurdekasv 0,618 €. 2. E: −0,445 €, T: −2,93 €, K: 0,336 €, N: 3,36 €, R: 2,85 €. 3. Teisipäeval, sest siis on hind keskmiselt kõige rohkem langenud.

A.10.8 2. 10 kuu libiseva keskmise järgi 1,400; 20 kuu libiseva keskmise järgi 1,356. 3. 20 päeva libisev keskmine. 4. Joonis ÜV.17.

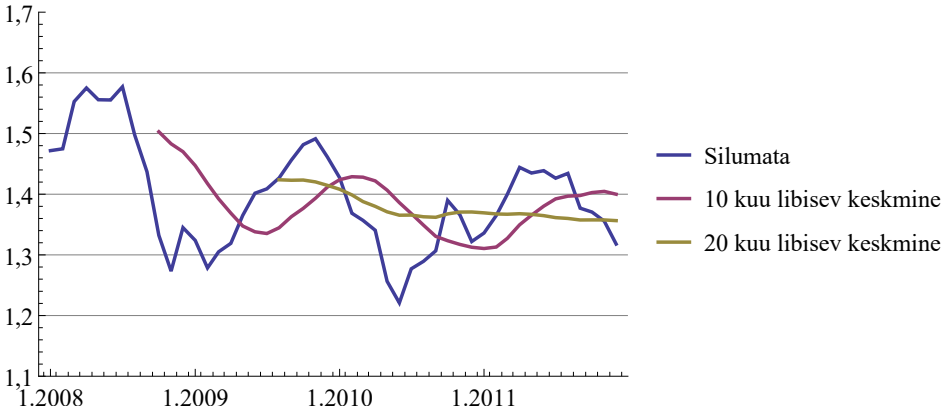
A.10.9 1. 7 päeva libisev keskmine. 2. Joonis ÜV.18 3. 47,7 €.

A.10.10 Bollingeri piirid on tabelis.

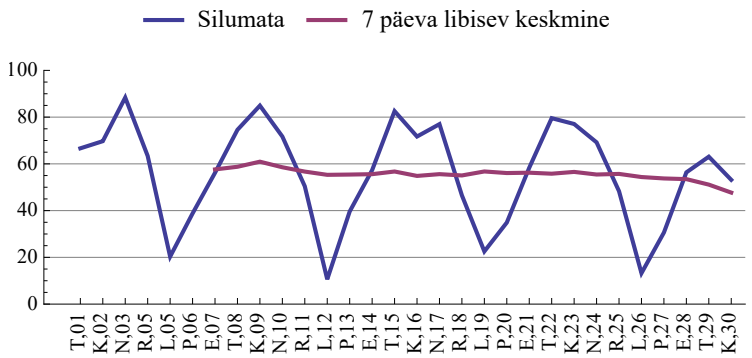
Kuupäev	Alumine piir	Ülemine piir
30.01.2015	0,8010	0,9710
2.02.2015	0,8071	0,9771
3.02.2015	0,8071	0,9953

A.10.11 Viimased silutud väärtused 30.12.2015: 895,96, kui $w = 0{,}2$ ja 898,65, kui $w = 0{,}7$.

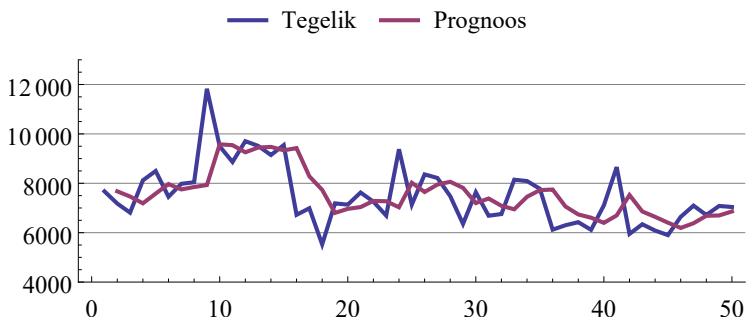
A.10.12 1. Viimane silutud väärtus 6923,65 dollarit. 2. 50. nädala prognoos 6845,82 dollarit. 3. Prognoosiviga 50. nädalal 194,57 dollarit. 4. $MSE = 1\,175\,013$. 5. Kui $w = 0{,}3$, siis $MSE = 1\,191\,260$. Kui $w = 0{,}5$, siis $MSE = 1\,180\,599$. MSE on kõige väiksem, kui $w = 0{,}4$. 6. $w \approx 0{,}4227$, $MSE = 1\,174\,458$ 7. Vt joonis ÜV.19.



Joonis ÜV.17. Ülesande A.10.8 vastuse juurde



Joonis ÜV.18. Ülesande A.10.9 2. osa vastus



Joonis ÜV.19. Ülesande A.10.12 7. osa vastus

A.10.13 1. Ida-Viru $\hat{y} = 91{,}9t + 1762{,}6$ ja Lääne-Viru $\hat{y} = 59{,}5t + 971{,}3$, kus y on ettevõtete arv ja t aastates, $t = 0$ aastal 2000. 2. Ida-Viru maakonnas lisandus keskmiselt 91,9 ja Lääne-Viru maakonnas 59,5 ettevõtet aastas. 3. Prognoos Ida-Viru jaoks 2314 ja Lääne-Viru jaoks 1328 ettevõtet. 4. Erinevus prognoosist: Ida-Viru 5,4% ja Lääne-Viru 11,9%.

A.10.14 1. $r = 0{,}285$. 2. $\hat{y} = 9{,}65t + 602{,}3$, kus y on banaanide käive tuhandetes eurodes ja t aeg kvartalites, $t = 1$ 1999. aasta I kv. 4. $r = 0{,}978$. 5. Kui ühe tunnuse aegrida sisaldab sesoonsust ja teise tunnuse aegrida mitte, siis seose tugevuse hindamiseks tuleb sesoonset aegrida enne siluda.

A.10.15 1. Tähistagu y THI-d ja t aega kuudes ($t = 1$ jaan 2004). Lineaarne mudel $\hat{y} = 0{,}2752t + 112{,}32$, $R^2 = 0{,}9126$; ruutpolünoom $\hat{y} = 0{,}0045t^2 + 0{,}055t + 114{,}16$, $R^2 = 0{,}9499$; eksponentsiaalne $\hat{y} = 112{,}47e^{0{,}0023t}$. 2. Parim mudel on ruutpolünoom, sest R^2 on kõige suurem. 3. Prognoos 127,66.

A.10.16 Parim mudel on eksponentsiaalne $\hat{y} = 5527{,}7e^{0{,}0751t}$, $R^2 = 0{,}9938$, kus y on raha pakkumine mln kr ja t aeg kvartalites, $t = 1$ 1994. a I kv.

A.10.17 1. Vt joonis 10.16. 2. $\hat{N} = 13{,}349t + 1056{,}9$, $R^2 = 0{,}9885$, kus N on rahvaarv tuhandetes ja aeg t aastates, $t = 0$ aastal 1950. Keskmine rahvaarvu suurenemine *ca* 13,3 tuhat elanikku aastas. 3. $\hat{N} = -0{,}1301t^2 + 18{,}813t + 1017{,}8$, $R^2 = 0{,}999$. 1960. aastal suurenemine *ca* 16,2 tuhat elanikku aastas, 1990. aastal suurenemine *ca* 8,4 tuhat elanikku aastas. 4. $\hat{N} = -5{,}5074t + 1396{,}3$, $R^2 = 0{,}9566$, kus N on rahvaarv tuhandetes ja aeg t aastates, $t = 0$ aastal 2000. Keskmine rahvaarvu vähenemine *ca* 5,5 tuhat elanikku aastas. 5. $\hat{N} = 0{,}2563t^2 - 10{,}12t + 1410{,}9$, $R^2 = 0{,}996$. Aastal 2000 vähenemine *ca* 10,1 tuhat elanikku aastas, aastal 2015 vähenemine *ca* 2,4 tuhat elanikku aastas.

A.10.18 1. $\hat{y} = 127{,}54e^{0{,}3488t}$, $R^2 = 0{,}97$, kus y on laenujääk mln eurot ja t aeg aastates, $t = 1$ aastal 1997. 3. $\ln \hat{y} = 0{,}3488t + 4{,}848$. 4. $a = 0{,}3488$, $y_0 = 127{,}5$. 5. Parameetrid langevad kokku.

A.10.19 2.(a) 40,55 kg; (b) 37,73 kg; (c) 39,37 kg; (d) 39,81 kg. 3. Lineaarne trend, prognoosi viga 0,08 kg. Tarbimine kõigub ümber lineaarse trendi.

A.10.20 2.(a) 49,69 mln €; (b) 47,86 mln €; (c) 50,35 mln €; (d) 49,45 mln €. 3. Kaalutud libisev keskmine, prognoosi viga 10,4 mln €.

A.10.21 1. Eksponentsiaalne trend $\hat{y} = 1233{,}9e^{0{,}0189t}$, kus y on sõitjate arv tuhandetes ja t aeg kvartalites, $t = 1$ 2005. a I kvartalis. Selle mudeli $R^2 = 0{,}2883$. 4.–5. Keskmised sesoonsed komponendid ja prognoos 2013. aastaks on toodud tabelis. 6. $MSE = 62347{,}2$.

Kvartal	Keskmine sesoonne komponent
I	−494,48
II	205,31
III	697,02
IV	−168,16

Kvartal	Prognoos
I	1807,75
II	2551,47
III	3087,95
IV	2268,38

A.10.22 1. Parim on eksponentsiaalne mudel $\hat{y} = 290{,}25e^{0{,}0155t}$, kus t on aeg kuudes, $t = 1$ jaan. 1997. 3.–4. Keskmised sesoonsed komponendid ja prognoos 2001. aasta juulini on tabelites. 5. $MSE \approx 1268{,}5$.

Kuu	Keskmine sesoonne komponent	Kuu	Keskmine sesoonne komponent
1	1,002	7	0,856
2	0,910	8	0,882
3	0,966	9	1,046
4	0,977	10	0,997
5	1,264	11	1,060
6	0,973	12	1,119

Kuu	Prognoos
01.2001	621,79
02.2001	573,23
03.2001	617,93
04.2001	634,91
05.2001	834,15
06.2001	652,23
07.2001	582,51

A.10.23 1.–2. vt tabelit, kus on arvutused kuni 2006. a IV kvartalini. 3. $MSE \approx 256873$. 4. Silumiskonstandid: $w \approx 0{,}194$, $v \approx 0{,}591$, $\alpha \approx 0{,}780$. 5. 2013. aasta IV kvartali prognoos on 1944,7 tuhat. 6. $MSE \approx 2644{,}5$. 7. Eksponentsilumine trendi ja sesoonsusega sobib rohkem, sest mudeli testimine andis väiksema MSE.

Aasta	Kvartal	t	Sõitjate arv, tuh	E	T	S	Prognoos F	Prognoosi viga
2005	I	1	863,9					
	II	2	1553,9	1553,9	690,0	0,0		
	III	3	1983,8	2217,9	677,0	−234,1		
	IV	4	1135,2	2718,9	589,0	−1583,7		
2006	I	5	792,4	3056,4	463,2	−2264,0		
	II	6	1423,1	3310,0	358,4	−1132,1		
	III	7	2117,0	3536,6	292,5	−945,4	3434,3	−1317,3
	IV	8	1518,5	3756,5	256,2	−1976,3	2245,5	−727,0
...	...	...	...	...	...	...	...	...

A.10.24 1.–2. ja 5. osa vt tabeleid, kus on arvutused jaanuarist kuni aprillini 1997 ning veebruarist kuni aprillini 1998 ja prognoos valimist välja kuni juulini 2001. 3. $MSE \approx 6174{,}8$. 4. Silumiskonstandid: $w \approx 0{,}245$, $v \approx 0{,}149$, $\alpha \approx 0{,}776$. 6. $MSE \approx 2431{,}6$. 7. Kompleksanalüüs regressioonjoone ning sesoonsusega sobib rohkem, sest selle mudeli testimine andis väiksema $MSE \approx 1268{,}5$.

Kuu	t	Käive, tuh €	E	T	S	Prognoos F	Prognoosi viga
01.1997	1	309,92					
02.1997	2	252,08	252,08	−57,836	1		
03.1997	3	309,44	222,52	−53,624	1,391		
04.1997	4	297,11	200,36	−48,936	1,483		
...	...	...	...	...	...	...	...
02.1998	14	364,19	284,69	−1,518	1,217		
03.1998	15	343,40	270,51	−3,405	1,297	393,79	−50,38
04.1998	16	347,97	247,97	−6,256	1,421	396,09	−48,12
...	...	...	...	...	...	...	...

Prognoos valimist välja

Kuu	t	Käive, tuh €	Prognoos F	Prognoosi viga
01.2001	1	601,22	548,72	52,50
02.2001	2	540,20	527,95	12,25
...	...	...	...	...
07.2001	7	593,58	580,56	13,02

A.10.25 $MSE = 121{,}9$ reisijat^2, $RMSE = 11{,}04$ reisijat, $MAPE = 2{,}37\%$.

A.10.26 $MAPE$ väärtused on toodud tabelis. Kõige parema prognoosi andis eksponentsilumine konstandiga 0,5.

Libisev keskmine			Eksponentsilumine			Lineaarne
$k = 3$	$k = 5$	$k = 7$	$w = 0{,}3$	$w = 0{,}5$	$w = 0{,}8$	regressioon
10,2%	12,9%	16,9%	13,1%	7,8%	11,4%	12,7%

11. ptk. Indeksid

11.1 Vastused on tabelis.

	2010	2011	2012	2013
			Alusindeks	
Harju maakond	1	1,12	1,21	1,29
Tartu maakond	1	1,08	1,18	1,27
			Ahelindeks	
Harju maakond		1,12	1,08	1,06
Tartu maakond		1,08	1,09	1,08

11.2 1. 72%. 2. 2,2%. 3. Vähenemine 2%. **11.3** 1. Vähenes aastatel 2003 ja 2007–2010. 2. Vähenes 10%. 3. Suurenes *ca* 15%. 4. Suurenes *ca* 33%. **11.4** Hinnaindeks 0,9, mahuindeks 1,3, käibeindeks 1,17. **11.5** Hinnaindeksid: bensiin 95E 0,910, bensiin 98 0,923, diislikütus 0,843. Kõige enam odavnes diislikütus. **11.6** 1,0255. **11.7** Vähenesid 16,5%. **11.8** 2%. **11.9** Toode A 30% ja toode B 70%. **11.10** Toodangu

maksumuse koondindeks 1,024 iseloomustab kogumuutust: maksumus suurenes 2,4%. Koguseindeks 1,068 näitab ainult koguste muutusest tingitud maksumuse muutust ja see oli 6,8%. Hinnaindeks 0,959 näitab hindade muutumisest tingitud muutust ning need põhjustasid maksumuse vähenemist 4,1%. **11.11** 1. 31,5%. 2. 2014. aastal 16,2 km ja 2015. aastal 20,4 km. 3. 4,1%. 4. 26,3%. **11.12** Keskmine omahind väheneb 6,4%. Üksikute toodete omahinna muutus põhjustas keskmise omahinna tõusu 4,0%. Muutused üksikute toodete tootmismahtudes põhjustasid keskmise omahinna languse 10,0%. **11.13** 1. 2015. aastal 8,06 mln tonni ja 2016. aastal 4,58 mln tonni. 2. 2015. aastal 5,0 t/ha ja 2016. aastal 2,9 t/ha. 3. Vähenes 43,2%. 4. Vähenes 44,6%. 5. Suurenes 2,6%. **11.14** Paasche indeks 1,98, Laspeyresi indeks 1,66 ja Fisheri indeks 1,81. Kõige suuremat inflatsiooni näitab Paasche indeks. **11.15** 1. 2,52%. 2. 1,01%.

A.11.1 8,6%.

A.11.2 a) 1,032; b) 1,079; c) 0,956; d) 1850 €; e) 4440 €; f) −2590 €.

A.11.3 a) Maksumuse kogumuutus 42,0%. b) Hindade muutumisest põhjustatud osamuutus 1,8% ja koguste muutumisest põhjustatud osamuutus 39,5%. c) Keskmise hinna kogumuutus 29,3%. d) Hindade muutumisest põhjustatud keskmise hinna osamuutus 1,8% ja koguste muutusest tingitud osamuutus 27,0%.

A.11.4 a) 2014. aastal 2546 €/m^2, 2015. aastal 2507 €/m^2; b) vähenemine 1,6%; c) suurenemine 1,3%; d) vähenemine 2,9%.

A.11.5 1. Piimatoodete hind tõusis keskmiselt 19,3% ehk 20 senti. 2. 14,8% (16 senti) tõusis hind üksikute toodete hindade muutumise tõttu. 3. 4% (4 senti) tõusis hind ostukoguste muutumise tõttu.

A.11.6 Muutuva struktuuri indeks 0,99, järelikult keskmine netokäive ühe töötaja kohta väheneb 1%. Püsiva struktuuri indeks 0,917 ja struktuurinihete indeks 1,079. Ühe töötaja kohta tuleva netokäibe vähenemine üksikutes ettevõtetes põhjustas keskmise näitaja vähenemise 8,3% võrra. Töötajate arvu muutus ettevõtetes põhjustas keskmise näitaja suurenemise 7,9% võrra.

A.11.7 Struktuuriindeksid on tabelis.

Aasta	Muutuva struktuuri indeks	Püsiva struktuuri indeks	Struktuurinihete indeks
2009	0,950	0,945	1,005
2010	1,011	1,007	1,003
2011	1,059	1,057	1,002
2012	1,057	1,059	0,998
2013	1,070	1,070	1,000
2014	1,058	1,064	0,995
2015	1,060	1,055	1,005

Lisa A

Mõningate valemite tõestusi

A.1. Dispersiooni arvutusvalemid

Alapeatükis 3.3 tõime dispersiooni arvutusvalemi kujul

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

kus $\bar{x}$ on aritmeetiline keskmine. Näitame, et dispersiooni arvutamiseks võib kasutada ka valemit

$$\sigma^2 = \overline{x^2} - \bar{x}^2, \tag{A.2}$$

s.t dispersioon on ruutude aritmeetiline keskmine miinus aritmeetilise keskmise ruut.

Selleks teisendame valemi (A.1) lugejat:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} - \bar{x}^2\right) = \sum_{i=1}^{n}x_i^2 - \sum_{i=1}^{n}2x_i\bar{x} + \sum_{i=1}^{n}\bar{x}^2. \tag{A.3}$$

Teise summa teisendamisel arvestame, et $\sum_{i=1}^{n} x_i = n\bar{x}$:

$$\sum_{i=1}^{n}2x_i\bar{x} = 2\bar{x}\sum_{i=1}^{n}x_i = 2\bar{x}n\bar{x} = 2n\bar{x}^2. \tag{A.4}$$

Avaldises (A.3) olevas viimases summas on n ühesugust liidetavat $\bar{x}^2$, järelikult

$$\sum_{i=1}^{n}\bar{x}^2 = n\bar{x}^2. \tag{A.5}$$

Arvestades valemeid (A.4) ja (A.5), saame (A.3) jaoks

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum_{i=1}^{n}x_i^2 - n\bar{x}^2. \tag{A.6}$$

Vastavalt valemile (A.1) tuleb dispersiooni saamiseks see avaldis jagada n-ga:

$$\sigma^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \overline{x^2} - \bar{x}^2. \tag{A.7}$$

Saimegi valemi (A.2). Märgime, et sellistes kalkulaatorites, kus on olemas dispersiooni arvutus, tuleb väärtused x_i ükshaaval sisestada. Seal kasutatakse valemi (A.2) modifikatsiooni, kus aritmeetiliste keskmiste asemel on summad ning igal sisestamisel neid summasid uuendatakse:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x\right)^2. \tag{A.8}$$

A.2. Ristkülikjaotuse dispersioon

Lõigul $[a, b]$ esineva ristkülikjaotuse dispersiooni leidmiseks lähtume valemist (5.38), millesse paneme jaotustiheduse (5.41) ja keskväärtuse (5.43). Rajadeks võtame lõigu otspunktid a ja b, sest väljaspool lõiku on jaotustihedus ja järelikult ka integraal 0.

$$\begin{aligned}
\sigma^2 &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_a^b \left(x - \frac{1}{2}(a+b)\right)^2 \frac{1}{b-a}dx = \\
&= \frac{1}{b-a}\int_a^b \left(x^2 - x(a+b) + \frac{1}{4}(a+b)^2\right)dx = \\
&= \frac{1}{b-a}\left[\frac{x^3}{3} - (a+b)\frac{x^2}{2} + \frac{1}{4}(a+b)^2 x\right]_a^b = \\
&= \frac{1}{b-a}\left[\frac{b^3-a^3}{3} - (a+b)\frac{b^2-a^2}{2} + \frac{1}{4}(a+b)^2(b-a)\right] = \\
&= \frac{b^2+ab+a^2}{3} - \frac{(a+b)^2}{2} + \frac{1}{4}(a+b)^2 = \\
&= \frac{b^2+ab+a^2}{3} - \frac{(a+b)^2}{4} = \\
&= \frac{4b^2+4ab+4a^2-3a^2-6ab-3b^2}{12} = \\
&= \frac{b^2-2ab+a^2}{12} = \\
&= \frac{(b-a)^2}{12}.
\end{aligned}$$

Järelikult ristkülikjaotuse dispersioon

$$\sigma^2 = \frac{(b-a)^2}{12}. \tag{A.9}$$

A.3. Poissoni jaotus kui binoomjaotuse piirjuht

Näitame, et Poissoni jaotus on binoomjaotuse piirjuht, kui katsete arv n läheneb lõpmatusele. Lähtume binoomjaotuse korral kehtivast Bernoulli valemist (5.49):

$$P(X=m) = C_n^m p^m (1-p)^{n-m}, \tag{A.10}$$

kus C_n^m on kombinatsioonide arv n elemendist m kaupa:

$$C_n^m = \frac{n!}{m!(n-m)!}. \tag{A.11}$$

Positiivse tulemuse tõenäosuse p avaldame binoomjaotuse keskväärtuse kaudu valemist (5.53):

$$p = \frac{\bar{x}}{n}. \tag{A.12}$$

Kompaktsuse huvides võtame keskväärtuse jaoks kasutusele tähistuse $\lambda = \bar{x}$ ja kirjutame valemi (A.10) välja keskväärtuse kaudu. Ühtlasi asetame sinna ka kombinatsioonide arvu avaldise (A.11)

$$P(X=m) = \frac{n!}{m!(n-m)!}\left(\frac{\lambda}{n}\right)^m \left(1-\frac{\lambda}{n}\right)^{n-m}. \tag{A.13}$$

Avaldise (A.13) teisendamiseks arvestame, et faktoriaal

$$n! = 1{\cdot}2{\cdot}\ldots{\cdot}(n-m){\cdot}(n-m+1){\cdot}\ldots{\cdot}(n-1){\cdot}n = (n-m)!{\cdot}(n-m+1){\cdot}\ldots{\cdot}(n-1){\cdot}n \tag{A.14}$$

ja järelikult

$$\frac{n!}{m!(n-m)!} = \frac{(n-m+1)\cdot\ldots\cdot(n-1)\cdot n}{m!}.$$

Tehes selle teisenduse avaldises (A.13), saame

$$\begin{aligned} P(X=m) &= \frac{(n-m+1)\cdot\ldots\cdot(n-1)\cdot n}{m!}\left(\frac{\lambda}{n}\right)^m\left(1-\frac{\lambda}{n}\right)^{n-m} = \\ &= \frac{(n-m+1)\cdot\ldots\cdot(n-1)\cdot n}{m!}\frac{\lambda^m}{n^m}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-m} = \\ &= \frac{(n-m+1)\cdot\ldots\cdot(n-1)\cdot n}{n^m}\frac{\lambda^m}{m!}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-m}. \end{aligned}$$

Läheme üle piirile, kus katsete arv $n \to \infty$.

$$\begin{aligned} \lim_{n\to\infty} P(X=m) &= \lim_{n\to\infty}\left[\frac{(n-m+1)\cdot\ldots\cdot(n-1)\cdot n}{n^m}\frac{\lambda^m}{m!}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-m}\right] = \\ &= \lim_{n\to\infty}\frac{(n-m+1)\cdot\ldots\cdot(n-1)\cdot n}{n^m}\cdot\lim_{n\to\infty}\frac{\lambda^m}{m!}\cdot\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n\cdot \\ &\cdot\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-m}. \end{aligned} \tag{A.15}$$

Leiame esimese teguri piirväärtuse:

$$\lim_{n\to\infty} \frac{(n-m+1)\cdot \ldots \cdot (n-1)\cdot n}{n^m} = ?$$

Lugejas on korrutis, milles on m tegurit. Ka nimetaja võime lahti kirjutada korrutisena, milles on m tegurit: $n^m = n \cdot n \cdot \ldots \cdot n$. Nii saame m tegurit, millest igaühest tuleb võtta piirväärtus:

$$\lim_{n\to\infty} \frac{(n-m+1)\cdot \ldots \cdot (n-1)\cdot n}{n^m} = \lim_{n\to\infty} \left[\frac{n-m+1}{n} \cdot \ldots \cdot \frac{n-1}{n} \cdot \frac{n}{n}\right] =$$
$$= \lim_{n\to\infty} \left(1 - \frac{m+1}{n}\right) \cdot \ldots \cdot \lim_{n\to\infty} \left(1 - \frac{1}{n}\right) \cdot 1.$$

On näha, et kõikide tegurite piirväärtus on 1. Järelikult

$$\lim_{n\to\infty} \frac{(n-m+1)\cdot \ldots \cdot (n-1)\cdot n}{n^m} = 1. \tag{A.16}$$

Teine piirväärtus avaldises (A.15)

$$\lim_{n\to\infty} \frac{\lambda^m}{m!} = \frac{\lambda^m}{m!}, \tag{A.17}$$

sest piirväärtuse all olev avaldis n-st ei sõltu. Kolmas piirväärtus avaldises (A.15) on

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n. \tag{A.18}$$

Tuletame meelde, et Euleri arvu $e = 2{,}718\ldots$ definitsioon on

$$e = \lim_{n\to\infty} \left(1 + \frac{1}{x}\right)^x. \tag{A.19}$$

Tehes piirväärtuses (A.18) asenduse $x = -n/\lambda$, saame

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = \lim_{n\to\infty} \left(1 + \frac{1}{x}\right)^{-\lambda x} = \lim_{n\to\infty} \left[\left(1 + \frac{1}{x}\right)^x\right]^{-\lambda} = e^{-\lambda}. \tag{A.20}$$

Viimane, neljas piirväärtus avaldises (A.15)

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{-m} = 1. \tag{A.21}$$

Võttes kokku valemid (A.16), (A.17), (A.20) ja (A.21), saame

$$P(X = m) = e^{-\lambda} \frac{\lambda^m}{m!}, \tag{A.22}$$

mis on tuntud Poissoni valemina.

A.4. Normaaljaotuse jaotustiheduse analüüs

Normaaljaotuse jaotustihedus väärtusel x on

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{A.23}$$

kus μ on jaotuse keskväärtus ja σ standardhälve.

1. Mediaan. Analüüsime jaotustiheduse väärtust mõlemal pool keskväärtust kaugusel a, s.t punktides $x = \mu - a$ ja $x = \mu + a$. Analüüsiks piisab, kui vaatame arvu e astendaja lugejat:

$$\begin{aligned} x = \mu - a: &\quad (x-\mu)^2 = (\mu - a - \mu)^2 = (-a)^2 = a^2, \\ x = \mu + a: &\quad (x-\mu)^2 = (\mu + a - \mu)^2 = a^2. \end{aligned}$$

Siit järeldub, et arvu e astendaja on punktides $x = \mu - a$ ja $x = \mu + a$ ühesugune ja järelikult $f(\mu - a) = f(\mu + a)$. See tähendab, et jaotustiheduse graafik on sümmeetriline keskväärtuse μ suhtes. Järelikult asub graafiku keskpunkt punktis μ ning mediaan on võrdne keskväärtusega μ.

2. Mood. Moodi leidmiseks tuleb leida jaotustiheduse maksimumkoht. Selleks võtame avaldisest (A.23) tuletise:

$$f'(x) = \frac{1}{\sigma\sqrt{2\pi}}\left(e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)' = \frac{1}{\sigma\sqrt{2\pi}}\left(-\frac{x-\mu}{\sigma^2}\right)e^{-\frac{(x-\mu)^2}{2\sigma^2}} = -\frac{1}{\sigma^3\sqrt{2\pi}}(x-\mu)e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{A.24}$$

Maksimumkohas peab 1. järku tuletis olema 0. Avaldis (A.24) on 0 siis, kui

$$x - \mu = 0.$$

Järelikult on mood kohas

$$x = \mu.$$

3. Käänukoht. Käänukohas on funktsiooni 2. järku tuletis 0. Leiame 2. järku tuletise:

$$f''(x) = -\frac{1}{\sigma^3\sqrt{2\pi}}\left((x-\mu)e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)'.$$

$f''(x) = 0$, kui sulgudes oleva avaldise tuletis on 0. Leiame selle tuletise, kasutades korrutise tuletise valemit:

$$\begin{aligned} \left((x-\mu)e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)' &= (x-\mu)'e^{-\frac{(x-\mu)^2}{2\sigma^2}} + (x-\mu)\left(e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)' = \\ &= e^{-\frac{(x-\mu)^2}{2\sigma^2}} + (x-\mu)\left(-\frac{x-\mu}{\sigma^2}\right)e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \\ &= \left(1 - \frac{(x-\mu)^2}{\sigma^2}\right)e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{\sigma^2 - (x-\mu)^2}{\sigma^2}e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \end{aligned}$$

Viimane avaldis on 0 siis, kui murru lugejas olev avaldis

$$\sigma^2 - (x-\mu)^2 = 0,$$
$$(\sigma - (x-\mu))\,(\sigma + (x-\mu)) = 0.$$

Siit saame kaks võrrandit, millest leiame kaks x väärtust:

$$\sigma - (x-\mu) = 0 \quad \Rightarrow \quad x = \mu + \sigma,$$
$$\sigma + (x-\mu) = 0 \quad \Rightarrow \quad x = \mu - \sigma.$$

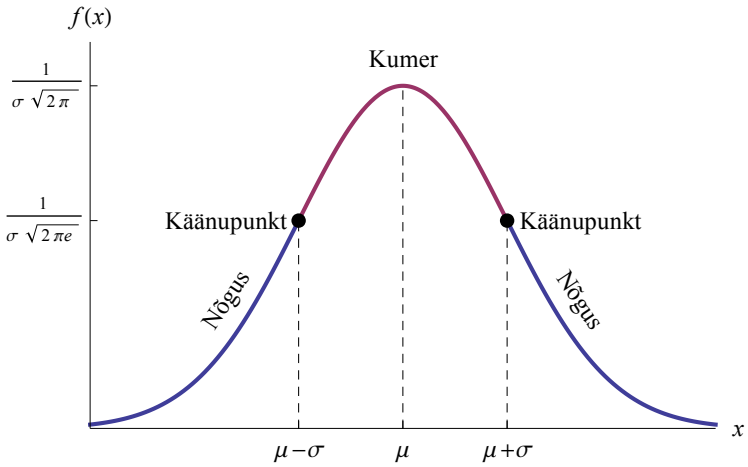
Järelikult on normaaljaotuse jaotustiheduse graafikul kaks käänupunkti kohtades

$$x_1 = \mu + \sigma,$$
$$x_2 = \mu - \sigma \tag{A.25}$$

ehk siis keskväärtusest μ kummalgi pool standardhälbe σ kaugusel.

Käänupunktis, kus kumerus läheb üle nõgususeks, on funktsiooni II tuletis null ja neid punkte on kaks: $\bar{x} - \sigma$ ja $\bar{x} + \sigma$. Nendes punktides jaotustiheduse funktsiooni väärtus (joonis A.1):

$$f(\bar{x} - \sigma) = f(\bar{x} + \sigma) = \frac{1}{\sigma\sqrt{2\pi e}}.$$



Joonis A.1. Normaaljaotuse jaotustiheduse kõvera maksimum ning käänupunktid

A.5. Normaaljaotusest tuletatud jaotused

χ^2-jaotus

Olgu meil juhuslik suurus X, mis allub standardiseeritud normaaljaotusele $N(0,1)$. Juhusliku suuruse

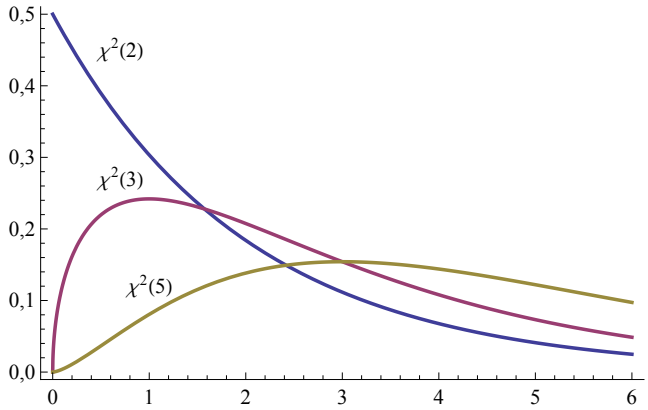
$$Y_1 = X^2 \tag{A.26}$$

jaotust nimetatakse ühe vabadusastmega χ^2-jaotuseks ($\chi^2(1)$-jaotus).

Kui meil on ν sõltumatut juhuslikku suurust $X_1, X_2, \ldots, X_\nu$, mis kõik alluvad standardiseeritud normaaljaotusele, siis suuruse

$$Y_\nu = \sum_{i=1}^{\nu} X_i^2 \tag{A.27}$$

jaotus on ν vabadusastmega χ^2-jaotus ($\chi^2(\nu)$-jaotus). Joonisel A.2 on esitatud χ^2-jaotuse tihedusfunktsiooni graafikud mõningate vabadusastmete korral.



Joonis A.2. $\chi^2(\nu)$-jaotuse tihedusfunktsiooni graafikud mõningate vabadusastmete korral

$\chi^2(\nu)$-jaotust kasutatakse mitmete testide kriitiliste väärtuste leidmisel, samuti multinominaalse tunnuse osakaalude usalduspiiride arvutamisel.

χ-jaotus

Kui juhuslik suurus Y_ν allub $\chi^2(\nu)$ jaotusele, siis juhusliku suuruse

$$Z_\nu = \sqrt{Y_\nu} \tag{A.28}$$

jaotust nimetatakse $\chi(\nu)$-jaotuseks. Selle jaotuse põhiline tähtsus seisneb selles, et selle abil tuletatakse t-jaotus.

t-jaotus

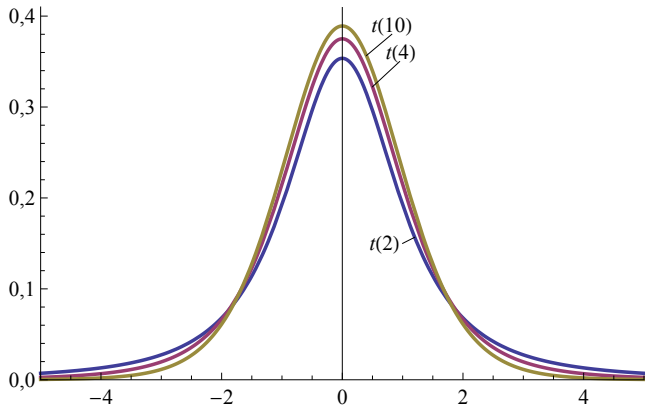
Olgu X ja Y sõltumatud juhuslikud suurused ja allugu nad vastavalt standardiseeritud normaaljaotusele ning χ-jaotusele:

$$X \sim N(0,1), \quad Y \sim \chi(\nu).$$

Juhusliku suuruse

$$Z_\nu = \sqrt{\nu}\frac{X}{Y} \tag{A.29}$$

jaotust nimetatakse t-jaotuseks vabadusastmete arvuga ν.



Joonis A.3. t-jaotuse jaotustiheduse graafikud mõningate vabadusastmete korral

F-jaotus

Olgu juhuslik suurus X_k^* sõltumatute standardiseeritud normaaljaotusele $N(0,1)$ alluvate juhuslike suuruste $X_1, X_2, \ldots, X_k$ ruutude summa

$$X_k^* = \sum_{i=1}^{k} X_i^2. \tag{A.30}$$

Teine juhuslik suurus Y_l^* olgu samuti sõltumatute normaaljaotusele $N(0,1)$ alluvate juhuslike suuruste $Y_1, Y_2, \ldots, Y_l$ ruutude summa

$$Y_l^* = \sum_{i=1}^{l} Y_i^2. \tag{A.31}$$

Juhuslik suurus

$$Z_{k,l} = \frac{X_k^*/k}{Y_l^*/l} \tag{A.32}$$

allub siis F-jaotusele vabadusastmetega k ja l:

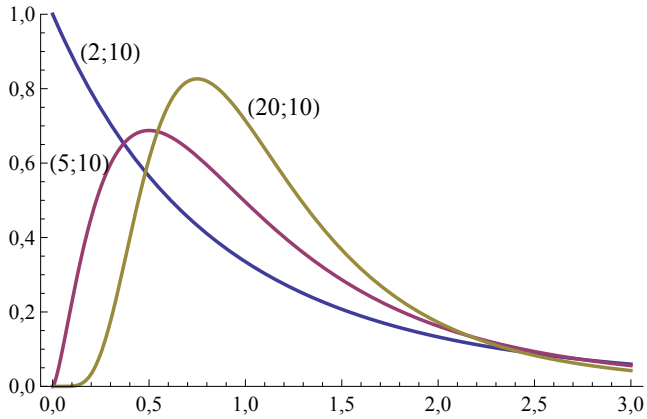
$$Z_{k,l} \sim F(k,l). \tag{A.33}$$

Joonisel A.4 on toodud jaotuste $F(2,10)$, $F(5,10)$ ja $F(20,10)$ tihedusfunktsiooni graafikud. F-jaotust kasutatakse kahe valimi dispersiooni testimisel F-testiga ja dispersioonanalüüsis.

A.6. Vabadusastmete arv

Valikvaatluste puhul kasutatakse paljude valemite korral sellist suurust nagu vabadusastmete arv (*degrees of freedom*). Vabadusastmete arv on lühidalt öeldes sõltumatute muutujate arv[1]. Püüame seda mõistet selgitada mõne näite abil.

[1] Füüsikas tähendab vabadusastmete arv parameetrite arvu, mis määravad ära süsteemi oleku.



Joonis A.4. F-jaotuse tihedusfunktsiooni graafikud mõningate vabadusastmete korral (sulgudes)

1. Kitsendav tingimus ja vabadusastmete arv

Oletame, et meil tuleb kõikide arvude hulgast välja valida neli arvu: x, y, z ja w. Sellisel juhul on meil neli sõltumatut muutujat.

Nüüd kehtestame ühe kitsendava tingimuse: nende arvude summa peab olema 10:

$$x + y + z + w = 10. \tag{A.34}$$

Sellisel juhul saame vabalt valida vaid kolm arvu. See tähendab, et vabadusastmete arv on 3.

Paneme lisaks peale teise kitsendava tingimuse:

$$x + y = 4. \tag{A.35}$$

Nüüd saame vabalt valida vaid kaks arvu, vabadusastmete arv on 2. Järelikult vähendab iga kitsendav tingimus vabadusastmete arvu ühe võrra.

Sama kehtib siis, kui me peame valima n arvu $x_1, x_2, \ldots, x_n$ ning kehtib kitsendav tingimus

$$\sum_{i=1}^{n} x_i = C, \tag{A.36}$$

kus C on konstant. Vabalt saame valida $n-1$ arvu, s.t vabadusastmete arv on $n-1$. Kui me kehtestame k kitsendavat tingimust, siis vabadusastmete arv on $n-k$.

2. Vabadusastmete arv kogumi dispersiooni hindamisel

Olgu meil arvukogum, mille keskväärtus on teada, $\mu = 6$. Soovime hinnata kogumi dispersiooni ja võtame kogumist juhuslikult ühe elemendi. Selle väärtus on 8. Võrdleme seda väärtust kogumi keskväärtusega ja saame kogumi dispersiooni hinnanguks $(8-6)^2 = 4$. See hinnang põhineb n-ö ühel infotükil ja vabadusastmete arv on 1. Võtame kogumist juhuslikult teise elemendi. Selle väärtuseks on 5. Leiame kogumi dispersiooni hinnangu nende kahe elemendi põhjal:

$$\frac{(8-6)^2 + (5-6)^2}{2} = 2{,}5.$$

Selle hinnangu korral on vabadusastmete arv 2.

Kui me aga ei tea kogumi keskväärtust, siis tuleb ka seda valimi põhjal hinnata. Üheelemendilise valimi $\{8\}$ põhjal on kogumi keskväärtuse hinnanguks 8 ja kogumi dispersiooni me hinnata ei saa. Kaheelemendilise valimi $\{8, 5\}$ põhjal on kogumi keskväärtuse hinnanguks valimi keskväärtus:

$$\bar{x} = \frac{8+5}{2} = 6{,}5.$$

Kogumi dispersiooni hindamiseks leiame kaks hinnangut:

$$\begin{aligned} &\text{hinnang 1} \quad (8-6{,}5)^2 = 2{,}25, \\ &\text{hinnang 2} \quad (5-6{,}5)^2 = 2{,}25. \end{aligned}$$

Kas need hinnangud on sõltumatud? Vastus on ei, sest näiteks väärtus 8 mõjutab nii hinnangut 1 kui ka keskväärtuse kaudu hinnangut 2. Kui näiteks esimene väärtus oleks mitte 8, vaid 10, siis valimi keskväärtus oleks 7,5 ja hinnang 2 oleks $(5-7{,}5)^2 = 6{,}26$, mitte 2,25. See tähendab, et vabadusastmete arv on 1. Kui me tahame leida nende kahe hinnangu keskmist, siis peame nende hinnangute summa jagama arvuga 1, mitte arvuga 2. Sest summas

$$(8-6{,}5)^2 + (5-6{,}5)^2$$

on ainult üks sõltumatu liidetav. Selles võib veenduda ka nii, et kui me teame ühe elemendi väärtust 8 ja valimi keskväärtust 6,5, siis saame leida teise elemendi väärtuse, see ei ole sõltumatu.

Kui meil on valimis 12 elementi ja samamoodi peame selle valimi põhjal hindama nii kogumi keskväärtust kui ka dispersiooni, siis dispersiooni hindamisel on vabadusastmete arv $12-1=11$. Kui valimi maht on n, siis kogumi dispersiooni hindamisel on vabadusastmete arv $n-1$.

Matemaatiliselt range tõestus valimi dispersiooni valemi jaoks on järgmises lisas A.7.

A.7. Valimi dispersiooni valemi tuletamine

Olgu meil kogum keskväärtusega μ ja dispersiooniga σ^2. Kasutame valimit $x_1, x_2, \ldots, x_n$ kogumi dispersiooni σ^2 hindamiseks. Olgu kogumi dispersiooni hinnanguks hälvete ruutude aritmeetiline keskmine

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}, \tag{A.37}$$

kus n on valimi maht ja $\bar{x}$ valimi keskmine. Leiame selle hinnangu nihke kogumi dispersiooni σ^2 suhtes. Selleks tuleb leida suuruse (A.37) keskväärtus üle kõikvõimalike valimite:

$$E[S^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[\left(x_i - \frac{1}{n}\sum_{j=1}^{n} x_j\right)^2\right]. \tag{A.38}$$

Siin kasutasime keskväärtuse aditiivsuse omadust (5.18). Sulgudes olevast avaldisest keskväärtuse leidmiseks tuleb summast $\sum_{j=1}^{n} x_j$ eraldada väärtus x_i:

$$\sum_{j=1}^{n} x_j = x_i + \sum_{\substack{j=1 \\ j\neq i}}^{n} x_j. \tag{A.39}$$

Nüüd saame valemis (A.38) sulgudes oleva avaldise ümber kirjutada

$$x_i - \frac{1}{n}\sum_{j=1}^{n} x_j = x_i - \frac{1}{n}x_i - \frac{1}{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} x_j = \frac{n-1}{n}x_i - \frac{1}{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} x_j. \tag{A.40}$$

Kuna meil on vaja võrrelda suuruse S^2 keskväärtust kogumi dispersiooniga σ^2, mille arvutamisel kasutatakse hälbeid kogumi keskväärtusest μ, toome sisse kogumi keskväärtuse μ. Arvestame seda, et

$$\frac{1}{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} (x_j - \mu) = \frac{1}{n}\left(\sum_{\substack{j=1 \\ j\neq i}}^{n} x_j - (n-1)\mu\right) = \frac{1}{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} x_j - \frac{n-1}{n}\mu. \tag{A.41}$$

Järelikult võime avaldisse A.40 tuua kogumi keskväärtuse nii, et liidame ja lahutame (A.41) viimase liikme:

$$\begin{aligned}\frac{n-1}{n}x_i - \frac{1}{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} x_j &= \frac{n-1}{n}x_i - \frac{n-1}{n}\mu + \frac{n-1}{n}\mu - \frac{1}{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} x_j = \\ &= \frac{n-1}{n}(x_i - \mu) - \frac{1}{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} (x_j - \mu)\,.\end{aligned} \tag{A.42}$$

Arvestades valemeid (A.39)–(A.42), võime valemis (A.38) sulgudes ja ruudus oleva avaldise nüüd lahti kirjutada:

$$\begin{aligned}\left(x_i - \frac{1}{n}\sum_{j=1}^{n} x_j\right)^2 &= \left(\frac{n-1}{n}(x_i - \mu) - \frac{1}{n}\sum_{\substack{j=1 \\ j\neq i}}^{n}(x_j - \mu)\right)^2 = \\ &= \frac{(n-1)^2}{n^2}(x_i - \mu)^2 - \frac{2(n-1)}{n^2}\sum_{\substack{j=1 \\ j\neq i}}^{n}(x_i - \mu)(x_j - \mu) + \\ &+ \frac{1}{n^2}\left(\sum_{\substack{j=1 \\ j\neq i}}^{n}(x_j - \mu)\right)^2.\end{aligned} \tag{A.43}$$

Viimase summa ruut on

$$\left(\sum_{\substack{j=1\\j\neq i}}^{n}(x_j-\mu)\right)^2 = \sum_{\substack{j=1\\j\neq i}}^{n}(x_j-\mu)\sum_{\substack{k=1\\k\neq i}}^{n}(x_k-\mu) = \\ = \sum_{\substack{j=1\\j\neq i}}^{n}(x_j-\mu)^2 + \sum_{j\neq i}\sum_{k\neq i,j}(x_j-\mu)(x_k-\mu). \tag{A.44}$$

Pannes (A.44) avaldisse (A.43), saame nüüd leida keskväärtuse (A.38):

$$E[S^2] = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{(n-1)^2}{n^2}E\left[(x_i-\mu)^2\right] - \frac{2(n-1)}{n^2}\sum_{\substack{j=1\\j\neq i}}^{n}E[(x_i-\mu)(x_j-\mu)] + \right. \\ \left. + \frac{1}{n^2}\sum_{\substack{j=1\\j\neq i}}^{n}E\left[(x_j-\mu)^2\right] + \frac{1}{n^2} + \sum_{j\neq i}\sum_{k\neq i,j}E[(x_j-\mu)(x_k-\mu)].\right\}. \tag{A.45}$$

Avaldises (A.45) olevate keskväärtuste leidmiseks arvestame järgmisi võrdusi:

$$E\left[(x_j-\mu)^2\right] = \sigma^2, \tag{A.46}$$

$$E[(x_j-\mu)(x_k-\mu)] = 0 \quad (j\neq k). \tag{A.47}$$

Võrdus (A.46) on kogumi dispersiooni σ^2 definitsioonvalem keskväärtuse kaudu. Võrdus (A.47) aga kehtib sellepärast, et kuna $(x_j-\mu)$ ja $(x_k-\mu)$ on sõltumatud, siis saame kasutada keskväärtuse multiplikatiivsuse omadust (5.19), aga $E[(x_j-\mu)]=0$ ja $E[(x_k-\mu)]=0$.

Arvestades võrdusi (A.46) ja (A.47), saame keskväärtuse jaoks

$$E[S^2] = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(n-1)^2}{n^2}\sigma^2 + \frac{1}{n^2}\sum_{\substack{j=1\\j\neq i}}^{n}\sigma^2\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(n-1)^2}{n^2}\sigma^2 + \frac{1}{n^2}(n-1)\sigma^2\right) = \\ = \frac{1}{n}\cdot n\left(\frac{(n-1)^2}{n^2}\sigma^2 + \frac{1}{n^2}(n-1)\sigma^2\right) = \frac{n-1}{n}\sigma^2. \tag{A.48}$$

Tulemus tähendab seda, et kui me kasutame kogumi dispersiooni hinnanguks suurust (A.37), saame me nihkega hinnangu. Nihketa hinnangu saamiseks tuleb see läbi jagada arvuga $(n-1)/n$. Kogumi dispersiooni nihketa hinnang on

$$s^2 = \frac{n}{n-1}S^2 = \frac{n}{n-1}\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}. \tag{A.49}$$

A.8. Lineaarse mudeli parameetrite tõlgendus

Lineaarse mudeli üldkuju on

$$y = ax + b. \tag{A.50}$$

Vabaliikme b tõlgenduse saame, kui valemis (A.50) võtame x väärtuseks 0:

$$y(0) = a \cdot 0 + b = b.$$

Järelikult, vabaliige b näitab y väärtust, kui $x = 0$.

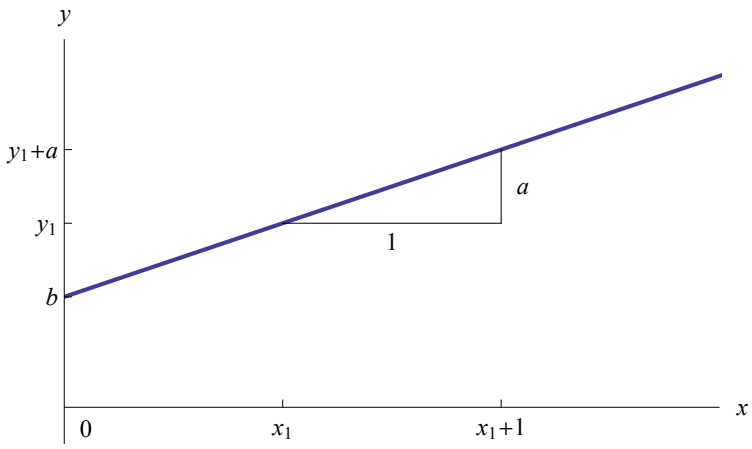
Lineaarliikme kordaja a tõlgenduse saamiseks uurime, kuidas muutub Y, kui X suureneb 1 võrra. Selleks leiame suuruse Y väärtused kahe erineva X väärtuse x_1 ja $x_1 + 1$ korral:

$$\begin{aligned} y_1 &= ax_1 + b, \\ y_2 &= a(x_1 + 1) + b = ax_1 + a + b. \end{aligned}$$

Nüüd leiame suuruse Y väärtuste erinevuse:

$$y_2 - y_1 = (ax_1 + a + b) - (ax_1 + b) = ax_1 + a + b - ax_1 - b = a.$$

Näeme, et kui X suureneb 1 võrra, siis Y muutub a võrra. See ongi kordaja a tõlgendus.



Joonis A.5. Lineaarse mudeli parameetrite tõlgendused

A.9. Lihtsa lineaarse regressioonmudeli parameetrite leidmine vähimruutude meetodil

Vähimruutude meetodi korral tuleb minimeerida hälvete ruutude summat:

$$S(a, b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2 \to \min. \tag{A.51}$$

Siin kasutame summeerimise sümbolit koos summeerimisindeksiga i, sest arvutustes on oluline summa liikmete arv n. Kahe muutuja funktsiooni $S(a,b)$ miinimumkoha leidmiseks võrdsustame osatuletised nulliga ja saame võrrandisüsteemi:

$$\begin{cases} \dfrac{\partial S(a,b)}{\partial a} = 0 \\ \dfrac{\partial S(a,b)}{\partial b} = 0. \end{cases} \tag{A.52}$$

Osatuletiste leidmiseks kirjutame lahti ruudu summas (A.51):

$$\begin{aligned} S(a,b) &= \sum_{i=1}^{n} (y_i - (ax_i + b))^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2 = \\ &= \sum_{i=1}^{n} (y_i^2 + a^2 x_i^2 + b^2 - 2ax_i y_i - 2by_i + 2abx_i). \end{aligned}$$

Nüüd võtame sellest summast osatuletised a ja b järgi. Arvestame, et summa tuletis on liidetavate tuletiste summa:

$$\begin{aligned} \frac{\partial S(a,b)}{\partial a} &= \frac{\partial}{\partial a} \sum_{i=1}^{n} (y_i^2 + a^2 x_i^2 + b^2 - 2ax_i y_i - 2by_i + 2abx_i) = \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial a} (y_i^2 + a^2 x_i^2 + b^2 - 2ax_i y_i - 2by_i + 2abx_i) = \sum_{i=1}^{n} (2ax_i^2 - 2x_i y_i + 2bx_i), \\ \frac{\partial S(a,b)}{\partial b} &= \frac{\partial}{\partial b} \sum_{i=1}^{n} (y_i^2 + a^2 x_i^2 + b^2 - 2ax_i y_i - 2by_i + 2abx_i) = \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial b} (y_i^2 + a^2 x_i^2 + b^2 - 2ax_i y_i - 2by_i + 2abx_i) = \sum_{i=1}^{n} (2b - 2y_i + 2ax_i). \end{aligned}$$

Et saaksime lahendada võrrandisüsteemi (A.52) tundmatute a ja b suhtes, tuleb need tundmatud summa märgi alt välja tuua:

$$\begin{aligned} \sum_{i=1}^{n} (2ax_i^2 - 2x_i y_i + 2bx_i) &= \sum_{i=1}^{n} 2ax_i^2 - \sum_{i=1}^{n} 2x_i y_i + \sum_{i=1}^{n} 2bx_i = \\ &= 2a \sum_{i=1}^{n} x_i^2 - 2 \sum_{i=1}^{n} x_i y_i + 2b \sum_{i=1}^{n} x_i, \\ \sum_{i=1}^{n} (2b - 2y_i + 2ax_i) &= \sum_{i=1}^{n} 2b - \sum_{i=1}^{n} 2y_i + \sum_{i=1}^{n} 2ax_i = 2nb - 2 \sum_{i=1}^{n} y_i + 2a \sum_{i=1}^{n} x_i. \end{aligned}$$

Viimase summa teisendamisel arvestasime seda, et $\sum_{i=1}^{n} b = b + b + \ldots + b$, kus on n ühesugust liidetavat b. Järelikult $\sum_{i=1}^{n} b = nb$.

Saadud avaldiste lihtsustamiseks arvestame, et vastavalt aritmeetilise keskmise valemile $\sum_{i=1}^{n} x_i = n\bar{x}$ ja $\sum_{i=1}^{n} y_i = n\bar{y}$. Nüüd võime osatuletised kirjutada kujul

$$\frac{\partial S(a,b)}{\partial a} = 2a\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i y_i + 2bn\bar{x}, \tag{A.53}$$

$$\frac{\partial S(a,b)}{\partial b} = 2nb - 2n\bar{y} + 2an\bar{x}. \tag{A.54}$$

Kasutades valemeid (A.53), saame võrrandisüsteemi (A.52) kirjutada välja kujul, mis võimaldab selle lahendamist:

$$\begin{cases} 2a\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i y_i + 2bn\bar{x} = 0 \\ 2nb - 2n\bar{y} + 2an\bar{x} = 0. \end{cases} \tag{A.55}$$

Järgneb võrrandsüsteemi lahendamine tundmatute a ja b suhtes. Paneme tähele, et esimese võrrandi võib läbi jagada arvuga 2 ning teise võrrandi avaldisega $2n$:

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i + bn\bar{x} = 0 \\ b - \bar{y} + a\bar{x} = 0. \end{cases}$$

Avaldame teisest võrrandist tundmatu b:

$$b = \bar{y} - a\bar{x}. \tag{A.56}$$

Saadud avaldise paneme esimesse võrrandisse ning avaldame sealt tundmatu a:

$$a\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i + (\bar{y} - a\bar{x})n\bar{x} = 0$$

$$a\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i + n\bar{y}\bar{x} - an\bar{x}^2 = 0$$

$$a\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x},$$

$$a = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}. \tag{A.57}$$

Võrrandisüsteem on lahendatud. Näeme, et lineaarse regressioonmudeli parameetrite leidmiseks tuleb algul välja arvutada aritmeetilised keskmised $\bar{x}$ ja $\bar{y}$ ning summad $\sum_{i=1}^{n} x_i y_i$ ja $\sum_{i=1}^{n} x_i^2$. Seejärel saame valemist (A.57) leida parameetri a ning siis valemist (A.56) parameetri b.

Parameetri a valem esitatakse tihti kujul, kus on kasutatud tunnuste x ja y üksikute väärtuste erinevusi aritmeetilisest keskmisest:

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{A.58}$$

Näitame, et valemid (A.57) ja (A.58) on ekvivalentsed. Selleks vaatame eraldi nende lugejaid ja nimetajaid.

Teisendame valemi (A.58) lugejat:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^{n}(x_i y_i - \bar{x}y_i - \bar{y}x_i + \bar{x}\bar{y}) = \\ &= \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n}\bar{x}y_i - \sum_{i=1}^{n}\bar{y}x_i + \sum_{i=1}^{n}\bar{x}\bar{y} = \\ &= \sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} x_i + n\bar{x}\bar{y} = \\ &= \sum_{i=1}^{n} x_i y_i - \bar{x}n\bar{y} - \bar{y}n\bar{x} + n\bar{x}\bar{y} = \\ &= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}.\end{aligned}$$

Saime valemi (A.57) lugeja. Nüüd teisendame valemi (A.58) nimetajat:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\bar{x} + \sum_{i=1}^{n}\bar{x}^2 = \\ &= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2 = \\ &= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.\end{aligned}$$

Saime valemi (A.57) nimetaja. Kui valemite (A.57) ja (A.58) lugejad ja nimetajad on võrdsed, on need valemid võrdsed.

A.10. Koguhajuvus lineaarse regressioonmudeli korral

Näitame, et lineaarse regressioonmudeli korral kehtib seos

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{A.59}$$

Lähtume valemist (9.28), kus i-nda punkti kõrvalekalle aritmeetilisest keskmisest on jagatud kaheks komponendiks:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i). \tag{A.60}$$

Võtame võrduse (A.60) mõlemad pooled ruutu ja summeerime üle i:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} 2(\hat{y}_i - \bar{y})(y_i - \hat{y}_i). \tag{A.61}$$

Näeme, et võrduse (A.61) paremal pool olevad esimesed kaks summat langevad kokku võrduse (A.59) parema poolega. Järelikult tuleb meil näidata, et viimane liige $\sum_{i=1}^{n} 2(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 0$.

Kuna $\hat{y}_i = ax_i + b$ ja valemist (A.56) lineaarse regressioonmudeli vabaliige $b = \bar{y} - a\bar{x}$, siis

$$\hat{y}_i = ax_i + \bar{y} - a\bar{x}. \tag{A.62}$$

Avaldist (A.62) kasutades asendame $\hat{y}_i$:

$$\begin{aligned}
\sum_{i=1}^{n} 2(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) &= \sum_{i=1}^{n} 2(ax_i + \bar{y} - a\bar{x} - \bar{y})(y_i - ax_i - \bar{y} + a\bar{x}) = \\
&= \sum_{i=1}^{n} 2a(x_i - \bar{x})(y_i - \bar{y} - a(x_i - \bar{x})) = \\
&= \sum_{i=1}^{n} 2a((x_i - \bar{x})(y_i - \bar{y}) - a(x_i - \bar{x})^2) = \\
&= 2a\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - a\sum_{i=1}^{n}(x_i - \bar{x})^2\right).
\end{aligned}$$

Sulgudes oleva teise liikme korral kasutame parameetri a valemit (A.57) ning näitame, et sulgudes olev avaldis võrdub nulliga:

$$\begin{aligned}
&\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - a\sum_{i=1}^{n}(x_i - \bar{x})^2 = \\
&= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0.
\end{aligned}$$

Järelikult avaldises (A.61) olev viimane summa

$$\sum_{i=1}^{n} 2(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 0$$

ja kehtib (A.59).

A.11. Kahe argumenttunnusega lineaarse regressioonmudeli parameetrite hindamine

Kahe argumenttunnuse X_1 ja X_2 korral on i-nda objekti y-koordinaat leitav seosest

$$y_i = a_0 + a_1 x_{1i} + a_2 x_{2i} + u_i. \tag{A.63}$$

Regressioonanalüüsi käigus tuleb leida kolm parameetrit a_0, a_1 ja a_2. Seosest (A.63) näeme, et hälve u_i on

$$u_i = y_i - (a_0 + a_1 x_{1i} + a_2 x_{2i}). \tag{A.64}$$

Minimeerime hälvete ruutude summat, mis sõltub kolmest tundmatust: a_0, a_1 ja a_2:

$$S(a_0, a_1, a_2) = \sum_{i=1}^{n} (y_i - (a_0 + a_1 x_{1i} + a_2 x_{2i}))^2 \to \min. \tag{A.65}$$

Kolme muutuja funktsiooni $S(a_0, a_1, a_2)$ miinimumkoha leidmiseks võrdsustame osatuletised nulliga ja saame kolmest võrrandist koosneva võrrandisüsteemi:

$$\begin{cases} \dfrac{\partial S}{\partial a_0} = 0 \\ \dfrac{\partial S}{\partial a_1} = 0 \\ \dfrac{\partial S}{\partial a_2} = 0. \end{cases}$$

Võttes summast (A.65) osatuletised ja võrdsustades need nulliga, saame kolmest võrrandist koosneva lineaarvõrrandisüsteemi:

$$\begin{cases} a_0 n + a_1 \sum\limits_{i=1}^{n} x_{1i}^2 + a_2 \sum\limits_{i=1}^{n} x_{2i}^2 = \sum\limits_{i=1}^{n} y_i \\ a_0 \sum\limits_{i=1}^{n} x_{1i} + a_1 \sum\limits_{i=1}^{n} x_{1i}^2 + a_2 \sum\limits_{i=1}^{n} x_{1i} x_{2i} = \sum\limits_{i=1}^{n} x_{1i} y_i \\ a_0 \sum\limits_{i=1}^{n} x_{2i} + a_1 \sum\limits_{i=1}^{n} x_{1i} x_{2i} + a_2 \sum\limits_{i=1}^{n} x_{2i}^2 = \sum\limits_{i=1}^{n} x_{2i} y_i. \end{cases} \tag{A.66}$$

Süsteemi (A.66) nimetatakse ka normaalvõrrandite süsteemiks. Selle süsteemi lahendamine tundmatute a_0, a_1 ja a_2 suhtes annab meile valemid parameetrite hinnangute leidmiseks:

$$\begin{cases} a_1 = \dfrac{SP_{yx1} SS_{x2} - SP_{yx2} SP_{x1x2}}{SS_{x1} SS_{x2} - (SP_{x1x2})^2} \\ a_2 = \dfrac{SP_{yx2} SS_{x1} - SP_{yx1} SP_{x1x2}}{SS_{x1} SS_{x2} - (SP_{x1x2})^2} \\ a_0 = \bar{y} - a_1 \bar{x}_1 - a_2 \bar{x}_2, \end{cases} \tag{A.67}$$

kus on kasutatud järgmisi tähistusi (*Sum of Squares SS*, *Sum of Products SP*):

$$\begin{aligned} SS_y &= \sum_{i=1}^{n}(y_i - \bar{y})^2, \\ SS_{x1} &= \sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2, \\ SS_{x2} &= \sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2, \\ SP_{yx1} &= \sum_{i=1}^{n}(y_i - \bar{y})(x_{1i} - \bar{x}_1), \\ SP_{yx2} &= \sum_{i=1}^{n}(y_i - \bar{y})(x_{2i} - \bar{x}_2), \\ SP_{x1x2} &= \sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2). \end{aligned} \tag{A.68}$$

A.12. Korrigeeritud determinatsioonikordaja valemi tuletamine

Lähtume determinatsioonikordaja valemist (9.34):

$$R^2 = 1 - \frac{SSE}{SST}, \tag{A.69}$$

kus jääkhajuvust SSE ja koguhajuvust SST kirjeldavad vahede ruutude summad:

$$SSE = \sum(y_i - \hat{y}_i)^2, \quad SST = \sum(y_i - \bar{y})^2.$$

Valemi (A.69) võib ümber kirjutada kujul

$$R^2 = 1 - \frac{SSE/n}{SST/n}, \tag{A.70}$$

kus n on valimi maht. Koguhajuvust kirjeldava vahede ruutude summa SST läbijagamisel valimi mahuga n saame dispersiooni (vt valem (3.3)):

$$\frac{SST}{n} = \frac{\sum(y_i - \bar{y})^2}{n} = \sigma^2. \tag{A.71}$$

Jääkhajuvust kirjeldava summa SSE jagatist valimi mahuga n võime vaadelda jääkdispersioonina:

$$\frac{SSE}{n} = \frac{\sum(y_i - \hat{y})^2}{n} = \sigma_E^2. \tag{A.72}$$

Dispersioonide (A.71) ja (A.72) abil võime determinatsioonikordaja valemi (A.70) ümber kirjutada kujul

$$R^2 = 1 - \frac{\sigma_E^2}{\sigma^2}. \tag{A.73}$$

Dispersioonid (A.71) ja (A.72) on leitud valimi põhjal. Sellisel kujul annavad need kogumi vastavate dispersioonide jaoks nihkega hinnangud. Et valimi põhjal leitud dispersioon annaks kogumi dispersiooni nihketa hinnangu, tuleb jagada mitte valimi mahuga n, vaid vastava vabadusastmete arvuga. Koguhajuvust kirjeldava dispersiooni σ^2 jaoks on vabadusastmete arv $n-1$ (vt valem (6.4)). Jääkhajuvust kirjeldava dispersiooni σ_E^2 korral on vabadusastmete arv väiksem: iga regressor vähendab vabadusastmete arvu. Kui regressoreid on mudelis k tükki, siis vastav vabadusastmete arv on $n-1-k$.

Asendades determinatsioonikordaja valemis (A.73) nihkega dispersioonide hinnangud nende nihketa hinnangutega, saamegi valemi korrigeeritud determinatsioonikordaja jaoks:

$$R_a^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}. \tag{A.74}$$

Teisendame valemit (A.74) nii, et toome sinna sisse determinatsioonikordaja. Valemist (A.69) saame, et $\frac{SSE}{SST} = 1 - R^2$. Teeme vastava asenduse:

$$R_a^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - \frac{SSE}{SST}\frac{n-1}{n-k-1} = 1 - (1-R^2)\frac{n-1}{n-k-1}.$$

Saime valemi (9.58):

$$R_a^2 = 1 - (1-R^2)\frac{n-1}{n-k-1}, \tag{A.75}$$

kus n on valimi maht ja k regressorite arv mudelis.

A.13. Regressioonmudeli kordaja tõlgendus, kui sõltuv tunnus on logaritmitud

Olgu meil k argumenttunnusega regressioonmudel, kus modelleeritavaks suuruseks on funktsioontunnuse Y naturaallogaritm:

$$\ln y = b + a_1 x_1 + \ldots + a_k x_k + \varepsilon. \tag{A.76}$$

Leiame kordaja a_1 tõlgenduse. Selleks vaatame kaht objekti: A ja B, mis erinevad teineteisest ainult tunnuse X_1 väärtuse poolest, nii et

$$x_{1B} = x_{1A} + 1. \tag{A.77}$$

Ülejäänud tunnuste väärtused olgu neil objektidel ühesugused. Leiame vastavad mudelväärtused:

$$\ln \hat{y}_A = b + a_1 x_{1A} + a_2 x_{2A} + \ldots + a_k x_{kA}, \tag{A.78}$$

$$\ln \hat{y}_B = b + a_1 (x_{1A} + 1) + a_2 x_{2A} + \ldots + a_k x_{kA}. \tag{A.79}$$

Lahutame avaldisest (A.79) avaldise (A.78) ja teisendame:

$$\begin{aligned} \ln \hat{y}_B - \ln \hat{y}_A &= a_1, \\ \ln\left(\frac{\hat{y}_B}{\hat{y}_A}\right) &= a_1, \\ \frac{\hat{y}_B}{\hat{y}_A} &= e^{a_1}. \end{aligned}$$

Leiame funktsioontunnuse väärtuste suhtelise erinevuse:

$$\frac{\hat{y}_B - \hat{y}_A}{\hat{y}_A} = \frac{\hat{y}_B}{\hat{y}_A} - 1 = e^{a_1} - 1. \tag{A.80}$$

Saadud tulemust võime üldistada suvalise argumenttunnuse X_l jaoks. Tähistades sõltuva tunnuse absoluutset muutust $\Delta\hat{y} = \hat{y}_B - \hat{y}_A$, saame suhtelise muutuse jaoks valemi

$$\frac{\Delta\hat{y}}{\hat{y}} = e^{a_l} - 1. \tag{A.81}$$

Järgnevalt kasutame rittaarendust:

$$e^a = \sum_{n=0}^{\infty} \frac{a^n}{n!} = 1 + a + \frac{a^2}{2} + \frac{a^3}{3} + \dots. \tag{A.82}$$

Kui astendaja a on väike, võime piirduda rittaarenduse (A.82) esimese kahe liikmega. Sellisel juhul

$$e^a - 1 \approx 1 + a - 1 = a.$$

Järelikult, väikese kordaja a_l korral on see ligikaudu võrdne funktsioontunnuse suhtelise muuduga argumenttunnuse X_l ühikulisel muutumisel:

$$\frac{\Delta\hat{y}}{\hat{y}} \approx a_l. \tag{A.83}$$

A.14. Õppimiskõver

Kui mingit tegevust sooritada korduvalt, siis lüheneb selle sooritamiseks kulunud aeg. Uuringud on näidanud, et iga kord, kui soorituste arv kahekordistub, lüheneb ühe soorituse jaoks kulunud aeg kindla protsendi võrra. Sellise nähtuse modelleerimiseks kasutatakse õppimiskõverat, mida mõnikord nimetatakse ka kogemuskõveraks.

Olgu esimesel korral tegevuse sooritamiseks kulunud aeg t_1 ja teisel korral kulub aega 15% vähem, siis $t_2 = (1 - 0{,}15)t_1 = 0{,}85t_1$. Kui sama tegevust sooritatakse neljandat korda, kulub aega jällegi 15% vähem: $t_4 = (1 - 0{,}15)t_2 = 0{,}85t_2$.

Tuletame õppimiskõvera mudeli üldjuhul, kui soorituste arvu n kahekordistumisel on aja t vähenemise määr r:

$$\begin{aligned} n &= 1 && t_1, \\ n &= 2 && t_2 = (1-r)t_1, \\ n &= 4 && t_4 = (1-r)t_2 = (1-r)^2 t_1, \\ n &= 8 && t_8 = (1-r)t_4 = (1-r)^3 t_1. \end{aligned}$$

Logaritmi kasutades saame avaldada iga $n > 1$ korral avaldise $(1-r)$ astmenäitaja n kaudu:

$$\begin{aligned} n &= 2 && t_2 = (1-r)t_1 && 1 = \log_2 2, \\ n &= 4 && t_4 = (1-r)^2 t_1 && 2 = \log_2 4, \\ n &= 8 && t_8 = (1-r)^3 t_1 && 3 = \log_2 8. \end{aligned}$$

Näeme, et kui soorituste arv on n, siis vahe $(1-r)$ astmenäitaja on $\log_2 n$. Nüüd võime kirjutada

$$t_n = (1-r)^{\log_2 n} t_1, \tag{A.84}$$

kus t_1 on aeg esmakordsel sooritusel, n soorituste arv, t_n sooritamiseks kulunud aeg n-ndal korral ja r aja vähenemise määr soorituste arvu kahekordistumisel.

Sellisel kujul õppimiskõverat tavaliselt ei esitada. Levinum on kuju

$$t_n = t_1 n^b, \tag{A.85}$$

kus parameeter $b < 0$. Leiame, kuidas on selle mudeli parameeter b seotud aja vähenemise määraga r valemis (A.84). Selleks paneme avaldiste (A.84) ning (A.85) paremad pooled võrduma ning avaldame b:

$$\begin{aligned} t_1 n^b &= (1-r)^{\log_2 n} t_1, \\ n^b &= (1-r)^{\log_2 n}, \\ b &= \log_n (1-r)^{\log_2 n}, \\ b &= \log_2 n \cdot \log_n (1-r). \end{aligned}$$

Kasutame üleminekut logaritmi ühelt aluselt teisele ja läheme teises teguris üle logaritmile alusel 2:

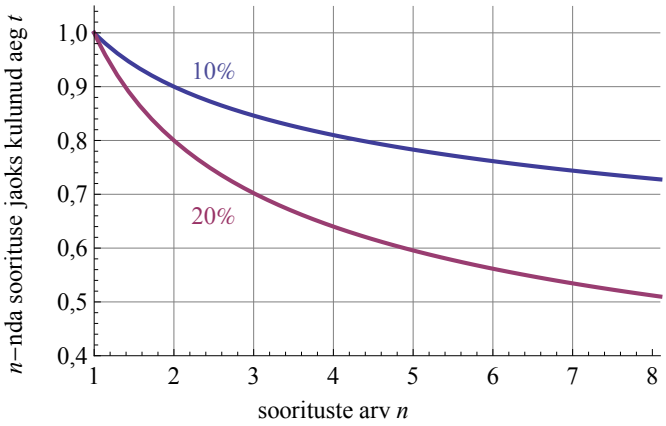
$$\log_n (1-r) = \frac{\log_2 (1-r)}{\log_2 n}.$$

Saame järgmise seose:

$$b = \log_2 (1-r). \tag{A.86}$$

Joonisel A.6 on toodud kaks erineva vähenemise määraga õppimiskõverat.

Tootmises on ajakulu otseselt seotud tööjõukuluga ning õppimiskõver väljendab ühiku tööjõukulu vähenemist tootmismahu suurenemisel.



Joonis A.6. Õppimiskõverad, kui tegevuse sooritamiseks kulunud aja vähenemise määr on 10% ja 20% ning esimene kord kulus aega 1 ajaühik

Lisa B

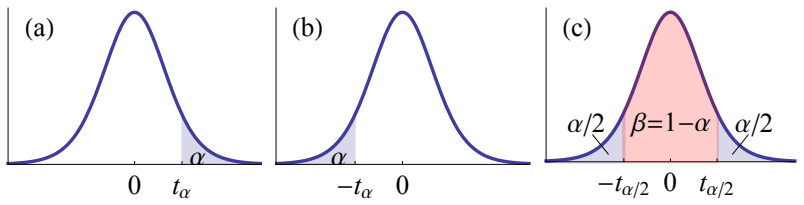
Tabelid

B.1. t-jaotuse täiendkvantiilid

t-jaotuse täiendkvantiile $t_\alpha(\nu)$ kasutatakse t-statistiku kriitilise väärtuse leidmisel t-testi korral ja keskväärtuse usaldusvahemiku leidmisel, kui valim on väike. α-täiendkvantiil t_α on määratud $t(\nu)$-jaotusest nii, et $P(X > t_\alpha) = \alpha$ (vt joonis (a)). Suurus ν on t-jaotuse vabadusastmete arv.

- Ühepoolse hüpoteesi korral on olulisuse nivool α t-testi kriitiliseks väärtuseks $t_\alpha(\nu)$ (joonised (a) ja (b)).
- Kahepoolse hüpoteesi korral on olulisuse nivool α t-testi kriitiliseks väärtuseks $t_{\alpha/2}(\nu)$ (vt joonis (c)).
- Usaldatavusele β vastava usaldusvahemiku leidmisel on tõenäosuskordajaks $t_{\alpha/2}(\nu)$, kus $\alpha = 1 - \beta$ (joonis (c)). Kui näiteks valimi maht $n = 10$, siis vabadusastmete arv $\nu = n - 1 = 9$ ja usaldatavusele $\beta = 0{,}95$ vastav tõenäosuskordaja on $t_{0{,}025}(9) = 2{,}26$.

Tabelarvutuses saab t-jaotuse kvantiile ja täiendkvantiile leida funktsiooniga T.INV(p;ν), mis leiab vabadusastmete arvuga ν t-jaotuse p-kvantiili. α-täiendkvantiil on sama, mis $1 - \alpha$ kvantiil ning α-täiendkvantiili leidmiseks peab T.INV esimene argument olema $p = 1 - \alpha$: T.INV($1 - \alpha$; ν).

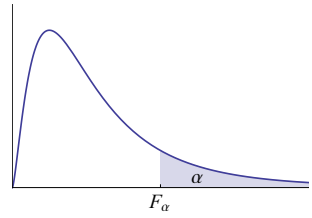


Tabel B.1. Täiendkvantiilid $t_\alpha(\nu)$

ν \ α	0,25	0,2	0,1	0,05	0,025	0,02	0,01	0,005
1	1,00	1,376	3,08	6,31	12,71	15,89	31,82	63,66
2	0,816	1,061	1,89	2,92	4,30	4,85	6,96	9,92
3	0,765	0,978	1,64	2,35	3,18	3,48	4,54	5,84
4	0,741	0,941	1,53	2,13	2,78	3,00	3,75	4,60
5	0,727	0,920	1,48	2,02	2,57	2,76	3,36	4,03
6	0,718	0,906	1,44	1,94	2,45	2,61	3,14	3,71
7	0,711	0,896	1,41	1,89	2,36	2,52	3,00	3,50
8	0,706	0,889	1,40	1,86	2,31	2,45	2,90	3,36
9	0,703	0,883	1,38	1,83	2,26	2,40	2,82	3,25
10	0,700	0,879	1,37	1,81	2,23	2,36	2,76	3,17
50	0,679	0,849	1,30	1,68	2,01	2,11	2,40	2,68
100	0,677	0,845	1,29	1,66	1,98	2,08	2,36	2,63
500	0,675	0,842	1,28	1,65	1,96	2,06	2,33	2,59

B.2. F-jaotuse täiendkvantiilid

F-jaotuse täiendkvantiilid $F_\alpha(\nu_1, \nu_2)$ on F-statistiku kriitiliseks väärtuseks kogumite dispersioonide võrdlemisel F-testiga ja dispersioonanalüüsi ANOVA korral. α-täiendkvantiil on määratud $F(\nu_1, \nu_2)$-jaotusest nii, et $P(X > F_\alpha) = \alpha$. Vabadusastmete arv ν_1 vastab F-statistiku lugejale ning ν_2 nimetajale.



Tabelarvutuses on F-jaotuse täiendkvantiilide leidmiseks funktsioon F.INV.RT(α;ν_1;ν_2).

Tabel B.2. $F_\alpha(\nu_1, \nu_2)$, kui $\alpha = 0{,}1$

ν_2 \ ν_1	1	2	3	4	5	10	15	20	50	100
1	39,86	49,50	53,59	55,83	57,24	60,19	61,22	61,74	62,69	63,01
2	8,53	9,00	9,16	9,24	9,29	9,39	9,42	9,44	9,47	9,48
3	5,54	5,46	5,39	5,34	5,31	5,23	5,20	5,18	5,15	5,14
4	4,54	4,32	4,19	4,11	4,05	3,92	3,87	3,84	3,80	3,78
5	4,06	3,78	3,62	3,52	3,45	3,30	3,24	3,21	3,15	3,13
10	3,29	2,92	2,73	2,61	2,52	2,32	2,24	2,20	2,12	2,09
15	3,07	2,70	2,49	2,36	2,27	2,06	1,97	1,92	1,83	1,79
20	2,97	2,59	2,38	2,25	2,16	1,94	1,84	1,79	1,69	1,65
50	2,81	2,41	2,20	2,06	1,97	1,73	1,63	1,57	1,44	1,39
100	2,76	2,36	2,14	2,00	1,91	1,66	1,56	1,49	1,35	1,29

Tabel B.3. $F_\alpha(\nu_1, \nu_2)$, kui $\alpha = 0{,}05$

ν_2 \ ν_1	1	2	3	4	5	10	15	20	50	100
1	161	200	216	225	230	242	246	248	252	253
2	18,51	19,00	19,16	19,25	19,40	19,43	19,45	19,48	19,49	19,49
3	10,13	9,55	9,28	9,12	8,79	8,70	8,66	8,58	8,55	8,54
4	7,71	6,94	6,59	6,39	6,26	5,96	5,86	5,80	5,70	5,66
5	6,61	5,79	5,41	5,19	5,05	4,74	4,62	4,56	4,44	4,41
10	4,96	4,10	3,71	3,48	3,33	2,98	2,85	2,77	2,64	2,59
15	4,54	3,68	3,29	3,06	2,90	2,54	2,40	2,33	2,18	2,12
20	4,35	3,49	3,10	2,87	2,71	2,35	2,20	2,12	1,97	1,91
50	4,03	3,18	2,79	2,56	2,40	2,03	1,87	1,78	1,60	1,52
100	3,94	3,09	2,70	2,46	2,31	1,93	1,77	1,68	1,48	1,39

Tabel B.4. $F_\alpha(\nu_1, \nu_2)$, kui $\alpha = 0{,}01$

ν_2 \ ν_1	1	2	3	4	5	10	15	20	50	100
1	4052	5000	5403	5625	5764	6056	6157	6209	6303	6334
2	98,50	99,00	99,17	99,25	99,30	99,40	99,43	99,45	99,48	99,49
3	34,12	30,82	29,46	28,71	28,24	27,23	26,87	26,69	26,35	26,24
4	21,20	18,00	16,69	15,98	15,52	14,55	14,20	14,02	13,69	13,58
5	16,26	13,27	12,06	11,39	10,97	10,05	9,72	9,55	9,24	9,13
10	10,04	7,56	6,55	5,99	5,64	4,85	4,56	4,41	4,12	4,01
15	8,68	6,36	5,42	4,89	4,56	3,80	3,52	3,37	3,08	2,98
20	8,10	5,85	4,94	4,43	4,10	3,37	3,09	2,94	2,64	2,54
50	7,17	5,06	4,20	3,72	3,41	2,70	2,42	2,27	1,95	1,82
100	6,90	4,82	3,98	3,51	3,21	2,50	2,22	2,07	1,74	1,60

B.3. Märgitesti kriitilised väärtused

Märgitesti kriitilised väärtused leitakse binoomjaotusest.

Tabelis B.5 on toodud vasakpoolsed kriitilised väärtused n_{krv} ühepoolse testi korral. n on korrigeeritud valimi maht ja α olulisuse tõenäosus ühepoolse testi korral. Kahepoolse testi korral tuleb kriitiline väärtus valida veerust, mille päises on $\alpha/2$. Näiteks kui kahepoolse testi korral on olulisuse nivoo 0,05, siis vasakpoolne kriitiline väärtus tuleb võtta veerust 0,025. Parempoolne kriitiline väärtus leitakse valemist

$$n_{krp} = n - n_{krv},$$

kus n on korrigeeritud valimi maht.

Tabelarvutuses kasutatakse märgitesti kriitiliste väärtust leidmiseks funktsiooni BINOM.INV(n;0,5;α).

Tabel B.5. Märgitesti vasakpoolsed kriitilised väärtused

n \ α	0,01	0,025	0,05	0,1	n \ α	0,01	0,025	0,05	0,1
5	0	0	1	1	23	6	7	8	8
6	0	1	1	1	24	6	7	8	9
7	1	1	1	2	25	7	8	8	9
8	1	1	2	2	26	7	8	9	10
9	1	2	2	3	27	8	8	9	10
10	1	2	2	3	28	8	9	10	11
11	2	2	3	3	29	8	9	10	11
12	2	3	3	4	30	9	10	11	11
13	2	3	4	4	31	9	10	11	12
14	3	3	4	5	32	9	10	11	12
15	3	4	4	5	33	10	11	12	13
16	3	4	5	5	34	10	11	12	13
17	4	5	5	6	35	11	12	13	14
18	4	5	6	6	36	11	12	13	14
19	5	5	6	7	37	11	13	14	15
20	5	6	6	7	38	12	13	14	15
21	5	6	7	8	39	12	13	14	16
22	6	6	7	8	40	13	14	15	16

B.4. χ^2-jaotuse täiendkvantiilid

χ^2-testi kriitiline väärtus olulisuse nivool α on χ^2-jaotuse täiendkvantiil $\chi^2_\alpha(\nu)$. Vabadusastmete arv ν leitakse teoreetilise jaotuse sobivuse testimisel valemist (7.63) ning kahe tunnuse vahelise seose testimisel valemist (7.70). Kui teststatistik $\chi^2 > \chi^2_\alpha(\nu)$, on nullhüpotees ümber lükatud.

Tabelarvutuses leitakse need täiendkvantiilid funktsiooni CHISQ.INV.RT abil, kus *Probability* on α ja *Deg_freedom* vabadusastmete arv ν.

Tabel B.6. χ^2-jaotuse täiendkvantiilid

ν \ α	0,01	0,05	0,1	ν \ α	0,01	0,05	0,1
1	6,63	3,84	2,71	15	30,58	25,00	22,31
2	9,21	5,99	4,61	16	32,00	26,30	23,54
3	11,34	7,81	6,25	17	33,41	27,59	24,77
4	13,28	9,49	7,78	18	34,81	28,87	25,99
5	15,09	11,07	9,24	19	36,19	30,14	27,20
6	16,81	12,59	10,64	20	37,57	31,41	28,41
7	18,48	14,07	12,02	30	50,89	43,77	40,26
8	20,09	15,51	13,36	40	63,69	55,76	51,81
9	21,67	16,92	14,68	50	76,15	67,50	63,17
10	23,21	18,31	15,99	60	88,38	79,08	74,40
11	24,72	19,68	17,28	70	100,43	90,53	85,53
12	26,22	21,03	18,55	80	112,33	101,88	96,58
13	27,69	22,36	19,81	90	124,12	113,15	107,57
14	29,14	23,68	21,06	100	135,81	124,34	118,50

B.5. Lineaarse korrelatsioonikordaja kriitilised väärtused

Kriitilise korrelatsioonikordaja valem saadakse valemist (8.10), avaldades sealt korrelatsioonikordaja r:

$$r_{kr} = \frac{1}{\sqrt{1 + \dfrac{\nu}{\left(t_{\alpha/2}(\nu)\right)^2}}}, \tag{B.1}$$

Erineva vabadusastmete arvu $\nu = n-2$ ja olulisuse nivoo α korral saab leida t-jaotuse täiendkvantiilid $t_{\alpha/2}(\nu)$ ning seejärel valemist (B.1) korrelatsioonikordaja kriitilised väärtused. Tabeli esimeses veerus on valimi maht n.

Korrelatsioonikordaja r absoluutväärtust tuleb võrrelda kriitilise väärtusega. Kui

$|r| \le r_{kr}$, võtta vastu H_0, korrelatsioon puudub;
$|r| > r_{kr}$, võtta vastu H_1, korrelatsioon esineb.

Tabel B.7. Lineaarse korrelatsioonikordaja kriitilised väärtused r_{kr}

n \ α	0,1	0,05	0,01	n \ α	0,1	0,05	0,01
5	0,805	0,878	0,959	20	0,378	0,444	0,561
6	0,729	0,811	0,917	25	0,337	0,396	0,505
7	0,669	0,754	0,875	30	0,306	0,361	0,463
8	0,621	0,707	0,834	40	0,264	0,312	0,403
9	0,582	0,666	0,798	50	0,235	0,279	0,361
10	0,549	0,632	0,765	60	0,214	0,254	0,330
11	0,521	0,602	0,735	70	0,198	0,235	0,306
12	0,497	0,576	0,708	80	0,185	0,220	0,286
13	0,476	0,553	0,684	90	0,174	0,207	0,270
14	0,458	0,532	0,661	100	0,165	0,197	0,256
15	0,441	0,514	0,641	200	0,117	0,139	0,182
16	0,426	0,497	0,623	300	0,095	0,113	0,149
17	0,412	0,482	0,606	400	0,082	0,098	0,129
18	0,400	0,468	0,590	500	0,074	0,088	0,115
19	0,389	0,456	0,575	1000	0,052	0,062	0,081

Lisa C

Juhendeid tabelarvutuse kasutamiseks

C.1. Arvude suurusjärk tabelarvutuses

Kummas veerus on arvud paremini loetavad?

250 000 000 000	$2{,}5 \cdot 10^{11}$
38 900 000 000 000 000	$3{,}89 \cdot 10^{16}$
0,00000017	$1{,}7 \cdot 10^{-7}$
0,0000000000059	$5{,}9 \cdot 10^{-12}$

Väga suurte ja väga väikeste arvude esitamisel on mugav kasutada **arvu standardkuju**, mis on arvu üleskirjutus korrutisena

$$a \cdot 10^b,$$

kus $1 \leq a < 10$ on arvu tüvi ja b täisarvuline astendaja. Arvu 10 astet kasutatakse seepärast, et see on kümnendsüsteemi alus.

Tabelarvutuses ei ole kümne astmenäitaja kuvamine astmenäitaja asukohas tehniliselt võimalik. Seepärast kasutatakse kümne astmenäitaja eraldamiseks arvu tüvest tähte E, mis tuleb ingliskeelsest sõnast *exponent* (aste). Arvude vormindamine sellisel moel on *Scientific format*.

Scientific format tabelarvutuses	Arvu standardkuju
2,5E+11	$2{,}5 \cdot 10^{11}$
3,89E+16	$3{,}89 \cdot 10^{16}$
1,7E−7	$1{,}7 \cdot 10^{-7}$
5,9E−12	$5{,}9 \cdot 10^{-12}$

Tabelarvutuse aruannete tõlgendamisel ja analüüsi ülevaadete kirjutamisel on vaja kirjutada sealt välja vajalikke arve. Kasutada tuleb siis matemaatiliselt korrektset standardkuju.

C.2. Programmi Excel analüüsivahendite komplekt *Data Analysis*

Andmeanalüüsi komplekti *Data Analysis* kasutamiseks peab see olema aktiveeritud. Aktiveerimine käib lisandmoodulite (*Add-Ins*) haldamise valikus, lisada tuleb *Analysis Toolpak*. Ekraanivideot õpetusega, kuidas lisada andmeanalüüsi vahendit, võib vaadata õpiku autori kodulehel[1]. Kui andmeanalüüsi komplekt on aktiveeritud, siis asub nupureal „Andmed“ („*Data*“) nupp „*Data Analysis*“. Sellega kutsutakse välja andmeanalüüsi vahendite komplekt ja valitakse sealt sobiv vahend. Analüüsikomplekt sisaldab järgmisi vahendeid:

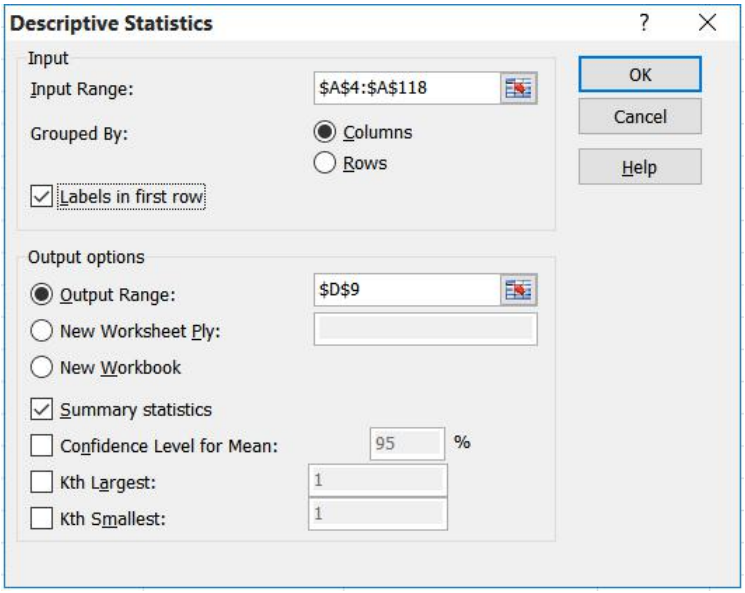
- ANOVA ehk dispersioonanalüüs:
  - ühefaktoriline dispersioonanalüüs *Anova: Single Factor*;
  - kahefaktoriline dispersioonanalüüs, kordumistega *Anova: Two-Factor With Replication*;
  - kahefaktoriline dispersioonanalüüs, kordumisteta *Anova: Two-Factor Without Replication*;
- korrelatsioonimaatriksi arvutamine *Correlation*;
- kovariatsioonimaatriksi arvutamine *Covariation*;
- kirjeldava statistika suurused *Descriptive Statistics*;
- eksponentsiaalne silumine *Exponential Smoothing*;
- dispersioonide võrdlemine F-testi abil *F-Test: Two-Sample for Variances*;
- harmooniline ehk Fourier' analüüs *Fourier Analysis*;
- sagedustabeli ja histogrammi koostamine *Histogram*;
- libisev keskmine *Moving Average*;
- juhuslike arvude genereerimine *Random Number Generation*;
- astaku ja protsentiilide leidmine *Rank and Percentile*;
- regressioonanalüüs *Regression*;
- juhuvalimite moodustamine *Sampling*;
- t-test:
  - sõltuvad valimid *t-Test: Paired Two Sample for Means*;
  - sõltumatud valimid, võrdse dispersiooniga *t-Test: Two-Sample Assuming Equal Variances*;
  - sõltumatud valimid, erineva dispersiooniga *t-Test: Two-Sample Assuming Unequal Variances*.

C.3. Kirjeldava statistika näitajad programmi Excel vahendiga *Descriptive Statistics*

Andmeanalüüsi vahendi *Descriptive Statistics* kasutamiseks peab Excelis olema aktiveeritud andmeanalüüsi komplekt *Data Analysis* (vt lisa C.2). Nupureal „Andmed“ („*Data*“) oleva nupuga „*Data Analysis*“ kutsutakse välja andmeanalüüsi vahendite

[1] `http://www.sauga.pri.ee/statistika_excelis/`

komplekt ja valitakse sealt „*Desciptive Statistics*“. Näidatakse, millistes lahtrites on andmed (*Input Range*), kas viidatud lahtrite esimene rida sisaldab veeru pealkirja (*Labels in First Row*) ning kuhu paigutatakse väljund (*Output options*). Tuleb teha ka valik *Summary statistics*.



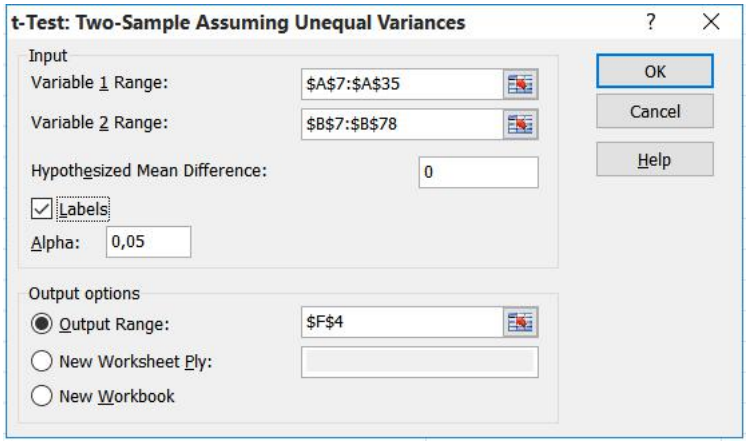
Tulemusena väljastatakse tabel erinevate kirjeldava statistika näitajatega. Eeldatakse, et tegemist on valimiga ning arvutamiseks kasutatakse valimi statistilisi näitajaid (valimi dispersioon, valimi standardhälve). Võib leida ka mitme erineva tunnuse näitajad korraga. Selleks peavad tunnuste väärtused olema kõrvuti veergudes ning *Input Range* sisestamisel viidatakse piirkonnale, kus asuvad kõikide tunnuste väärtused.

Näitena on toodud alapeatükis 3.7 joonisel 3.8 esitatud sõiduauto Audi erinevate mudelite hindade kirjeldav statistika. Lisatud on terminite eestikeelsed vasted. Vastavad andmed on failis N03Varieerumine lehel J3.8.

	Hind, tuh EUR	
Aritmeetiline keskmine	Mean	47,72
Standardviga	Standard Error	2,86
Mediaan	Median	39,93
Mood	Mode	34,05
Valimi standardhälve	Standard Deviation	30,6
Valimi dispersioon	Sample Variance	935,7
Püstakuse kordaja	Kurtosis	5,63
Asümmeetria kordaja	Skewness	2,19
Variatsioonamplituud	Range	158,82
Miinimum	Minimum	15,88
Maksimum	Maximum	174,7
Summa	Sum	5440,26
Maht	Count	114

C.4. Programmi Excel analüüsivahend *t-Test: Two-Sample Assuming Unequal Variances*

Andmeanalüüsi vahendi *t-Test: Two-Sample Assuming Unequal Variances* kasutamiseks peab Excelis olema aktiveeritud andmeanalüüsi komplekt *Data Analysis* (vt lisa C.2). Nupureal „Andmed“ („*Data*“) oleva nupuga „*Data Analysis*“ kutsutakse välja andmeanalüüsi vahendite komplekt ja valitakse sealt „*t-Test: Two-Sample Assuming Equal Variances*“. Näidatakse, millistes lahtrites on kummagi valimi andmed (*Variable 1 Range* ja *Variable 2 Range*), nullhüpoteesiga püstitatud keskväärtuste erinevus (*Hypothesized Mean Difference*, kui soovime keskväärtuste võrdsust kontrollida või ümber lükata, siis 0), kas lahtrite esimene rida sisaldab veeru pealkirja (*Labels*), olulisuse nivoo *Alpha* ning kuhu paigutatakse väljund (*Output options*).



Tulemusena väljastatakse tabel *t*-testi tulemustega.

t-Test: Two-Sample Assuming Unequal Variances

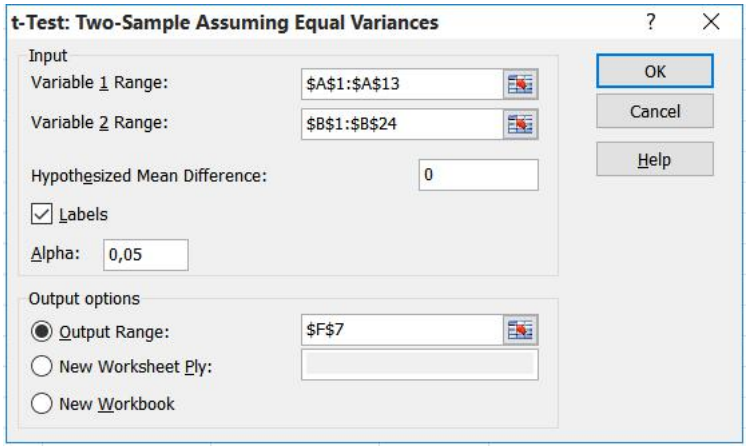
		Mehed	*Naised*
Valimi keskmine	Mean	9337,23	6757,73
Valimi dispersioon	Variance	43292030	12846605
Valimi maht	Observations	28	71
Nullhüpoteesiga püstitatud erinevus	Hypothesized Mean Difference	0	
Vabadusastmete arv	df	34	
t-statistik	t Stat	1,963	
Olulisuse tõenäosus ühepoolse hüpoteesi korral	P(T<=t) one-tail	0,029	
Parempoolne kriitiline väärtus ühepoolse hüpoteesi korral	t Critical one-tail	1,691	
Olulisuse tõenäosus kahepoolse hüpoteesi korral	P(T<=t) two-tail	0,058	
Parempoolne kriitiline väärtus kahepoolse hüpoteesi korral	t Critical two-tail	2,032	

t-statistik leitakse valemist (7.21) ja vabadusastmete arv valemist (7.22). Tabelis on toodud näites 7.11 läbiviidud t-testi tulemused (vt ka faili N07Hüpoteesid). Lisatud on eestikeelsed seletused.

Kui püstitatud hüpotees oli ühepoolne, kasutatakse ühepoolset olulisuse tõenäosust või kriitilist väärtust. Kahepoolse hüpoteesi korral kasutatakse vastavalt kahepoolseid väärtusi. Vasakpoolsed kriitilised väärtused nii ühepoolse kui ka kahepoolse hüpoteesi korral on parempoolsete vastandarvud.

C.5. Programmi Excel analüüsivahend *t-Test: Two-Sample Assuming Equal Variances*

Andmeanalüüsi vahendi *t-Test: Two-Sample Assuming Equal Variances* kasutamiseks peab Excelis olema aktiveeritud andmeanalüüsi komplekt *Data Analysis* (vt lisa C.2). Nupureal „Andmed“ („*Data*“) oleva nupuga „*Data Analysis*“ kutsutakse välja andmeanalüüsi vahendite komplekt ja valitakse sealt „*t-Test: Two-Sample Assuming Equal Variances*“. Näidatakse, millistes lahtrites on kummagi valimi andmed (*Variable 1 Range* ja *Variable 2 Range*), nullhüpoteesiga püstitatud keskväärtuste erinevus (*Hypothesized Mean Difference*, kui soovime keskväärtuste võrdsust kontrollida või ümber lükata, siis 0), kas viidatud lahtrite esimene rida sisaldab veeru pealkirja (*Labels*), olulisuse nivoo *Alpha* ning kuhu paigutatakse väljund (*Output options*).



Tulemusena väljastatakse tabel t-testi tulemustega. Ühendatud dispersioon leitakse valemist (7.23), t-statistik valemist (7.24) ja vabadusastmete arv valemist (7.25). Tabelis on toodud ülesande A.7.5 a) osa t-testi tulemused (andmed failis ÜL07Hüpoteesid). Lisatud on eestikeelsed seletused.

t-Test: Two-Sample Assuming Equal Variances

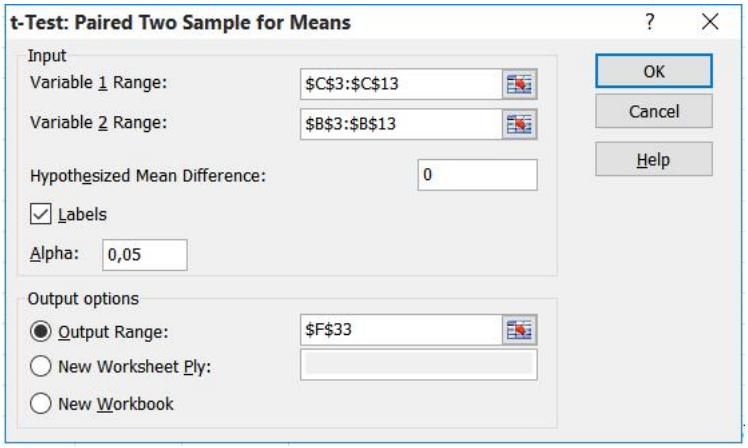
		Lennuki-tööstus	*Masina-tööstus*
Valimi keskmine	Mean	309,24	135,53
Valimi dispersioon	Variance	127438	57252
Valimi maht	Observations	12	23
Ühendatud dispersioon	Pooled Variance	80647	
Nullhüpoteesile vastav erinevus	Hypothesized Mean Difference	0	
Vabadusastmete arv	df	33	
t-statistik	t Stat	1,718	
Olulisuse tõenäosus ühepoolse hüpoteesi korral	P(T<=t) one-tail	0,048	
Parempoolne kriitiline väärtus ühepoolse hüpoteesi korral	t Critical one-tail	1,308	
Olulisuse tõenäosus kahepoolse hüpoteesi korral	P(T<=t) two-tail	0,095	
Parempoolne kriitiline väärtus kahepoolse hüpoteesi korral	t Critical two-tail	1,692	

Kui püstitatud hüpotees oli ühepoolne, kasutatakse ühepoolset olulisuse tõenäosust või kriitilist väärtust. Kahepoolse hüpoteesi korral kasutatakse vastavalt kahepoolseid väärtusi. Vasakpoolsed kriitilised väärtused nii ühepoolse kui ka kahepoolse hüpoteesi korral on parempoolsete vastandarvud.

C.6. Programmi Excel analüüsivahend *t-Test: Paired Two Sample for Means*

Andmeanalüüsi vahendi *t-Test: Paired Two Sample for Means* kasutamiseks peab Excelis olema aktiveeritud andmeanalüüsi komplekt *Data Analysis* (vt lisa C.2). Nupureal „Andmed“ („*Data*“) oleva nupuga „*Data Analysis*“ kutsutakse välja andmeanalüüsi vahendite komplekt ja valitakse sealt „*t-Test: Paired Two Sample for Means*“. Näidatakse, millistes lahtrites on kummagi valimi andmed (*Variable 1 Range* ja *Variable 2 Range*), nullhüpoteesiga püstitatud keskväärtuste erinevus (*Hypothesized Mean Difference*, kui soovime keskväärtuste võrdsust kontrollida või ümber lükata, siis 0), kas viidatud lahtrite esimene rida sisaldab veeru pealkirja (*Labels*), olulisuse nivoo *Alpha* ning kuhu paigutatakse väljund (*Output options*).

Väljastatakse tabel t-testi tulemustega. t-statistik leitakse valemist (7.37) ja vabadusastmete arv valemist (7.38). Tabelis on toodud näites 7.12 läbiviidud t-testi tulemused (vt ka faili N07Hüpoteesid). Lisatud on eestikeelsed seletused.



t-Test: Paired Two Sample for Means

		2009	*2008*
Valimi keskmine	Mean	115,66	135,13
Valimi dispersioon	Variance	7090,6	11538,7
Valimi maht	Observations	10	10
Pearsoni korrelatsioonikordaja	Pearson Correlation	0,982	
Nullhüpoteesile vastav erinevus	Hypothesized Mean Difference	0	
Vabadusastmete arv	df	9	
t-statistik	t Stat	2,10	
Olulisuse tõenäosus ühepoolse hüpoteesi korral	P(T<=t) one-tail	0,033	
Parempoolne kriitiline väärtus ühepoolse hüpoteesi korral	t Critical one-tail	1,83	
Olulisuse tõenäosus kahepoolse hüpoteesi korral	P(T<=t) two-tail	0,065	
Parempoolne kriitiline väärtus kahepoolse hüpoteesi korral	t Critical two-tail	2,26	

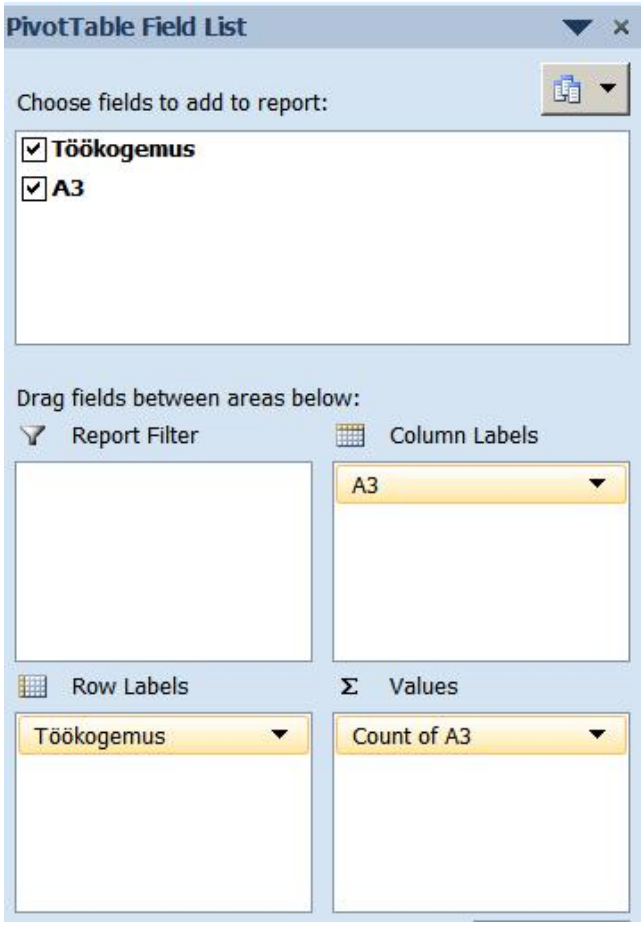
Kui püstitatud hüpotees oli ühepoolne, kasutatakse ühepoolset olulisuse tõenäosust või kriitilist väärtust. Kahepoolse hüpoteesi korral kasutatakse vastavalt kahepoolseid väärtusi. Vasakpoolsed kriitilised väärtused nii ühepoolse kui ka kahepoolse hüpoteesi korral on parempoolsete vastandarvud.

C.7. Risttabeli loomine programmi Excel vahendiga *PivotTable*

Risttabeli loomisel lähtutakse andmetabelist, kus veergudes on erinevad tunnused ja ridades erinevad objektid. Lähtume näites 7.23 toodud andmetabelist, kus on

kahe tunnuse väärtused: töötaja töökogemus ning hinnang väitele A3 (vt ka faili N07Hüpoteesid).

Risttabeli loomiseks tuleb andmetabel välja valida ning nupurealt *Insert* valida *PivotTable*. Aknas *Pivot Table Field List* lohistada üks tunnus piirkonda *Row Labels* ning teine tunnus piirkonda *Column Labels*. Seejärel lohistada teine tunnus ka piirkonda *Values* ning seal valida *Value Field Settings* ning *Count*.

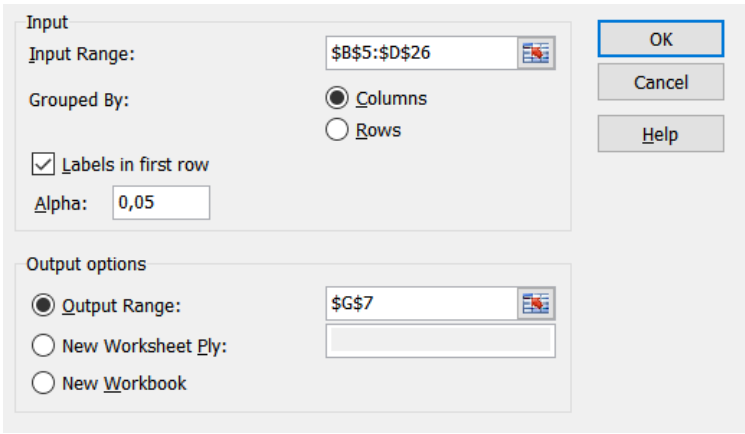


Tulemuseks saadakse järgmine risttabel (vt ka faili N07Hüpoteesid):

Count of A3	Column Labels			
Row Labels	2	3	4	Grand Total
1	7	17	17	41
2	16	8	21	45
3	14	17	20	51
Grand Total	**37**	**42**	**58**	**137**

C.8. Programmi Excel analüüsivahend *ANOVA: Single Factor*

Andmeanalüüsi vahendi *ANOVA: Single Factor* kasutamiseks peab Excelis olema aktiveeritud andmeanalüüsi komplekt *Data Analysis* (vt lisa C.2). Nupureal „Andmed“ („*Data*“) oleva nupuga „*Data Analysis*“ kutsutakse välja andmeanalüüsi vahendite komplekt ja valitakse sealt „*ANOVA: Single Factor*“. Piirkond *Input Range* on kõikide valimite andmed, soovitavalt koos pealkirjadega. Valimid võivad olla erineva mahuga, ristkülikukujulise piirkonna suurus on määratud valimite arvu ja kõige suurema mahuga valimi elementide arvuga (pluss rida veergude pealkirjadega). *Grouped by Columns* valitakse siis, kui andmed on veergudes, ja *Grouped by Rows*, kui andmed on ridades. Kui valiti välja ka veergude pealkirjad, siis märkida *Labels in first row*. Olulisuse nivoo on *Alpha*. See, kuhu paigutatakse aruanne, määratakse valikutes *Output options*.



Väljastatakse kaks tabelit. Tabelis *Summary* on valimite maht, elementide summa, valimite keskmised ja dispersioonid. Tabelis ANOVA on dispersioonanalüüsi tulemused. Arvutused teostatakse alapeatükis 7.14 toodud valemite (7.74)–(7.80) järgi.

Järgnevas tabelis on toodud näite 7.25 dispersioonanalüüsi tulemused (vt ka faili N07Hüpoteesid leht N7.25,25). Lisatud on eestikeelsed seletused.

Anova: Single Factor

SUMMARY	Valimi maht	Valimi elementide summa	Valimi keskmine	Valimi dispersioon
Groups	*Count*	*Sum*	*Average*	*Variance*
A	18	1477	82,06	50,76
B	21	1694	80,67	57,73
C	19	1666	87,68	27,34

ANOVA

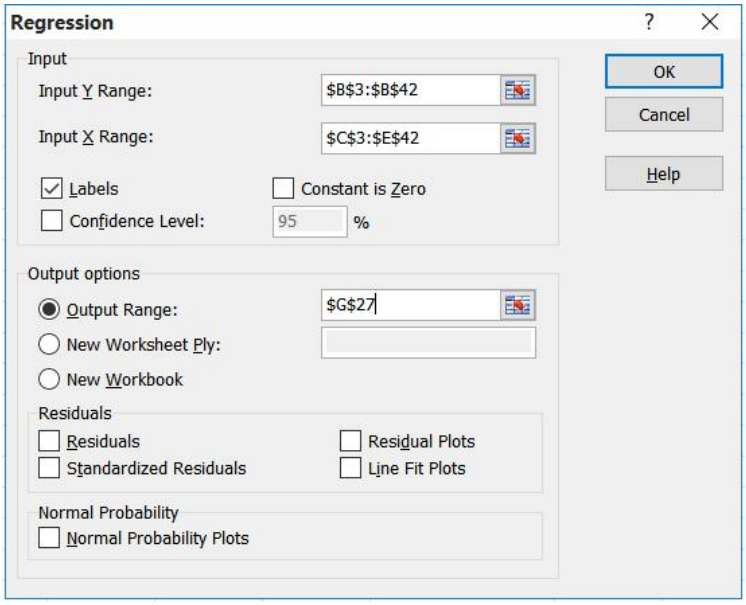
Varieeruvuse allikas		Vabadus-astmete arv	Keskruut	F-statistik	Olulisuse tõenäosus	Kriitiline F väärtus
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	538,2	2	269,1	5,90	0,0048	3,16
Within Groups	2509,7	55	45,6			
Total	3047,9	57				

Veerus „Varieeruvuse allikas (*Source of Variation*)“ on

Between Groups	rühmadevaheline varieerumine;
Within Groups	rühmasisene varieerumine;
Total	summaarne varieerumine.

C.9. Programmi Excel analüüsivahend *Regression*

Lineaarseks regressioonanalüüsiks programmis Excel sobib hästi andmeanalüüsi vahend *Regression*. Selle kasutamiseks peab Excelis olema aktiveeritud andmeanalüüsi komplekt *Data Analysis* (vt lisa C.2). Nupureal „Andmed“ („*Data*“) oleva nupuga „*Data Analysis*“ kutsutakse välja andmeanalüüsi vahendite komplekt ja valitakse sealt „*Regression*“.



Põhilised valikud, mis tuleb teha:

Input Y Range näidatakse, millistes lahtrites on sõltuva tunnuse Y väärtused. *Input X Range* näidatakse, millistes lahtrites on seletavate tunnuste $X_1, X_2, \ldots X_k$ väärtused. Mitme seletava tunnuse korral peavad kõik andmeveerud asuma kõrvuti.

Labels märgitakse, kui eelnevalt märgiti ära ka veerude pealkirjad (soovitav).

Output options näidatakse, kuhu paigutada väljund.

Regressioonanalüüsi aruanne koosneb kolmest osast: summaarne statistika (*SUMMARY OUTPUT*), ANOVA tabel ning kordajate tabel. Esitame näites 9.16 hinnatud mudeli aruande koos eestikeelsete seletustega (vt ka faili N09Regressioon leht N9.12,16,17).

SUMMARY OUTPUT

Regression Statistics		
Multiple R	0,846	korrelatsioonikordaja
R Square	0,715	determinatsioonikordaja
Adjusted R Square	0,691	korrigeeritud determinatsioonikordaja
Standard Error	35,61	mudeli standardviga
Observations	39	vaatluste arv

ANOVA

Varieeruvuse allikas	Vabadus-astmete arv	Hälvete ruutude summa	Keskruut	F- statistik	F-statistiku olulisuse tõenäosus
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	111392,6	37130,9	29,28	1,18E-09
Residual	35	44390,6	1268,3		
Total	38	155783,2			

	Parameeter	Standard-viga	t-statistik	Olulisuse tõenäosus p	Usaldus-vahemiku alumine piir	Usaldus-vahemiku ülemine piir
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2445	93,6	26,113	1,54E-24	2254,7	2634,9
TTASU	−47,6	23,0	−2,070	0,0459	−94,32	−0,91
VARAD	0,0264	0,00393	6,720	8,8E-08	0,018	0,034
VANUS	−8,66	1,71	−5,078	1,27E-05	−12,13	−5,20

Järgnevalt lisavõimalused, mis saab märkida aknas *Regression*.

Constant is Zero märgitakse, kui regressioonsirge peab läbima nullpunkti.

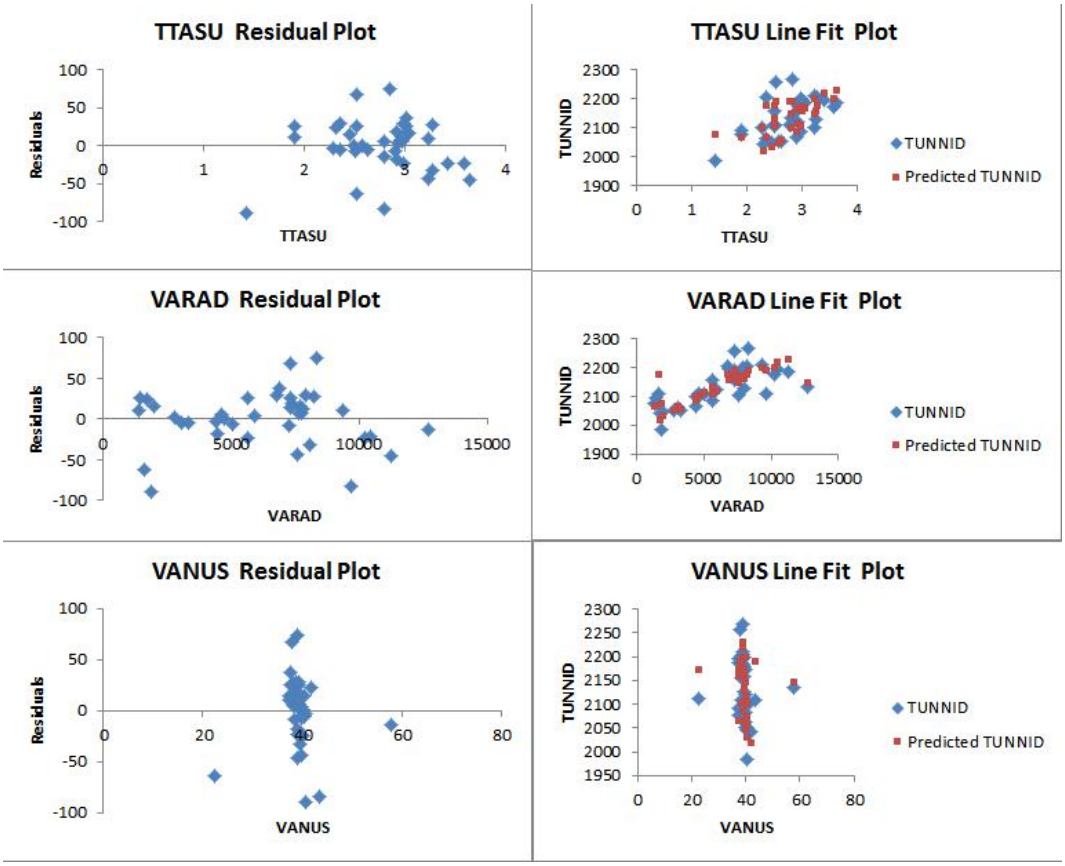
Confidence Level märgitakse, kui soovitakse saada parameetrite usalduspiire usaldatavusega, mis ei ole 95% (näiteks 90% või 99%). Usalduspiirid usaldatavusega 95% leitakse vaikimisi alati.

Residuals: lisatakse tabel, kus on mudelväärtused (*Predicted*) ja regressioonijäägid (*Residuals*).

Standardized Residuals: leitakse standardiseeritud jäägid (valem (9.53)).

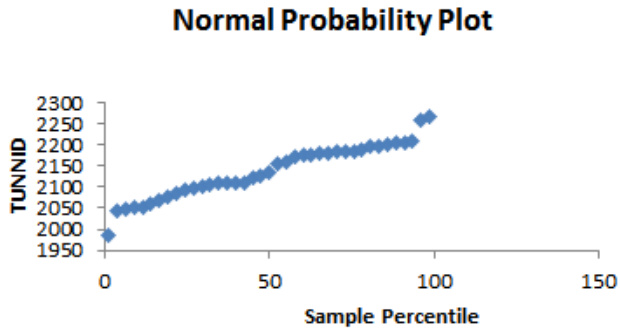
Residual Plots: regressioonijääkide põhjal konstrueeritakse jääkide diagramm, kus horisontaalteljel on seletav tunnus. Jääkide diagramme on sama palju kui seletavaid tunnuseid (joonis C.1 vasakul).

Line Fit Plots: iga seletava tunnuse jaoks konstrueeritakse diagramm, kus horisontaalteljel on seletava tunnuse väärtused ning vertikaalteljel sõltuva tunnuse väärtused (joonis C.1 paremal). Diagrammil esitatakse empiirilised andmed ning mudelväärtused (*Predicted*). Ühe seletava tunnuse korral asuvad mudelväärtused sirgel. Mitmese regressioonmudeli korral, kui on k seletavat tunnust, asuvad mudelväärtustele vastavad punktid k-mõõtmelisel tasandil $k+1$ mõõtmelises ruumis (vt joonis 9.26).



Joonis C.1. Vasakul jääkide diagrammid, paremal mudelväärtuste võrdlus tegelike väärtustega

Normal Probability Plots: lisatakse sõltuva tunnuse protsentiilide diagramm (joonis C.2). Kui see on sirge, siis allub sõltuv tunnus normaaljaotusele. Seda diagrammi kasutatakse vähe, sest see ei abista regressioonmudeli koostamisel. Olulisem oleks jääkide normaaljaotuse diagramm, kuid seda Excel ei väljasta.



Joonis C.2. Sõltuva tunnuse protsentiilide diagramm

C.10. Mitmene regressioonanalüüs: funktsiooni LINEST kasutamine

Funktsioon LINEST leiab mitmese lineaarse regressioonmudeli parameetrite hinnangud ja teised statistilised suurused. LINEST on massiivifunktsioon (*array function*) ja selle kasutamine erineb tavaliste tabelarvutuse funktsioonide kasutamisest. Kui tavaline funktsioon (näiteks AVERAGE) väljastab ühe väärtuse lahtrisse, mis on parajasti aktiivne, siis massiivifunktsioon väljastab korraga mitu väärtust. Järelikult tuleb enne funktsiooni valimist ära märkida kõik need lahtrid, kuhu väärtused väljastatakse. Märkima peab $k+1$ veergu ja 5 rida, kus k on regressorite arv. $k+1$ veergu on selleks, et mudelis on ka vabaliige. Näiteks mudeli $y = ax + b$ korral tuleb märkida 2×5 lahtrit, mudeli $y = a_0 + a_1x_1 + a_2x_2$ korral 3×5 lahtrit jne. Seejärel otsida üles funktsioon LINEST. Funktsiooni argumendid on järgmised:

Known_ y's funktsioontunnuse Y väärtuste piirkond;

Known_ x's kõigi argumenttunnuste X_l väärtuste piirkond;

Const kui mudelis on vabaliige, siis 1, nullpunkti läbiva joone korral 0;

Stats märkida 1 lisastatistika (standardvead, determinatsioonikordaja jm) saamiseks.

Lõpetamiseks tuleb valida klahvikombinatsioon Ctrl+Shift+Enter.

Näiteks kahe regressoriga mudeli $y = a_0 + a_1x_1 + a_2x_2$ korral väljastatakse järgmine tabel:

Kordajad	a_2	a_1	a_0
Standardvead	$se(a_2)$	$se(a_1)$	$se(a_0)$
	R^2	se	#N/A
	F	$n-k-1$	#N/A
	SSR	SSE	#N/A

Siin R^2 on determinatsioonikordaja, se mudeli standardviga, F on F-statistik, $n-k-1$ jääkhajuvuse vabadusastmete arv, SSR regressioonhajuvus ja SSE jääkhajuvus. Ülejäänud lahtrites informatsioon puudub.

Parameetrite statistilise olulisuse testimiseks tuleb leida valemist (9.69) t-statistikud. Vastavad olulisuse tõenäosused leitakse tabelarvutuse funktsiooni T.DIST.2T abil, kus vabadusastmete arv on $n - k - 1$.

Näites 9.11 hinnati ravimimüügi keti Walgreens käibemudelit, milles on 5 seletavat tunnust. Funktsiooni LINEST kasutamine selle mudeli hindamiseks käib järgmiselt (vt ka faili N09Regressioon lehte N9.11).

1. Valida lehe tühjas osas välja 6×5 lahtrit.
2. Otsida funktsioonide nimekirjast üles funktsioon LINEST.
3. Märkida ära Y andmete massiiv (*Known_ y's*) lahtrid A4:A30.
4. Märkida ära X andmete massiiv (*Known_ x's*) lahtrid B4:F30.
5. *Const* sisestada 1.
6. *Stats* sisestada 1.
7. Vajutada Ctrl+Shift+Enter.

Väljastatakse järgmine tabel:

-5,31	13,6	11,5	0,175	16,2	-18,9
1,705	1,770	2,532	0,0576	3,544	30,150
0,99318	17,65	#N/A	#N/A	#N/A	#N/A
611,6	21	#N/A	#N/A	#N/A	#N/A
952538,9	6541,41	#N/A	#N/A	#N/A	#N/A

Kuna funktsiooni LINEST kasutamine on tülikas ja see ei väljasta kogu regressioonanalüüsi jaoks vajalikku informatsiooni (testimiseks tuleb teha lisaarvutusi), siis programmis Excel on soovitav alati kasutada analüüsivahendit *Regression* komplektist *Data Analysis*. Funktsiooni LINEST tuleb kasutada vaid LibreOffice Calc korral.

C.11. Tabelarvutuse funktsioonid

Järgnevalt on õpikus kirjeldatud tabelarvutuse funktsioonid programmides MS Excel (versioon 2010) ja LibreOffice Calc (versioon 5.1.0.3) rühmitatud kasutusotstarbe järgi. Teises veerus on viidatud lehekülgedele, kus asuvad juhised vastava funktsiooni kasutamise kohta. Viimases veerus on vastava funktsiooni nimetus tarkvara vanemates versioonides (Exceli versioonid kuni 2007 ja LibreOffice Calc versioonid kuni 4.1). Neid on hea teada, kui tuleb avada vanemates versioonides loodud faile. Kui varasemates versioonides vastavat funktsiooni pole, siis on viimases veerus märgitud „Puudub“. Kui viimases veerus pole midagi märgitud, siis on funktsioon sama, mis vanemates versioonides.

Sagedustabelid

Funktsioon	Lk	Kirjeldus	Vana
COUNTIF	32	Leiab tingimusele *criteria* vastavale piirkonnas *range* olevate väärtuste arvu. Kriteeriumiks võib olla nii arv kui tekst	
FREQUENCY	33	Sageduste leidmine intervallitud variatsioonrea korral. Kasutamisõpetust võib vaadata õpiku autori kodulehel olevalt ekraanivideolt `http://www.sauga.pri.ee/statistika_excelis`	

Keskmised

Funktsioon	Lk	Kirjeldus	Vana
AVERAGE	43	Aritmeetiline keskmine	
GEOMEAN	77	Geomeetriline keskmine	
HARMEAN	73	Harmooniline keskmine	
MEDIAN	51	Mediaan	
MODE.MULT	67	Moodide leidmine multimodaalse kogumi korral. Tegemist on massiivifunktsiooniga	Puudub
MODE.SNGL	67	Mood	MODE
PERCENTILE.EXC	62	Protsentiilid. Väljastab väärtuse a, nii et $P(X < a) = k$	Puudub
PERCENTILE.INC	62	Protsentiilid. Väljastab väärtuse a, nii et $P(X \leq a) = k$	PERCENTILE
PERCENTRANK.EXC	62	Leiab, milline protsentiili järk k vastab arvukogumist valitud arvule a, nii et $P(X < a) = k$. Funktsiooni PERCENTILE.EXC pöördfunktsioon	Puudub
PERCENTRANK.INC	62	Leiab, milline protsentiili järk k vastab arvukogumist valitud arvule a, nii et $P(X \leq a) = k$. Funktsiooni PERCENTILE.INC pöördfunktsioon	PERCENTRANK
QUARTILE.EXC	62	Kvartiilid. Väljastab väärtuse a, nii et $P(X < a) = 25\%$ või 50% või 75%	Puudub
QUARTILE.INC	62	Kvartiilid, lisaks miinimum ($quart$=0) ja maksimum ($quart$=5)	QUARTILE
SUMSQ	79	Ruutude summa	

Varieerumine ja jaotuse kuju

Funktsioon	Lk	Kirjeldus	Vana
AVEDEV	93	Keskmine absoluuthälve	
DEVSQ	441	Hälvete ruutude summa	
KURT	104	Püstakuse kordaja	
MAX	92	Maksimum	
MIN	92	Miinimum	
SKEW	103	Asümmeetriakordaja	
STANDARDIZE	102	Standardiseeritud väärtus	
STDEV.P	95	Kogumi standardhälve	STDEVP
VAR.P	94	Kogumi dispersioon	VARP

Jaotusseadused

Funktsioon	Lk	Kirjeldus	Vana
BINOM.DIST	179	Binoomjaotuse tõenäosus või jaotusfunktsioon	BINOMDIST
BINOM.INV	184, 273, 342, 726	Leiab binoomjaotuse korral minimaalse positiivsete tulemuste arvu m, nii et $F(m) > \alpha$	CRITBINOM
CHISQ.INV.RT	268, 348, 727	$\chi^2(\nu)$-jaotuse täiendkvantiil, kus ν on vabadusastmete arv	CHIINV
EXPON.DIST	196	Eksponentjaotuse jaotustihedus või jaotusfunktsioon	EXPONDIST
F.INV	331	F-jaotuse kvantiil F_β, nii et $P(X < F_\beta) = \beta$	Puudub
F.INV.RT	331, 468, 724	F-jaotuse täiendkvantiil	FINV
NORM.DIST	201, 205	Normaaljaotuse jaotustihedus või jaotusfunktsioon	NORMDIST
NORM.INV	205	Funktsiooni NORM.DIST pöördfunktsioon. Leiab normaaljaotuse korral väärtuse x, mille puhul jaotusfunktsioon $F(x) = Probability$	NORMINV
NORM.S.DIST	214	Standardiseeritud normaaljaotuse jaotustihedus või jaotusfunktsioon	NORMSDIST

Funktsioon	Lk	Kirjeldus	Vana
NORM.S.INV	254, 300	Funktsiooni NORM.S.DIST pöördfunktsioon. Leiab standardiseeritud normaaljaotuse kvantiili z_β või täiendkvantiili vastandväärtuse $-z_\alpha$. Kvantiili leidmisel on parameeter *Probability* vasakpoolne tõenäosus $\beta = P(X < z_\beta)$. Täiendkvantiili leidmisel on *Probability* parempoolne tõenäosus $\alpha = P(X > z_\alpha)$	NORMSINV
POISSON.DIST	190	Poissoni jaotuse tõenäosus või jaotusfunktsioon	POISSON
RAND	171	Vahemikus $[0, 1]$ ühtlaselt jaotunud juhusliku suuruse väärtuste genereerimine	
T.DIST.2T	741	Leiab $t(\nu)$-jaotusele alluva juhusliku suuruse X korral tõenäosuse α, nii et $P(\|X\| > t_\alpha) = \alpha$. Ette tuleb anda t_α väärtus ja vabadusastmete arv. Funktsiooni T.INV.2T pöördfunktsioon	TDIST
T.INV	259, 315, 324, 723	$t(\nu)$-jaotuse kvantiil t_β, nii et parameeter *Probability* on vasakpoolne tõenäosus $\beta = P(X < t_\beta)$. Täiendkvantiili leidmisel on *Probability* parempoolne tõenäosus $\alpha = P(X > t_\alpha)$ ja tulemuseks on täiendkvantiili vastandväärtus $-t_\alpha$	Puudub
T.INV.2T	259, 315, 324, 441, 475	$t(\nu)$-jaotuse korral väärtus t_α, nii et $P(\|X\| > t_\alpha) = \alpha$. On seotud funktsiooniga T.INV järgmiselt: T.INV.2T(α; ν) = – T.INV($\alpha/2$; ν)	TINV

Valikvaatlused

Funktsioon	Lk	Kirjeldus	Vana
CONFIDENCE.NORM	254	Keskväärtuse usaldusvahemiku poollaius suure valimi korral	CONFIDENCE
CONFIDENCE.T	259	Keskväärtuse usaldusvahemiku poollaius väikese valimi korral	Puudub
SMALL	273	Leiab arvu, mille järjenumber kasvavalt järjestatud kogumis *Array* on K	

Funktsioon	Lk	Kirjeldus	Vana
STDEV.S	248	Valimi standardhälve	STDEV
VAR.S	248	Valimi dispersioon	VAR

Hüpoteeside kontrollimine

Funktsioon	Lk	Kirjeldus	Vana
CHISQ.TEST	355	Olulisuse tõenäosus χ^2-testi korral, kui risttabeli abil testitakse kahe tunnuse vahelist seost	CHITEST
F.TEST	331, 335	Olulisuse tõenäosus kahepoolse F-testi korral	FTEST
Z.TEST	313	Olulisuse tõenäosus kahe kogumi keskväärtuse testimisel ühepoolse z-testiga	ZTEST
T.TEST	329	Olulisuse tõenäosus kahe kogumi keskväärtuse testimisel t-testiga	TTEST

Korrelatsioon ja regressioon

Funktsioon	Lk	Kirjeldus	Vana
CORREL	398	Lineaarne korrelatsioonikordaja. Sama, mis PEARSON	
COVARIANCE.P	397	Kogumi kovariatsioon	COVAR
COVARIANCE.S	397	Valimi kovariatsioon	Puudub
INTERCEPT	427	Lihtsa lineaarse regressioonmudeli vabaliige	
LINEST	462, 741	Mitmese lineaarse regressioonmudeli parameetrid, standardvead, determinatsioonikordaja, F-statistik	
PEARSON	398	Lineaarne korrelatsioonikordaja. Sama, mis CORREL	
RSQ	436	Lihtsa lineaarse regressioonmudeli determinatsioonikordaja	
SLOPE	427	Lihtsa lineaarse regressioonmudeli lineaarliikme kordaja ehk sirge tõus	
STEYX	438, 441	Lihtsa lineaarse regressioonmudeli standardviga	
TREND	430	Lihtsa lineaarse regressioonmudeli mudelväärtus	

Lisa D

Õpikuga kaasasolevad failid

Õpikuga on kaasas tabelarvutusfailid õpikus esinevate näidetega ja ülesannete jaoks vajalike andmetega. Failid saab alla laadida TTÜ Raamatukogu digikogust aadressilt `https://digi.lib.ttu.ee/` jaotisest "Õpikud ja õppevahendid". Näited ja ülesannete andmed on failidesse jaotatud peatükkide kaupa. Kõik failid on kahes formaadis: Exceli (*.xlsx) ja avatud dokumendivormingus (*.ods).

Ptk	Näited	Ülesannete andmed
	Exceli failid	
1	N01Sissejuhatus.xlsx	ÜL01Arvjoonised.xlsx
		ÜL01Sagedustabelid.xlsx
2	N02Keskmised.xlsx	ÜL02Keskmised.xlsx
3	N03Varieerumine.xlsx	ÜL03Varieerumine.xlsx
5	N05Jaotused.xlsx	ÜL05Jaotused.xlsx
6	N06Valikvaatlused.xlsx	ÜL06Valikvaatlused.xlsx
7	N07Hüpoteesid.xlsx	ÜL07Hüpoteesid.xlsx
8	N08Korrelatsioon.xlsx	ÜL08Korrelatsioon.xlsx
9	N09Regressioon.xlsx	ÜL09Regressioon.xlsx
10	N10Aegread.xlsx	ÜL10Aegread.xlsx
11	N11Indeksid.xlsx	ÜL11Indeksid.xlsx
	Avatud dokumendivormingus failid	
1	N01Sissejuhatus.ods	ÜL01Sissejuhatus.ods
		ÜL01Sagedustabelid.ods
2	N02Keskmised.ods	ÜL02Keskmised.ods
3	N03Varieerumine.ods	ÜL03Varieerumine.ods
5	N05Jaotused.ods	ÜL05Jaotused.ods
6	N06Valikvaatlused.ods	ÜL06Valikvaatlused.ods
7	N07Hüpoteesid.ods	ÜL07Hüpoteesid.ods
8	N08Korrelatsioon.ods	ÜL08Korrelatsioon.ods
9	N09Regressioon.ods	ÜL09Regressioon.ods
10	N10Aegread.ods	ÜL10Aegread.ods
11	N11Indeksid.ods	ÜL11Indeksid.ods

Register

Kirjandus

20 aastat Eesti piimaturul 1995–2013 (2013). Ülevaade. Eesti Konjunktuuriinstituut.

Aarma, A. (2010). *Arvjoonised.* Tln: TTÜ Kirjastus. 48 lk.

Aarma, A. ja Vensel, V. (2005). *Statistika teooria põhikursus.* 2. tr. Tln: Külim. 214 lk.

Aasma, A. ja Levin, A. (2013). *Matemaatilised meetodid majanduses.* Argo. 270 lk.

Abišala, A. (2015). *Mudeli looja selgitab: kuidas kasutada börsi analüüsimisel Bollingeri koridori.* Postimees. URL: `http://majandus24.postimees.ee/2750796/mudeli-looja-selgitab-kuidas-kasutada-borsi-analuusimisel-bollingeri-koridori` (25.2.2015).

Agresti, A. and Coull, B. A. (1998). Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions. — *The American Statistician*, Vol. 52, No. 2, pp. 119–126. DOI: `10.1080/00031305.1998.10480550`.

Ahmed, A. M., Andersson, L., and Hammarstedt, M. (2012). Does age matter for employability? A field experiment on ageism in the Swedish labour market. — *Applied Economics Letters*, Vol. 19, No. 4, pp. 403–406. DOI: `10.1080/13504851.2011.581199`.

Allen, T. T. (2006). *Introduction to Engineering Statistics and Six sigma.* Springer. 529 pp.

Anderson, D., Sweeney, D., and Willams, T. (1999). *Statistics for Business and Economics.* 7th ed. South-Western College Publishing. 1080 pp.

Anderson, R. L. and Anderson, K. P. (1988). A comparison of women in small and large companies. — *American Journal of Small Business*, Vol. 12, No. 3, pp. 23–33.

Anscombe, F. J. (1973). Graphs in Statistical Analysis. — *The American Statistician*, Vol. 27, No. 1, pp. 17–21. DOI: `10.2307/2682899`.

Antin, A. (2002). *Ainetöö statistikas.* Magistritöö. Tallinn, Audentese Kõrgem Ärikool.

Arroba, T. and Wedgwood-Oppenheim, F. (1994). Do Senior Managers Differ in the Public and Private Sector?: An Examination of Team Role Preferences. — *Journal of Managerial Psychology*, Vol. 9, No. 1, pp. 13–16. DOI: `10.1108/02683949410051468`.

Ashenfelter, O., Ashmore, D., and Lalonde, R. (1995). Bordeaux Wine Vintage Quality and the Weather. — *Chance*, Vol. 8, No. 4, pp. 7–14. DOI: 10.1080/09332480.1995.10542468.

Audi (2014). *Audi 2014 Annual Report*. Annual Report. Audi AG.

Baker, T. K. and Collier, D. A. (1999). A comparative revenue analysis of hotel yield management heuristics. — *Decision Sciences*, Vol. 30, No. 1, pp. 239–263.

Bakker, A. and Gravemeijer, K. P. E. (2006). An Historical Phenomenology of Mean and Median. — *Educational Studies in Mathematics*, Vol. 62, No. 2, pp. 149–168. DOI: 10.1007/s10649-006-7099-8.

Baruch, Y. and Holtom, B. C. (2008). Survey response rate levels and trends in organizational research. — *Human Relations*, Vol. 61, No. 8, pp. 1139–1160. DOI: 10.1177/0018726708094863.

Bass, F. M. (1969). A New Product Growth for Model Consumer Durables. — *Management Science*, Vol. 15, pp. 215–227.

Belbin, R. M. (1981). *Management Team*. Heinemann.

Bell, J., Crick, D., and Young, S. (2004). Small Firm Internationalization and Business Strategy: An Exploratory Study of 'Knowledge-Intensive' and 'Traditional' Manufacturing Firms in the UK. — *International Small Business Journal*, Vol. 22, No. 1, pp. 23–56. DOI: 10.1177/0266242604039479.

Belmont, D. M. (1957). A Pattern of Interstation Air Travel. — *Transactions of the American Society of Civil Engineers*, Vol. 122, No. 1, pp. 864–871.

— (1958). A Study of Airline Interstation Traffic. — *Journal of Air Law and Commerce*, Vol. 25, pp. 361–368.

Beneplanc, G. and Rochet, J.-C. (2011). *Risk Management in Turbulent Times*. Oxford University Press. 224 pp.

Bitran, G. R. and Gilbert, S. M. (1996). Managing Hotel Reservations with Uncertain Arrivals. — *Operations Research*, Vol. 44, No. 2, pp. 35–49. DOI: 10.1287/opre.44.1.35.

BMW (2013). *Financial Statements of BMW AG Financial Year 2013*. Tech. rep. BMW AG.

Bolger, M. P., Egan, K., and Tao, S. S.-H. (2008). *Letter Concerning Arsenic in Pear Juice Products*. U.S. Food and Drug Administration. URL: http://www.fda.gov/ (1.2.2015).

Bollinger, J. (1992). Using Bollinger bands. — *Stocks & Commodities*, Vol. 10, No. 2, pp. 47–51.

Bolotin, V. A. (1994). "Telephone Circuit Holding Time Distributions". In: *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks: Proceedings of the 14th International Teletraffic Congress — ITC 14, Antibes Juan-les-Pins, France, 6–10 June, 1994*. Ed. by J. Labetoulle and J. W. Roberts. Colloquia Mathematica Societatis Janos Bolyai, pp. 125–134.

Box, G. E. P. and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day. 537 pp.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval Estimation for a Binomial Proportion. — *Statist. Sci.* Vol. 16, No. 2, pp. 101–133. DOI: 10.1214/ss/1009213286.

Brown, L. et al. (2005). Statistical Analysis of a Telephone Call Center. — *Journal of the American Statistical Association*, Vol. 100, No. 469, pp. 36–50. DOI: 10.1198/016214504000001808.

Cantor, R., Hamilton, D., and Tennant, J. (2007). *Confidence Intervals for Corporate Default Rates.* Moody's Special Comment. Moody's Investors Service.

Chatterjee, S. and Hadi, A. S. (2006). *Regression Analysis by Example.* A John Wiley & Sons, Inc. 374 pp.

Cifuentes, A. and O'Connor, G. (1996). *The Binomial Expansion Method Applied to CBO/ CLO Analysis.* Special report. Moody's Investors Service.

Cinca, C. S., Molinero, C. M., and Larraz, J. G. (2005). Country and size effects in financial ratios: A European perspective. — *Global Finance Journal*, Vol. 16, No. 1, pp. 26–47. DOI: 10.1016/j.gfj.2005.05.003.

Cincera, M. (1997). Patents, R&D, and technological spillovers at the firm level: some evidence from econometric count models for panel data. — *Journal of Applied Econometrics*, Vol. 12, No. 3, pp. 265–280. DOI: 10.1002/(SICI)1099-1255(199705)12:3<265::AID-JAE439>3.0.CO;2-J.

Clark, R v [2003] EWCA Crim 1020 (11 April 2003) (2003). England and Wales Court of Appeal (Criminal Division) Decisions. URL: http://www.bailii.org/ew/cases/EWCA/Crim/2003/1020.html (18.10.2015).

Cobb, C. and Douglas, P. (1928). A Theory of Production. — *American Economic Review*, Vol. 18, No. 1, pp. 139–165.

Consumer Reports Buying Guide 2013: Home Appliances (2013).

Daamen, W. and Hoogendoorn, S. (2007). Free speed distributions — Based on empirical data in different traffic conditions. In: *Pedestrian and Evacuation Dynamics 2005.* Ed. by N. Waldau et al. Springer Berlin Heidelberg, pp. 13–25. DOI: 10.1007/978-3-540-47064-9_2.

Dahiya, R. (2013). Indian Youth and Internet: An Empirical Study. — *Journal of Marketing & Communication*, Vol. 9, No. 2, pp. 19–28.

Davis, D. J. (1952). An Analysis of Some Failure Data. — *Journal of the American Statistical Association*, Vol. 47, No. 258, pp. 113–150. DOI: 10.1080/01621459.1952.10501160.

Dewenter, K. L. and Malatesta, P. H. (2001). State-Owned and Privately Owned Firms: An Empirical Analysis of Profitability, Leverage, and Labor Intensity. — *The American Economic Review*, Vol. 91, No. 1, pp. 320–334.

Diakomihalis, M. N. (2011). Financial Structure and Profitability Analysis of Greek Hotels. — *The Journal of Hospitality Financial Management*, Vol. 19, No. 1, pp. 51–70. DOI: 10.1080/10913211.2011.10653900.

Dominick's Database (2016). James M. Kilts Center for Marketing, University of Chicago Booth School of Business. URL: https://research.chicagobooth.edu/kilts/marketing-databases/dominicks (15.5.2016).

Dutton, J. M. and Thomas, A. (1984). Treating Progress Functions as a Managerial Opportunity. — *The Academy of Management Review*, Vol. 9, No. 2, pp. 235–247. DOI: 10.5465/AMR.1984.4277639.

Eamets, R. jt (2011). *Energeetika tööjõu uuring.* Uuringu lõpparuanne. Tartu Ülikool, SA Poliitikauuringute Keskus Praxis.

Eesti sotsiaaluuring (2013). Eesti Statistikaamet. URL: http://www.stat.ee/51917 (17.10.2014).

Eesti tööjõu-uuring (2012). Eesti Statistikaamet. URL: http://www.stat.ee/51920 (9.3.2014).

Eglit, T. (2014). *Obesity, impaired glucose regulation, metabolic syndrome and their associations with high-molecular-weight adiponectin levels.* PhD thesis. Tartu University.

Eisenberg, T., Sundgren, S., and Wells, M. T. (1998). Larger board size and decreasing firm value in small firms. — *Journal of Financial Economics*, Vol. 48, No. 1, pp. 35–54. DOI: 10.1016/S0304-405X(98)00003-8.

Eisenhauer, J. G. (2003). Regression through the Origin. — *Teaching Statistics*, Vol. 25, No. 3, pp. 76–80.

Emor (2001). *Hoiakud web'i keskkonnas tuludeklaratsiooni täitmise suhtes.* Uuringu aruanne. AS Emor.

Eskandari, H., Babolmorad, N., and Farrokhnia, N. (2013). "Bottleneck Analysis in a Pharmaceutical Production Line Using Simulation Approach". In: *Proceedings of the 2013 Summer Computer Simulation Conference.* SCSC '13. Society for Modeling and Simulation International, 12:1–12:8.

Ess, J. (2014). *Viimsi valla teede ja tänavate uuring.* Uuringu aruanne. AS Teede Tehnokeskus.

Ettredge, M., Gerdes, J., and Karuga, G. (2005). Using Web-based Search Data to Predict Macroeconomic Statistics. — *Communications of the ACM*, Vol. 48, No. 11, pp. 87–92. DOI: 10.1145/1096000.1096010.

Fitzpatrick, S. and Scott, A. (1987). Quick Simultaneous Confidence Intervals for Multinomial Proportions. — *Journal of the American Statistical Association*, Vol. 82, No. 399, pp. 875–878.

Fontanills, G. A. and Gentile, T. (2001). *Stock Market Course.* Wiley. 463 pp.

Franchetti, M. J. (2015). *Lean Six Sigma for Engineers and Managers.* CRC Press.

Frecka, T. J. and Hopwood, W. S. (1983). The Effects of Outliers on the Cross-Sectional Distributional Properties of Financial Ratios. — *The Accounting Review*, Vol. 58, No. 1, pp. 115–128.

Freedman, D. and Diaconis, P. (1981). On the histogram as a density estimator: L 2 theory. — *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, Vol. 57, No. 4, pp. 453–476. DOI: 10.1007/BF01025868.

Gallo, M. Á., Tàpies, J., and Cappuyns, K. (2004). Comparison of Family and Nonfamily Business: Financial Logic and Personal Preferences. — *Family Business Review*, Vol. 17, No. 4, pp. 303–318. DOI: 10.1111/j.1741-6248.2004.00020.x.

Galtung, J. (1967). *Theory and Methods of Social Research.* Columbia University Press.

GDP long-term forecast (2016). OECD. URL: http://dx.doi.org/10.1787/d927bc18-en (29.1.2016).

Goev, V. (2009). Estimating the hidden economy in Bulgaria. — *SEER — South-East Europe Review for Labour and Social Affairs*, Vol. 1, pp. 77–93.

Goodman, L. A. (1965). On Simultaneous Confidence Intervals for Multinomial Proportions. — *Technometrics*, Vol. 7, No. 2, pp. 247–254.

Green, P. (2002). *Letter from the President to the Lord Chancellor regarding the use of statistical evidence in court cases.*

Greenberg, D. H. and Kosters, M. (1970). *Income Guarantees and the Working Poor: The Effect of Income Maintenance Programs on the Hours of Work of Male Family Heads.* Tech. rep. Rand.

Greene, W. H. (2002). *Econometric Analysis.* 5th ed. Prentice Hall. 1026 pp.

Greenstein, S. M. and Wade, J. B. (1998). The Product Life Cycle in the Commercial Mainframe Computer Market, 1968–1982. — *The RAND Journal of Economics*, Vol. 29, No. 4, pp. 772–789.

Guzmán, G. (2011). Internet search behavior as an economic forecasting tool: The case of inflation expectations. — *The Journal of Economic and Social Measurement*, Vol. 36, No. 3, pp. 119–167. DOI: 10.3233/JEM-2011-0342.

Haidla, T. (2004). *Reisijate arvud ning lennule mitteilmuvad reisijad.* Ainetöö statistikas. Tallinn, Audentese Ülikool.

Hall, L. A. and Bagchi-Sen, S. (2002). A study of R&D, innovation, and business performance in the Canadian biotechnology industry. — *Technovation*, Vol. 22, No. 4, pp. 231–244. DOI: 10.1016/S0166-4972(01)00016-5.

Handeland, S. O., Imsland, A. K., and Stefansson, S. O. (2008). The effect of temperature and fish size on growth, feed intake, food conversion efficiency and stomach evacuation rate of Atlantic salmon post-smolts. — *Aquaculture*, Vol. 283, No. 1–4, pp. 36–42. DOI: 10.1016/j.aquaculture.2008.06.042.

Hartsenko, J. and Sauga, A. (2013). The role of financial support in SME and economic development in Estonia. — *Business and Economic Horizons*, Vol. 9, No. 2, pp. 10–22.

Higgins, H. N. (1998). Analyst Forecasting Performance in Seven Countries. — *Financial Analysts Journal*, Vol. 54, No. 3, pp. 58–62. DOI: 10.2469/faj.v54.n3.2181.

Hinnainfo (2016) 230.1.

Holzer, H. J. et al. (1993). Are Training Subsidies for Firms Effective? The Michigan Experience. — *Industrial and Labor Relations Review*, Vol. 46, No. 4, pp. 625–636.

Houthakker, H. (1951). Some Calculations on Electricity Consumption in Great Britain. — *Journal of the Royal Statistical Society. Series A (General)*, Vol. 114, No. 3, pp. 359–371.

Hunter, J. M. and Shannon, G. W. (1984). Exercises on Distance Decay Using Mental Health Historical Data. — *Journal of Geography*, Vol. 83, No. 6, pp. 277–285. DOI: 10.1080/00221348408980528.

Hyndman, R. and Athanasopoulos, G. (2013). *Forecasting: principles and practice*. OTexts.

Ideon, A. (2003). Valitsusparteide toetus kerkis. — *Postimees 17.02.2003*.

Income distribution statistics (2015). Eurostat. URL: `http://ec.europa.eu/eurostat/statistics-explained/index.php/Income_distribution_statistics#At-risk-of-poverty_rate_and_threshold` (12.7.2015).

Inselberg, K. (2012). *Kaubanduskettide kliendid eelistavad pangakaarti sularahale*. Postimees. Tarbija. URL: `http://tarbija24.postimees.ee/1049894/kaubanduskettide-kliendid-eelistavad-pangakaarti-sularahale` (2.11.2015).

Jarvis, E. (1866). Influence of distance from and nearness to an insane hospital on its use by the people. — *The American Journal of Psychiatry*, Vol. 22, No. 3, pp. 361–406. DOI: `10.1176/ajp.22.3.361`.

Jin, C. (2010). An empirical comparison of online advertising in four countries: Cultural characteristics and creative strategies. — *Journal of Targeting, Measurement and Analysis for Marketing*, Vol. 18, No. 3, pp. 253–261. DOI: `10.1057/jt.2010.18`.

Johnson, L. and Williamson, S. H. (2016). *What Was the U. S. GDP Then?* MeasuringWorth. URL: `http://www.measuringworth.org/usgdp12/` (29.1.2016).

Johnson, R. (1995). A mutliple regression project. — *Teaching Statistics*, Vol. 17, No. 2, pp. 64–66. DOI: `10.1111/j.1467-9639.1995.tb00871.x`.

Josing, M. (2014). Tarbijabaromeeter. — *Konjunktuur*, Kd. 191, Nr 4, lk. 48–53.

Jung, A. F. (1959). Price Variations Among Automobile Dealers in Chicago, Illinois. — *The Journal of Business*, Vol. 32, No. 4, pp. 315–326.

Jõgi, A. (2000). *Tõenäosusteooria, 1. ja 2. osa*. Tln: TTÜ Kirjastus. 416 lk.

Kaarma, O. ja Paas, T. (2000). Riski mõiste ja majandusriskid. Kogumikus: *Riskid Eesti majanduses*. Toim. T. Paas. Tartu Ülikool. Ptk 1, lk. 15–61.

Kaizoji, T. (2004). Inflation and deflation in financial markets. — *Physica A: Statistical Mechanics and its Applications*, Vol. 343, No. 0, pp. 662–668. DOI: `10.1016/j.physa.2004.06.137`.

Kala, R. (2004). *Töötajate motiveerimine väikeettevõttes*. Bakalaureusetöö. Tallinn, Audentese Ülikool.

Kaldaru, H. (1995). *Mikroökonoomika I. Majapidamisteooria ja firmateooria*. Tartu: TÜ Kirjastus. 148 lk.

Kandla, K. jt (2013). *Heategevusalaste hoiakute uuring*. Uuringuaruanne. TNS Emor.

Kehoe, P. J. and Midrigan, V. (2008). *Temporary price changes and the real effects of monetary policy*. Working Paper 14392. National Bureau of Economic Research.

Kerem, K. jt (1988). *Makroökonoomika teooriad ja mudelid*. Toim. M. Randveer. Tallinn: Tallinna Raamatutrükikoda. 286 lk.

Kiivet, V. (2000). *Statistiliste meetodite kasutamine Eesti väärtpaberiturul*. Diplomitöö. Tallinn, Audentese Ülikool.

Kinnucan, H. W. (1986). Demographic Versus Media Advertising Effects On Milk Demand: The Case Of The New York City Market Northeastern. — *Journal of Agricultural and Resource Economics*, Vol. 15, No. 1, pp. 66–74.

Kirss, L. jt (2011). *Eesti üliõpilaste eluolu 2010*. Uuringu aruanne. SA Poliitikauuringute Keskus PRAXIS.

Kitt, J. and Sträter, P. (2008). The Impact of Advertising in the U.S. Sweet Confection Market. — *Journal of Advertising Research*, Vol. 48, No. 1, pp. 22–29.

Knoblauch, M. (2014). *Internet Users Send 204 Million Emails Per Minute*. Mashable. URL: http://mashable.com/2014/04/23/data-online-every-minute/#XDxEpcYgRSqg (29.10.2015).

Kondjukova, J. (2016). *Tulude ebavõrdsuse mõju majanduskasvule*. Bakalaureusetöö. TTÜ.

Koonce, J. C. et al. (2008). Financial Information: Is It Related to Savings and Investing Knowledge and Financial Behavior of Teenagers? — *Journal of Financial Counseling and Planning*, Vol. 19, No. 2, pp. 19–28.

Kovačec, M., Pilipović, A., and Štefanić, N. (2010). Improving the quality of glass containers production with plunger process control. — *CIRP Journal of Manufacturing Science and Technology*, Vol. 3, No. 4, pp. 304–310. DOI: 10.1016/j.cirpj.2011.02.003.

Kroll, L. and Dolan, K. A. (2015). *The World's Billionaires*. Forbes. URL: http://www.forbes.com/billionaires/ (6.3.2015).

Kuiper, S. and College, G. (2008). Introduction to Multiple Regression: How Much Is Your Car Worth? — *Journal of Statistics Education*, Vol. 16, No. 3.

Lee, K.-H., Yang, S.-B., and Choi, M. (2009). The Association between Hospital Ownership and Technical Efficiency in a Managed Care Environment. — *Journal of Medical Systems*, Vol. 33 (4), pp. 307–315. DOI: 10.1007/s10916-008-9192-2.

Leibkonna eelarve uuring (2010). Eesti Statistikaamet. URL: http://www.stat.ee/51901 (11.4.2015).

Leibkonna eelarve uuring (2012). Eesti Statistikaamet. URL: http://www.stat.ee/51901 (20.4.2014).

Lu, W. et al. (2006). Estimation of U.S. Bark Generation and Implications for Horticultural Industries. — *Journal of Environmental Horticulture*, Vol. 24, No. 1, pp. 29–34.

Mahadeva, K., Passmore, R., and Woolf, B. (1953). Individual variations in the metabolic cost of standardized exercises: the effects of food, age, sex and race. — *The Journal of Physiology*, Vol. 121, No. 2, p. 225.

Majeske, K. and Herrin, G. (1998). "Determining warranty benefits for automobile design changes". In: *Annual Reliability and Maintainability Symposium. 1998 Proceedings. International Symposium on Product Quality and Integrity*, pp. 94–99. DOI: 10.1109/RAMS.1998.653636.

Makridakis, S., Wheelwright, S., and Hyndman, J. R. (1998). *Forecasting: Methods and Applications*. 3rd ed. Wiley.

Masanjala, W. H. and Papageorgiou, C. (2008). Rough and lonely road to prosperity: a reexamination of the sources of growth in Africa using Bayesian model averaging. — *Journal of Applied Econometrics*, Vol. 23, No. 5, pp. 671–682. DOI: 10.1002/jae.1020.

Masso, M. jt (2013). *Töölepingu seaduse uuring*. Uuringu aruanne. Poliitikauuringute keskus PRAXIS.

McAfee, A. and Brynjolfsson, A. (2012). Big data: The management revolution. — *Harvard Business Review*, Vol. 90, No. 10, pp. 60–66.

McCarthy, T. M. et al. (2006). The evolution of sales forecasting management: a 20-year longitudinal study of forecasting practices. — *Journal of Forecasting*, Vol. 25, No. 5, pp. 303–324. DOI: 10.1002/for.989.

McClave, J. T., Benson, P., and Sincich, T. (2005). *Statistics for business and economics*. Prentice Hall.

McDaniel, S. W. (1981). Multicollinearity in Advertising-related Data. — *Journal of Advertising Research*, Vol. 21, No. 3, pp. 59–63.

Mendenhall, W. and Sincich, T. (1993). *A Second Course in Business Statistics: Regression Analysis*. 4th ed. Dellen Publishing Company.

Mereste, U. ja Saarepera, M. (1983). *Arvjoonised*. Tallinn: Valgus. 248 lk.

Michelson, G., Niils, H. ja Kala, R. (2001). *Inimeste arvamus progresseeruva tulumaksu kohta*. Uurimustöö. Tallinn, Audentese Kõrgem Ärikool.

Minh, C. C., Sano, K., and Matsumoto, S. (2005). The speed, flow and headway analyses of motorcycle traffic. — *Journal of the Eastern Asia Society for Transportation Studies*, Vol. 6, pp. 1496–1508. DOI: 10.11175/easts.6.1496.

Moore, D. S. and McCabe, G. P. (1989). *Introduction to the Practice of Statistics*. New York: Freeman.

Moraga-González, J. L., Sándor, Z., and Wildenbeest, M. R. (2013). Semi-nonparametric estimation of consumer search costs. — *Journal of Applied Econometrics*, Vol. 28, No. 7, pp. 1205–1223. DOI: 10.1002/jae.2290.

Morton, J., Runciman, B., and Gordon, K. (2014). *Big Data : Opportunities and challenges*. BCS Learning & Development Limited. 60 pp.

Munnell, A. H. et al. (1996). Mortgage Lending in Boston: Interpreting HMDA Data. — *The American Economic Review*, Vol. 86, No. 1, pp. 25–53.

Narula, S. C. and Wellington, J. F. (1977). Prediction, Linear Regression and the Minimum Sum of Relative Errors. — *Technometrics*, Vol. 19, No. 2, pp. 185–190. DOI: 10.2307/1268628.

Nerlove, M. (1963). Returns to Scale in Electricity Supply. In: *Measurement in economics: studies in mathematical economics and econometrics in memory of Yehuda Grunfeld*. Ed. by C. F. Christ. Stanford University Press, pp. 167–198.

Niroomand, F. and Nissan, E. (2007). Socio-Economic Gaps within the EU: A Comparison. — *International Advances in Economic Research*, Vol. 13, No. 3, pp. 365–378. DOI: 10.1007/s11294-007-9092-0.

Noormägi, N. (2010). *Kalendriefektide esinemine NASDAQ OMX Tallinna Väärtpaberiturul*. Bakalaureusetöö. Tallinna Tehnikaülikool.

Paas, T. (1997). *Kvantitatiivsed meetodid majanduses*. Tartu: Tartu Ülikool. 268 lk.

Papp, K., Part, K. ja Tõrik, S. (2001). *KISS. Noorsoouuring 1999*. Uuringu aruanne. Eesti Pereplaneerimise Liit.

Parring, A.-M., Vähi, M. ja Käärik, E. (1997). *Statistilise andmetöötluse algõpetus*. Tartu: TÜ Kirjastus. 405 lk.

Pettuseriskide alane uuring Eestis (2014). Uuringu aruanne. Ernst & Young Baltic.

Pham, H., ed. (2006). *Springer Handbook of Engineering Statistics.* Springer. 1120 pp.

Pilinkiene, V. (2008). Selection of market demand forecast methods: Criteria and application. — *Engineering Economics*, Vol. 58, No. 3, pp. 19–25.

Pitsi, T., Kambek, L. ja Jõeleht, A. (2014). *NutriData toidu koostise andmebaas, väljaanne 6.* Tervise Arengu Instituut. URL: `www.nutridata.ee` (30.4.2016).

Poulsen, L. H. (2015). Suurandmed näitavad epideemia puhkemise kohta. — *Imeline teadus*, Nr 2, lk. 58–65.

Preis, T., Reith, D., and Stanley, H. E. (2010). Complex dynamics of our economic life on different scales: insights from search engine query data. — *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, Vol. 368, No. 1933, pp. 5707–5719. DOI: `10.1098/rsta.2010.0284`.

Provan, K. and Skinner, S. J. (1989). Interorganizational Dependence and Control as Predictors of Opportunism in Dealer-Seller Relations. — *Academy of Management Journal*, Vol. 32, No. 1, pp. 202–212. DOI: `10.2307/256427`.

Pukk, K. (2005). *Töörahulolu-uuringu läbiviimine Osaühingus Ehitus Service.* Ainetöö statistikas. Tallinn, Audentese Ülikool.

Pulliam, S. (1997). Bloated Inventories at Retailers May Mean Trouble for Investors. — *The Wall Street Journal, May 21*, p. C1.

Põldaru, R., Roots, J. ja Viira, A.-H. (2009). *Eesti põllumajanduse analüüs ja prognoos ökonomeetrilise modelleerimise abil.* Uuringuaruanne. Eesti Maaülikool.

Pärna, K. (2013). *Tõenäosusteooria algkursus.* Tartu: Tartu Ülikooli Kirjastus. 209 lk.

Randoja, P. ja Käerdi, H. (2011). Tulekahjudes hukkunute ajaline, piirkondlik ja sündmuskohaga seonduv statistiline analüüs. Kogumikus: *Sisekaitseakadeemia toimetised.* Kd. 5. Tallinn: Sisekaitseakadeemia, lk. 151–177.

Raudvere, R. (2011). Selgub, kui tark on eestlane. — *Maaleht 1. detsember 2011*, Nr 48, lk. 6.

Rezende, L. (2008). Econometrics of auctions by least squares. — *Journal of Applied Econometrics*, Vol. 23, No. 7, pp. 925–948. DOI: `10.1002/jae.1036`.

Robison, J. (2006). *Too Many Interruptions at Work?* Gallup Business Journal. URL: `http://www.gallup.com/businessjournal/23146/too-many-interruptions-work.aspx` (14.2.2015).

— (2014). *Why So Many New Companies Fail During Their First Five Years.* Gallup Business Journal. URL: `http://www.gallup.com/businessjournal/178787/why-new-companies-fail-during-first-five-years.aspx` (14.2.2015).

Román, F. J. (2011). A Case Study on Cost Estimation and Profitability Analysis at Continental Airlines. — *Issues in Accounting Education*, Vol. 26, No. 1, pp. 181–200. DOI: `10.2308/iace.2011.26.1.181`.

Roomets, S. (1999). *Arvjoonised.* Tln: Tallinna Pedagoogikaülikooli Kirjastus. 35 lk.

Rossman Allan, J. (1994). Televisions, Physicians, and Life Expectancy. — *Journal of Statistics Education*, Vol. 2, No. 2.

Rõõm, O. (2014). *Mis ja milleks on big data?* URL: http://lhv.delfi.ee/news/4781213?locale=et (2.2.2016).

Sander, P. (1999). *Portfelliteooria I.* Tartu: TÜ Kirjastus.

Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. — *Biometrics Bulletin*, Vol. 2, No. 6, pp. 110–114. DOI: 10.2307/3002019.

Sauga, A. (2011). New Product Diffusion in the Baltic States. — *Journal of Business Management*, Vol. 4, pp. 47–52.

Savur, S. R. (1937). "The use of the median in tests of significance". In: *Proceedings of the Indian Academy of Sciences — Section A*. Vol. 5. 6. Springer, pp. 564–576. DOI: 10.1007/BF03050163.

Scalas, E. et al. (2005). On the intertrade waiting-time distribution. — *Finance Letters*, Vol. 3, No. 1, pp. 38–43.

Scott, D. W. (1979). On optimal and data-based histograms. — *Biometrika*, Vol. 66, No. 3, pp. 605–610. DOI: 10.1093/biomet/66.3.605.

Servinski, M., Kivilaid, M. ja Tischler, G. (2009). *Linnad ja vallad arvudes 2009.* Eesti Statistikaamet.

Shao, J. et al. (2007). Quantitative relations between corruption and economic factors. — *The European Physical Journal B*, Vol. 56, No. 2, pp. 157–166. DOI: 10.1140/epjb/e2007-00098-2.

Singh, A. and Fagernäs, S. (2006). *Globalisation, Instability and Economic Insecurity.*

Sison, C. P. and Glaz, J. (1995). Simultaneous Confidence Intervals and Sample Size Determination for Multinomial Proportions. — *Journal of the American Statistical Association*, Vol. 90, No. 429, pp. 366–369. DOI: 10.1080/01621459.1995.10476521.

Squire, P. (1988). Why the 1936 Literary Digest Poll Failed. — *Public Opinion Quarterly*, Vol. 52, No. 1, pp. 125–133. DOI: 10.1086/269085.

Statistics New Zealand (2015). *Analytical retrospective superlative index based on New Zealand's CPI: 2014.* Available from www.stats.govt.nz.

Stedron, B. and Bínová, H. (2015). Economy and forecast for 2020: 3 key trends in the future. — *International Journal of Research in Engineering and Technology*, Vol. 4, pp. 271–276.

Stigler, G. J. (1961). The Economics of Information. — *Journal of Political Economy*, Vol. 69, No. 3, pp. 213–225. DOI: 10.1086/258464.

Stock, J. H. and Watson, M. W. (2003). *Introduction to Econometrics.* Addison-Wesley.

Sturges, H. A. (1926). The Choice of a Class Interval. — *Journal of the American Statistical Association*, Vol. 21, No. 153, pp. 65–66. DOI: 10.1080/01621459.1926.10502161.

Suits, D. B. (1958). The demand for new automobiles in the United States 1929–1956. — *The Review of Economics and Statistics*, Vol. 40, pp. 273–280.

Swartzman, G. (1970). The patient arrival process in hospitals: statistical analysis. — *Health services research*, Vol. 5, No. 4, pp. 320–329.

Zanden, J. van and Baten, J., eds. (2014). *How Was Life? Global Well-being since 1820.* OECD. DOI: 10.1787/9789264214262-en.

Taghizadegan, S. (2006). *Essentials of Lean Six Sigma.* Elsevier. 304 pp.

Tammeraid, I. (2004). *Tõenäosusteooria ja matemaatiline statistika.* Tln: TTÜ Kirjastus. 246 lk.

Taylor, G., ed. (2007). *Logistics Engineering Handbook.* CRC Press. 640 pp.

Tekkel, M. ja Veideman, T. (2013). *Eesti täiskasvanud rahvastiku tervisekäitumise uuring, 2012.* Uuringu raport. Tervise Arengu Instituut.

Terep, A. (2008). *Vedelkütuse ostuhinna riski juhtimine.* Magistritöö. TTÜ.

The 1993 Cars — Annual Auto Issue (1993). Consumer Reports. NY: Consumers Union.

Thompson, P. (2007). How Much Did the Liberty Shipbuilders Forget? — *Management Science*, Vol. 53, No. 6, pp. 908–918.

Thompson, W. R. (1936). On Confidence Ranges for the Median and Other Expectation Distributions for Populations of Unknown Distribution Form. — *The Annals of Mathematical Statistics*, Vol. 7, No. 3, pp. 122–128.

Thorndike, F. (1926). Applications of Poisson's probability summation. — *Bell System Technical Journal*, Vol. 5, No. 4, pp. 604–624. DOI: 10.1002/j.1538-7305.1926.tb00126.x.

Tiik, T. (2011). *Taksoteenuse kasutamise uurimuse tulemused.* Uuringu aruanne. MTÜ Tallinna Puuetega Inimeste Koda.

Tiit, E.-M. (1995). Rahvastiku prognoosist. Kogumikus: *Eesti Statistikaseltsi V Teabevihik.* TÜ Kirjastus, lk. 5–23.

— (2014). *2011. aasta rahva ja eluruumide loendus. Metoodika.* Eesti Statistikaamet.

Tiitso, R. (2015). *Eesti noorte säästmisharjumused.* Magistritöö. Tallinna Tehnikaülikool.

Tikva, P. ja Arnik, K. (2012). *2010. Leibkonna eelarve uuring. Metoodika.* Eesti Statistikaamet.

Toh, R. S. and Dekay, F. (2002). Hotel room-inventory management: an overbooking model. — *The Cornell Hotel and Restaurant Administration Quarterly*, Vol. 43, No. 4, pp. 79–90. DOI: 10.1016/S0010-8804(02)80044-1.

Total Diet Study. Elements Results Summary Statistics. Market Baskets 2006 through 2011 (2014). Tech. rep. U.S. Food and Drug Administration.

Traat, I. ja Inno, J. (1997). *Tõenäosuslik valikuuring.* Tartu Ülikooli kirjastus. 211 lk.

Täht, M. (2007). Aegridade sesoonne korrigeerimine. — *Eesti Statistika Kuukiri*, Kd. 5, lk. 153–164.

UNESCO (1990). *UNESCO Statistical Yearbook, 1990.*

Vainu, J. (2006). *Ökonomeetria. Lihtsad mudelid.* Tln: Külim. 174 lk.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. — *The Journal of Economic Perspectives*, Vol. 28, No. 2, pp. 3–27. DOI: 10.1257/jep.28.2.3.

Vella, F. and Verbeek, M. (1998). Whose wages do unions raise? A dynamic model of unionism and wage rate determination for young men. — *Journal of Applied Econometrics*, Vol. 13, No. 2, pp. 163–183. DOI: 10.1002/(SICI)1099-1255(199803/04)13:2<163::AID-JAE460>3.0.CO;2-Y.

Verganti, R. (1997). Order overplanning with uncertain lumpy demand: A simplified theory. — *International Journal of Production Research*, Vol. 35, No. 12, pp. 3229–3248. DOI: 10.1080/002075497194057.

Villsaar, K. jt (2014). *Ettevõtluse alustamise toetuse mõjuanalüüs*. Uuringu aruanne. Eesti Töötukassa.

Voog, A., Sarv, K. ja Männaste, K. (2012). *Välisturistide alkoholi ostumaht 2011. aastal*. Uuringu aruanne. TNS Emor.

Võrguteenuste kvaliteedinäitajad (2015). Elering. URL: http://elering.ee/vorguteenuste-kvaliteedinaitajad/ (14.2.2015).

Wang, H. (2008). Exact confidence coefficients of simultaneous confidence intervals for multinomial proportions. — *Journal of Multivariate Analysis*, Vol. 99, No. 5, pp. 896–911. DOI: 10.1016/j.jmva.2007.05.003.

Welch, B. L. (1951). On the Comparison of Several Mean Values: An Alternative Approach. — *Biometrika*, Vol. 38, No. 3/4, pp. 330–336. DOI: 10.2307/2332579.

Wessa, P. (2016). *Free Statistics Software*. Office for Research Development and Education. URL: http://www.wessa.net/.

Wheatley, J. J., Chiu, J. S. Y., and Stevens, A. C. (1980). Demographics to Predict Consumption. — *Journal of Advertising Research*, Vol. 20, No. 6, pp. 31–38.

Women and men in leadership positions in the European Union, 2013 (2013). Report. European Commission, Directorate-General for Justice. DOI: 10.2838/50821.

Wooldridge, J. M. (2002). *Introductory Econometrics: A Modern Approach*. 2nd. South-Western. 900 pp.

Worthington, A. C. (2006). Debt as a source of financial stress in Australian households. — *International Journal of Consumer Studies*, Vol. 30, No. 1, pp. 2–15. DOI: 10.1111/j.1470-6431.2005.00420.x.

Wright, T. P. (1936). Factors affecting the cost of airplanes. — *Journal of the Aeronautical Sciences*, Vol. 3, No. 4, pp. 122–128. DOI: 10.2514/8.155.

Übi, E. ja Keres, K. (2013). *Rakendusmatemaatika: õpik kõrgkoolidele*. Tln: TTÜ Kirjastus. 380 lk.

Üliõpilaste sotsiaalmajanduslik olukord 2005/2006 (2007). Küsitluse kokkuvõte. Klaster.

Xin, Y. et al. (2012). Energy consumption quota of four and five star luxury hotel buildings in Hainan province, China. — *Energy and Buildings*, Vol. 45, pp. 250–256. DOI: 10.1016/j.enbuild.2011.11.014.

Yang, Y. (2015). Development of the regional freight transportation demand prediction models based on the regression analysis methods. — *Neurocomputing*, Vol. 158, No. 0, pp. 42–47. DOI: 10.1016/j.neucom.2015.01.069.

Õpik on eeskätt mõeldud majanduserialade bakalaureuseõppes õppijaile, kuid sobib kasutamiseks kõigile, kellel on vaja teha statistilistele andmetele tuginevaid järeldusi. Tutvustatakse põhilisi statistilisi meetodeid ja nende kasutamist mitmesuguste äri- ning majandusalaste küsimuste lahendamisel. Antakse juhiseid statistiliseks andmetöötluseks tabelarvutusprogrammides Excel ja LibreOffice Calc. Õpikus on ligi 200 näidet, millest enamikuga on võimalik tutvuda ka õpiku juurde kuuluvates tabelarvutuse failides, kus on arvutused koos selgitavate kommentaaridega. Iseseisvaks lahendamiseks on üle 400 ülesande, millest ligikaudu pool tuleb lahendada tabelarvutuses, kasutades ülesannete failides olevaid andmeid. Kõik ülesanded on varustatud vastustega. Valdav enamik näidetest ja ülesannetes kasutatud andmetest pärineb reaalsest elust.

Autorist

Ako Sauga on lõpetanud Tartu Ülikooli füüsikaosakonna 1983. aastal, alates 2006. aastast ökoloogia doktor. Töötanud Tallinna Tehnikaülikooli füüsika instituudis, erakõrgkoolis Audentese Ülikool, Tallinna Ülikoolis ja aastast 2007 töötab Tallinna Tehnikaülikooli majandusteaduskonnas. Lugenud loenguid statistika, ökonomeetria, majandusmatemaatika, kvantitatiivsete meetodite ja füüsika vallast, osalenud gümnaasiumi matemaatikaalaste õppevahendite koostamisel ning viinud läbi matemaatikaõpetajate täienduskoolitusi. Avaldanud teadustöid stohhastiliste protsesside, populatsioonidünaamika ja mõjuanalüüsi kohta.