
**Tõenäosusteooria ja matemaatilise statistika alused**

**Eessõna**

Käesoleva kursuse aluseks on 2002/2003. õppeaasta sügissemestril Tartu Ülikoolis Matemaatika-Informaatikateaduskonna ja Füüsika-Keemiateaduskonna infotehnoloogia eriala üliõpilastele loetud kursus. Selle kursuse maht oli 2 punkti, seega hõlmas ta 8 2-tunnilist loengut ja sama palju praktikume, millest osa toimus arvutiklassis. Sellele vastabki õppevahendi struktuur: iga peatükk sisaldab täpselt üht loengut.

Et kursust loeti täppisteaduste esindajatele, on siin (erinevalt väikesemahuliste kursuste tavapraktikast) esitatud ka mõningad tõestused ja matemaatilised põhjendused, kuigi kõigi tõestuste esitamine nii väiksemahulise kursuse puhul pole mõeldav.

Kursuse eesmärgiks oli tutvustada niihästi matemaatilise statistika kui ka tõenäosusteooria põhitõdesid nii palju, et selle kursuse läbinud kuulajad suudaksid edaspidi vajalikus kirjanduses orienteeruda ja leida iseseisvalt konkreetsetele ülesannetele lahendusi. Loodetavasti aitab seda eesmärki saavutada ka käesolev CD-plaat, kusjuures sihtrühmaks võiksid olla erinevate erialade õppurid, kes on kuulanud suhteliselt väikese mahuga tõenäosusteooria, matemaatilise statistika, statistika jaS andmeanalüüsi kursusi.

*E.-M. Tiit, okt. 2003.*

Sisukord


TAGASI

**Sisukord**


**Tõenäosusteooria ja matemaatilise statistika alused**

**Eessõna**

Käesoleva kursuse aluseks on 2002/2003. õppeaasta sügissemestril Tartu Ülikoolis Matemaatika-Informaatikateaduskonna ja Füüsika-Keemiateaduskonna infotehnoloogia eriala üliõpilastele loetud kursus. Selle kursuse maht oli 2 punkti, seega hõlmas ta 8 2-tunnilist loengut ja sama palju praktikume, millest osa toimus arvutiklassis. Sellele vastabki õppevahendi struktuur: iga peatükk sisaldab täpselt üht loengut.

Et kursust loeti täppisteaduste esindajatele, on siin (erinevalt väikesemahuliste kursuste tavapraktikast) esitatud ka mõningad tõestused ja matemaatilised põhjendused, kuigi kõigi tõestuste esitamine nii väiksemahulise kursuse puhul pole mõeldav.

Kursuse eesmärgiks oli tutvustada niihästi matemaatilise statistika kui ka tõenäosusteooria põhitõdesid nii palju, et selle kursuse läbinud kuulajad suudaksid edaspidi vajalikus kirjanduses orienteeruda ja leida iseseisvalt konkreetsetele ülesannetele lahendusi. Loodetavasti aitab seda eesmärki saavutada ka käesolev CD-plaat, kusjuures sihtrühmaks võiksid olla erinevate erialade õppurid, kes on kuulanud suhteliselt väikese mahuga tõenäosusteooria, matemaatilise statistika, statistika jaS andmeanalüüsi kursusi.

*E.-M. Tiit, okt. 2003.*

Sisukord


TAGASI

**Sündmus. Klassikaline ja geomeetriline tõenäosus**


TAGASI

**Statistiline tõenäosus ja suurte arvude seadus. Sündmuste sõltuvus**

## STATISTILINE TÕENÄOSUS

Tõenäosuse üldine mõiste

Katseseeria ja statistiline tõenäosus

Statistilise tõenäosuse omadused

Statistilise tõenäosuse juhuslikkus

Tõenäosuse järgi koondumise graafiline pilt

Suurte arvude seadus

Suurte arvude seaduse seos statistilise tõenäosusega

Järelemõtlemiseks

## TINGLIK TÕENÄOSUS JA SÜNDMUSTE SÕLTUVUS

Tingimus

Tinglik tõenäosus

Sündmuste korrutise tõenäosus

Sündmuste sõltumatus

Sündmuste sõltuvus

Järelemõtlemiseks

## BAYESI TEOREEM

Täistõenäosuse valem

Bayesi valem

Järelemõtlemiseks


TAGASI

**Juhuslik suurus ja vektor. Jaotus ja tema esitused**

## JUHUSLIK SUURUS JA VEKTOR

Juhusliku suuruse määratlus
Juhusliku suuruse jaotus ja tõenäosusfunktsioon
Juhusliku suuruse abil defineeritud sündmused
Empiiriline jaotus
Juhuslik vektor ja selle jaotus
Juhusliku vektori komponentide sõltumatus
Juhusliku suuruse funktsioon
Järelemõtlemiseks

## DISKREETSED JAOTUSSEADUSED

Teoreetiline diskreetne jaotus
Bernoulli jaotus
Binoomjaotus
Diskreetne ühtlane jaotus
Hüpergeomeetriline jaotus
Geomeetriline jaotus
Hüpergeomeetriline jaotus
Poissoni jaotus
Multinomiaaljaotus
Järelemõtlemiseks

## PIDEVAD JAOTUSSEADUSED

Jaotusfunktsioon
Tihedusfunktsioon
Ühtlane jaotus
Eksponentjaotus
Järelemõtlemiseks


TAGASI

# Juhusliku suuruse jaotusparameetrid


TAGASI

**Normaaljaotus ja tsentraalne piirteoreem. Lineaarne korrelatsioonikordaja**

## NORMAALJAOTUS

Normaaljaotuse tihedusfunktsioon
Normaaljaotuse keskväärtus
Normaaljaotuse dispersioon
Normaaljaotuse jaotusfunktsioon
Normaaljaotuse teised arvkarakteristikud
Normaaljaotuse lineaarfunktsiooni jaotus
Normaaljaotuse standardiseerimine
Normaaljaotuse tabel
Normaaljaotuse kaudu defineeritud sündmuste tõenäosuste leidmine
Järelemõtlemiseks

## TSENTRAALSED PIIRTEOREEMID

Piirteoreemi mõiste
De Moivre'-Laplace'i piirteoreem
Üldine klassikaline tsentraalne piirteoreem
Poissoni piirteoreem
Järelemõtlemiseks

## NORMAALJAOTUSEGA JUHUSLIK VEKTOR. KORRELATSIOONIKORDAJA

Normaaljaotusega juhusliku vektori mõiste ja omadused
Kahemõõtmeline normaaljaotusega juhuslik vektor
Korrelatsioonikordaja
Järelemõtlemiseks


TAGASI

# Üldkogum ja valim. Hindamine

## ÜLDKOGUM JA VALIM

Matemaatilise statistika põhiülesanne
Üldkogum ja valim
Teoreetiline ja konkreetne valim
Matemaatilise statistika põhiülesande täpsustus
Valimi jaotus
Järelemõtlemiseks

## PUNKTHINNANG

Jaotusparameetrite hindamise ülesanne
Punkt- ja vahemikhinnang
Punkthinnangu arvutamine valimi põhjal
Hinnangu keskväärtus. Nihe
Dispersiooni hinnang. Nihke kõrvaldamine
Asümptootiline nihketus
Hinnangu hajuvus ja dispersioon
Standardviga ja suhteline viga
Hinnangu efektiivsus
Hinnangu mõjusus
Tõenäosuse hinnang ja hinnangu viga
Hindamismeetodid
Järelemõtlemiseks

## VAHEMIKHINNANG

Usalduspiirid ja usaldustõenäosus
Usalduspiiride konstrueerimine
Normaaljaotuse keskväärtuse usalduspiiride määramine normaaljaotuse abil
Põhistatistikute jaotuste defineerimine
t-statistiku jaotus
Normaaljaotuse keskväärtuse usalduspiiride määramine t-statistiku abil
Järelemõtlemiseks

# Statistiliste hüpoteeside kontrollimine

## STATISTILISE HÜPOTEESIDE KONTROLLIMISE TEOORIA PÕHIMÕISTED

Millal ja milleks on tarvis kontrollida statistilisi hüpoteese?
Statistiliste hüpoteeside paar
Vead statistiliste hüpoteeside kontrollimisel
Olulisuse nivoo ja võimsus
Hüpoteesi kontrollimise eeskirja (kriteeriumi) konstrueerimine ja otsuse vastuvõtmine
Seos hüpoteeside kontrollimise ja vahemikhindamise vahel
Järelemõtlemiseks

## HÜPOTEESIDE KONTROLLIMINE ÜLDKOGUMI KESKVÄÄRTUSE KOHTA

Kõige sagedamini kontrollitavad hüpoteesid ühe keskväärtuse kohta
Ühepoolse hüpoteesi $EX>c$ kontrollimine normaaljaotuse keskväärtuse kohta
Ühepoolse hüpoteesi $EX<c$ kontrollimine normaaljaotuse keskväärtuse kohta
Kahepoolse hüpoteesi $EX\neq c$ kontrollimine normaaljaotuse keskväärtuse kohta
Näited
Tulemuste üldistamine juhule, kui lähtejaotus erineb normaaljaotusest
Kahe normaaljaotuse keskmiste võrdlemine (sõltuvad vaatlused)
Kahe normaaljaotuse keskmiste võrdlemine (sõltumatud vaatlused)
Ühepoolse ja kahepoolse hüpoteesi vahekord
Järelemõtlemiseks

## MÕNINGATE MUUDE STATISTILISTE HÜPOTEESIDE KONTROLLIMINE

Hüpoteeside kontrollimine korrelatiivse seose kohta
Hüpoteeside kontrollimine jaotuste erinevuse kohta (jaotuste võrdlemine)
Empiirilise ja teoreetilise jaotuse võrdlemine $\chi^2$-statistiku abil
Kahe jaotuse võrdlemine $\chi^2$-statistiku abil
Tõenäosuste võrdlemine $\chi^2$-statistiku abil
Järelemõtlemiseks

**Statistiline sõltuvus ja statistiline mudel**

## STATISTILINE SÕLTUVUS

### Statistiline sõltuvus kahe tunnuse vahel

Matemaatilises statistikas nimetatakse juhuslikke suurusi sageli ***tunnusteks***. Enamasti on ühe korraga vaadeldavad tunnused mõõdetud samal objektide hulgal, seega moodustavad nad juhusliku vektori ehk ***tunnusvektori***. Kui juhusliku vektori komponendid ei ole sõltumatud, siis on nad sõltuvad, ***nende vahel on statistiline sõltuvus***.

Tunnused $X$ ja $Y$ on ***statistiliselt sõltumatud***, kui nad ei ole statistiliselt sõltumatud, st kui iga x ja y puhul kehtib võrdus

$$F_{XY}(x,y)=F_X(x)F_Y(y),$$

kus tunnuste $X$ ja $Y$ jaotusfunktsioonid on vastavalt

$$F_X(x),\ F_Y(y)$$

ning tunnusvektori $(X,\ Y)$ jaotusfunktsioon on $F_{XY}(x,y)$. Tunnused $X$ ja $Y$ on ***statistiliselt sõltuvad***, kui võrdus (1) ei kehti kõigi argumentide korral, st kui leidub mingi selliste väärtuste paar $(x,\ y)$, et võrdus (1) ei kehti.

Statistiline sõltuvus on statistika üks põhimõisteid. Tunnuste vahel on statistiline sõltuvus siis, kui ühe tunnuse käitumine (jaotus) sõltub teise tunnuse väärtustest. Statistiline sõltuvus on võimalik niihästi pidevate, diskreetsete kui ka pidevate ja diskreetsete tunnuste vahel.

•Diskreetsed tunnused $X$ ja $Y$ on statistiliselt sõltumatud, kui iga $i$ ja $j$ korral kehtib võrdus:

$$P(X=x_i,\ Y=y_j)=\ P(X=x_i)P(Y=y_j)\ \text{ehk}\ p_{ij}=p_i\cdot p_j,$$

kus $p_{ij}$ tähistab ühisjaotuse ning $p_i$ ja $p_j$ vastavalt marginaaljaotuste tõenäosusfunktsioone.

•Pidevad tunnused $X$ ja $Y$ on statistiliselt sõltumatud, kui iga $x$ ja $y$ korral kehtib võrdus:

$$f_{XY}(x,y)=\ f_X(x)f_Y(y),$$

kus $f_{XY}(xy)$ tähistab ühisjaotuse ning $f_X(x)$ ja $f_Y(y)$ vastavalt marginaaljaotuste tihedusfunktsioone.

Statistilise sõltuvuse mõiste on kasutatav ka üldisema tunnuse mõiste korral. Tunnuste väärtused ei tarvitse olla mitte üksnes arvud, vaid nendeks võivad olla ka muud objektid/ omadused. Oluline on see, et iga katsetulemuse korral on tunnuse väärtus üheselt määratud. Edaspidi vaatleme mõningates näidetes ka mittearvuliste väärtustega tunnuseid, kasutades nende puhul sama sõltumatuse definitsiooni nagu diskreetsete arvtunnuste korral.

***Näide 1.*** Lisatud tabel sisaldab andmeid kolmes vanuses õpilaste kontrolltöö hinnete kohta.

| Hinne | | | | | |
|---|---|---|---|---|---|
| Vanus | 2 | 3 | 4 | 5 | Kokku |
| 10 | 1 | 3 | 7 | 4 | 15 |
| 11 | 3 | 9 | 21 | 12 | 45 |
| 12 | 2 | 6 | 14 | 8 | 30 |
| Kokku | 6 | 18 | 42 | 24 | 90 |

Järgmises tabelis on antud tunnuste “Vanus” ja “Hinne” ühisjaotus ja kummagi tunnuse marginaaljaotused. Nende abil on põhimõtteliselt võimalik kontrollida, kas sõltumatuse tingimus on täidetud, kuid selleks tuleks kontrollida 12 võrduse kehtivust:

$$0{,}167 \times 0{,}067 = 0{,}011, \ldots \text{jne.}$$

Oluline on see, et ka siis, kui üksainus võrdus ei kehti, on tunnused sõltuvad.

| Hinne | | | | | |
|---|---|---|---|---|---|
| Vanus | 2 | 3 | 4 | 5 | Tingliku jaotuse summa |
| 10 | 0,011 | 0,033 | 0,078 | 0,044 | 0,167 |
| 11 | 0,033 | 0,100 | 0,233 | 0,133 | 0,500 |
| 12 | 0,022 | 0,067 | 0,156 | 0,089 | 0,333 |
| Hinde marginaaljaotus | 0,067 | 0,200 | 0,467 | 0,267 | 1,000 |

**Ühe tunnuse tinglikud jaotused teise tunnuse suhtes**

Leiame nüüd olemasolevate andmete põhjal eraldi 10-, 11- ja 12-aastaste õpilaste kontrolltööhinnete jaotuse, kasutades selleks ***tingliku tõenäosuse valemit***, kusjuures tingimuseks on õpilase vanus:

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

Esitame arvutustulemused alljärgnevas tabelis, mis sisaldab ***ridadena hinde tinglikke jaotusi sõltuvalt õpilase vanusest***:

| Hinne | | | | | |
|---|---|---|---|---|---|
| Tingimus | 2 | 3 | 4 | 5 | Vanuse marginaaljaotus |
| 10-aastaste hinde jaotus | 0,067 | 0,200 | 0,467 | 0,267 | 1,000 |
| 11-aastaste hinde jaotus | 0,067 | 0,200 | 0,467 | 0,267 | 1,000 |
| 12-aastaste hinde jaotus | 0,067 | 0,200 | 0,467 | 0,267 | 1,000 |
| Hinde marginaaljaotus | 0,067 | 0,200 | 0,467 | 0,267 | 1,000 |

Samal viisil võime arvutada erineva hinde saanud õpilaste vanuse tinglikud jaotused, mis paiknevad alljärgnevas tabelis veergu

| Tingimus | | | | | |
|---|---|---|---|---|---|
| Vanus | Hinde 2 saanute vanuse-jaotus | Hinde 3 saanute vanuse-jaotus | Hinde 4 saanute vanuse-jaotus | Hinde 5 saanute vanuse-jaotus | Vanuse marginaal-jaotus |
| 10-aastased | 0,167 | 0,167 | 0,167 | 0,167 | 0,167 |
| 11-aastased | 0,500 | 0,500 | 0,500 | 0,500 | 0,500 |
| 12-aastased | 0,333 | 0,333 | 0,333 | 0,333 | 0,333 |
| | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |

Kui tunnused on sõltumatud, siis tuleneb üldisest valemist vahetult alljärgnev järeldus, mida illustreerivad ka lisatud näited.

| |
|---|
| Sõltumatute tunnuste $X$ ja $Y$ korral ühtivad alati tunnuse $X$ tinglikud jaotused sama tunnuse marginaaljaotusega, kui tingimus on määratud tunnusega $Y$. |

See seos kehtib igatüüpi tunnuste korral. Viimast tingimust kasutades on lihtne defineerida ka pideva ja diskreetse tunnuse vahelist sõltuvus.
•Kui $X$ on diskreetne ja $Y$ pidev juhuslik suurus, siis on $X$ ja $Y$ sõltumatud parajasti siis, kui $Y$ tinglik tihedusfunktsioon $f_Y(y|x)$ ei sõltu $X$ väärtusest $x$.

Esitatud seos on kasulik ka selleks, et ilma arvutusteta tunda ära sõltumatute tunnuste ühisjaotust – selgub, et ***sõltumatute tunnuste ühisjaotuses on nii***

***read kui ka veerud võrdelised***.

Statistiline sõltuvus erineb ***põhjuslikust sõltuvusest*** selle poolest, et ***statistiline sõltuvus on vastastikune –*** kui tunnus *X* sõltub tunnusest *Y*, siis sõltub ka *Y* tunnusest *X*. Näiteks kui lapse pikkus sõltub põhjuslikult oluliselt tema vanusest, kuid pikkuse muutus ei põhjusta sisuliselt kuidagiviisi vanuse muutumist, siis statistiline sõltuvus pikkuse ja vanuse vahel on vastastikune.

**Statistilise sõltuvuse / sõltumatuse tähtsus**

1.Sõltumatud vaatlused sisaldavad ***maksimaalselt teavet*** uuritava objekti kohta, selletõttu jälgitakse valimit moodustades enamasti seda, et vaadeldavad objektid oleksid sõltumatud (näiteks ei võeta mõnikord valimisse sama leibkonna liikmeid).

2.Samadel objektidel mõõdetud erinevate tunnuste statistilise sõltuvuse abil on võimalik mitmetes eluvaldkondades ilmnevaid sõltuvusi/ nähtusi ***modelleerida***.

3.Kui eesmärgiks on ***konstrueerida mudel***, siis selgitatakse, kas uuritav juhuslik suurus ***sõltub statistiliselt*** mudeli argumentidest – vastasel korral on mudeli konstrueerimine mõttetu ja võimatu.

4.Mida ***tugevam*** on kahe tunnuse vaheline statistiline sõltuvus, seda rohkem teavet sisaldab üks tunnus teise kohta ja seda paremini (täpsemalt) on üht tunnust kasutades võimalik teise tunnuse väärtusi ***prognoosida***.

5.Statistiline sõltuvus on teatavas mõttes üldistuseks funktsionaalse sõltuvuse mõistele, statistilise sõltuvuse tugevust mõõdetakse tema läheduse kaudu funktsionaalsele sõltuvusele.

**Statistilise sõltuvuse tugevus ja olulisus**

Kui statistikas kõneldakse sõltuvusest kahe tunnuse vahel, siis tekib kaks tähtsat küsimust:

•Kui *tugev* see sõltuvus on? (ehk – kas seda sõltuvust saab kasutada mudelite koostamiseks ja tunnuse väärtuste ennustamiseks?);

•Kas valimi põhjal avastatud sõltuvus kehtib ka üldkogumis, ehk kas see sõltuvus on statistiliselt *oluline*?

Need küsimused on küll omavahel seotud, kuid ei ühti, sest sõltuvuse olulisus tuleneb mitte ainult valimi põhjal leitud sõltuvuse tugevusest, vaid ka valimi mahust. Selletõttu võivad suurte valimite puhul leitud suhteliselt nõrgad sõltuvused olla statistiliselt olulised, st kajastada vastavaid sõltuvusi üldkogumis. Seevastu väikeste valimite puhul võivad isegi suhteliselt tugevad sõltuvused kajastada üksnes valimi iseärasusi ning mitte olla üldistatavad üldkogumile.

**Statistilise seose tugevust iseloomustavad seosekordajad**

Statistilise sõltuvuse mõõtmiseks kasutatakse statistilise seose kordajaid, mida nimetatakse ka assotsiatsioonikordajateks või kontingentsuse kordajateks. Neid on erinevate autorite poolt defineeritud mitmeti, kuid (peaaegu) kõigil on teatavad ühisjooned.

1.Statistilise seose kordaja väärtus muutub 0 ja 1 vahel.

2.Kui tunnused on statistiliselt sõltumatud, siis on seosekordaja väärtus 0.

3.Kui tunnuste vahel on täielik statistiline sõltuvus, siis on seosekordaja väärtus 1.

4.Tugevamale seosele vastab suurem seosekordaja väärtus.

Statistilise seose kordajad arvutatakse valimi põhjal, nad on valimi statistikud. Ühe osa puhul neist on teada ka neile vastavate statistikute jaotus, selliseid seosekordajaid saab kasutada ka seose olulisuse kontrollimiseks.

**Täielik statistiline sõltuvus**

Selgitamist vajab see, mida mõista täieliku statistilise sõltuvuse vahel. Siin on kaks võimalust:

a.***Täielik vastastikune sõltuvus***, mis seisneb selles, et teades ühe tunnuse väärtust võib täpselt öelda ka teise tunnuse väärtuse ja vastupidi. Täielik vastastikune sõltuvus saab aset leida üksnes tunnuste vahel, millel on võrdne arv väärtusi (väärtusklasse) ja sel juhul on ühisjaotuse jaotustabelis igas reas ja igas veerus ainult üks nullist erinev arv (st et kõik tinglikud jaotused on mittejuhuslikud ja kõdunud konstandiks).

b.***Täielik ühepoolne sõltuvus***, mis seisneb selles, et teades ühe tunnuse väärtust võib täpselt öelda ka teise tunnuse väärtuse, kuid mitte vastupidi. Täieliku ühepoolse sõltuvuse korral on tunnuste väärtuste arvud erinevad, ühe tunnuse tinglikud jaotused on konstandiks kõdunud, teise omad aga mitte. Alati saab suurema väärtuste arvuga tunnuse väärtust teades täpselt öelda väiksema tunnuste arvuga tunnuse väärtus, ent mitte vastupidi.

***Näide 2.***

Kolmes erisuunitlusega klassis tehti õpilastele kolm testi, mis mõõtsid nende teadmisi ja oskusi erinevates valdkondades (vt tabel, kus on märgitud testi positiivselt sooritanud õpilaste arv).

| Testi tulemus | I blokk | II blokk | III blokk |
|---|---|---|---|
| Täppisteaduste klass | 20 | 0 | 0 |
| Loodusteaduste klass | 0 | 35 | 0 |
| Kunstiklass | 0 | 0 | 15 |

1.Tabelis on võrdne arv ridu ja veerge ning igas reas ja igas veerus on ainult üks nullist erinev sagedus. Sel juhul on kehtivad järgmised asjaolud:

•Iga esimese tunnuse väärtusega esineb koos ainult üks teise tunnuse väärtus.

•Kummagi tunnuse tinglikud jaotused teise tunnuse suhtes on mittejuhuslikud.

•Tunnuste vahel on üks-ühene vastavus, st et ühe tunnuse väärtust teades on võimalik täpselt määrata ka teise tunnuse väärtus.

***Näide 3.***

Viies erinevas klassis tehti õpilastele kolm testi, mis mõõtsid nende teadmisi ja oskusi erinevates valdkondades (vt tabel, kus on märgitud testi positiivselt sooritanud õpilaste arv).

Esitatud näites on tegemist mittearvuliste väärtustega tunnustega – ühe tunnuse väärtuseks on klassi nimetus või tüüp, teise väärtuseks – testiblokk.

| Lemmikharrastus | I blokk | II blokk | III blokk |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Täppisteaduste klass A | 20 | 0 | 0 |
| Loodusteaduste klass A | 0 | 35 | 0 |
| Kunstiklass | 0 | 0 | 15 |
| Täppisteaduste klass B | 10 | 0 | 0 |
| Loodusteaduste klass B | 0 | 15 | 0 |

Tabelis on ridu rohkem kui veerge, seega ei saa tunnuste vahel olla üks-ühest vastavust. Näeme, et klassi järgi on võimalik üheselt ennustada testitulemust, kuid vastupidine pole üldiselt võimalik: teades, et juhuslikult valitud õpilane tegi edukalt testibloki II, pole teada, kas ta käib loodusteaduste klassis A või B.

**Statistilise sõltuvuse olulisus**

Tunnuste *X* ja *Y* vahelist statistilist sõltuvust nimetatakse ***oluliseks*** (olulisuse nivool $\alpha$) siis, kui olulisuse nivool $\alpha$ õnnestub tõestada sisukas hüpotees $H_1$ alljärgnevast hüpoteesipaarist:

- $H_1$: Tunnused *X* ja *Y* ei ole üldkogumis sõltumatud.
- $H_0$: Tunnused *X* ja *Y* on üldkogumis sõltumatud.

Üks võimalusi statistilise sõltuvuse olulisuse tõestamiseks on hii-ruut statistiku kasutamine sarnaselt eelmises loengus kasutatud juhuga. Selleks valime ***teoreetiliseks jaotuseks*** ühisjaotuse, mis tekiks siis, kui vaadeldavate tunnuste empiirilised jaotused oleksid sõltumatud:

$$P(X = x_i,\ Y = y_j) = \frac{k_i}{n} \cdot \frac{k_j}{n}.$$

Sel juhul tuleks eelmises loengus kasutusele võetud hii-ruut-statistiku

$$H = \sum_{j=1}^{2} \sum_{i=1}^{m_j} \frac{(k_{ji} - n_j p_i)^2}{n_j p_i}$$

tuletamisega sarnase mõttekäigu abil saadud statistiku kuju järgmine:

$$H = \frac{1}{n} \sum_{j=1}^{m'} \sum_{i=1}^{m_2} \frac{(n\, k_{ij} - k_i k_j)^2}{k_i k_j}.$$

Selle statistiku asümptootiline jaotus nullhüpoteesi õigsuse korral on hii-ruut jaotus vabadusastmete arvul $(m_1 - 1)(m_2 - 1)$. Ühtasi on selge, et erijuhul, kui empiiriline jaotus on sõltumatu, on statistiku *H* väärtus 0. Hüpotees $H_0$ kummutatakse ja sisukas hüpotees võetakse vastu siis, kui statistiku *H* väärtus on nii suur, et sellele vastav olulisuse tõenäosus on suurem kui $\alpha$.

Kuna statistiku *H* jaotus on teada asümptootiliselt, siis tuleb selleks, et kasutada statistikut *H* statistilise seose olulisuse kontrollimiseks, jälgida seda, et sagedused $k_{i.}$ ja $k_{.j}$ ei oleks liiga väiksed (rusikareegel on, et igasse tabeli lahtrisse peaks teoreetiliselt sattuma vähemalt 2–5 vaatlust). Vastasel korral ei tarvitse tehtud järeldused kehtida.

**Statistilise sõltuvuse tugevus. Crameri seosekordaja**

Selleks, et teatava statistiku abil statistilise sõltuvuse tugevust mõõta, on otstarbekas teha selgeks, missugune on tema väärtus täieliku statistilise sõltuvuse korral. Arvutame hii-ruut statistiku väärtuse täieliku statistilise sõltuvuse korral, lähtudes alljärgnevast tabeli kujust, kus marginaaljaotused on diskreetese ühtlase jaotusega ja $n = km$:

| | 1 | 2 | ... | *m* | |
|---|---|---|---|---|---|
| 1 | *k* | | ... | | *k* |
| 2 | | *k* | ... | | *k* |
| ... | ... | ... | ... | ... | ... |
| *m* | | | ... | *k* | *k* |
| | *k* | *k* | ... | *k* | *n* |

Saame siis hii-ruut-statistiku väärtuseks:

$$H = \frac{m}{n}\left[\frac{(nk - k^2)^2}{k^2} + (m-1)\frac{(0 - k^2)^2}{k^2}\right] = \frac{m}{n}\left[k^2(m-1)^2 + (m-1)k^2\right] = m(m-1)k =$$
$$= n(m-1).$$

Loomulik on määrata seosekordaja väärtus nii, et tabelis esitatud lihtsaima kujuga sõltuvuse korral omandab ta väärtuse 1. Siit tulenebki üks populaarsemaid statistilise seose kordajaid, Crameri kordaja, mis on juhul, kui *X* ja *Y* väärtuste arv on sama, defineeritud seosega

$$V = \sqrt{\frac{H}{n(m-1)}}.$$

Crameri seosekordaja mõõdab statistilise seose tugevust, tema väärtused muutuvad 0 ja 1 vahel, kusjuures 0 vastab sõltumatule ja 1 täielikult sõltuvale (üks-üheses vastavuses olevale) kahe tunnuse ühisjaotusele. Tuleb aga tõdeda, et olukord, kus valimi põhjal moodustatud empiiriline jaotus on sõltumatu, on väga haruldane. Kui aga empiirilise jaotuse puhul on tegemist nõrga seosega, võib teoreetiline jaotus olla sõltumatu.

**Üldisemad statistilise seose kordajad**

Kui tunnustel *X* ja *Y* on väärtuste arvud erinevad, vastavalt $m_1$ ja $m_2$, siis kasutatakse Crameri seosekordaja jaoks avaldist,

$$V = \sqrt{\frac{H}{n\,min((m_1-1),(m_2-1))}},$$

mille muutumispiirkond on sama ja mis näitab ühepoolse seose tugevust (kui hästi on lühema skaalaga tunnus prognoositav pikema skaalaga tunnuse järgi).

Teine Crameri kordaja üldistus on Tšuprovi seosekordaja

$$T = \sqrt{\frac{H}{n\sqrt{(m_1-1)(m_2-1)}}}.$$

Selle valemi järgi arvutatud kordaja maksimaalne väärtus on 1 ainult sel juhul, kui $m_1 = m_2$.

Nende seosekordajate eeliseks on see, et nende arvutamisel kasutatakse hii-ruut-statistikut, mille abil on võimalik kontrollida statistilise seose olulisust.

**Järelemõtlemiseks**

1.Kuidas muutub statistiline sõltuvus siis, kui ühisjaotuse tabelis ridu ja veerge vahetada?

2.Kuidas muutub statistilise seose tugevus siis, kui kõiki sagedusi korrutada konstandiga $c$?

3.Kuidas muutub statistilise seose olulisus siis, kui kõiki sagedusi korrutada konstandiga $c > 1$?

4.Olgu ühel tunnusel m väärtust, teisel $2m$ väärtust. Kumma tunnuse (ühepoolne) sõltuvus teisest on üldjuhul tugevam?

**Sündmus. Klassikaline ja geomeetriline tõenäosus**

## SÜNDMUSE MÕISTE

### Mis on stohhastika?

Stohhastika on teadus juhuslikkusest, mis sisaldab niihästi tõenäosusteooria, matemaatilise statistika kui ka juhuslike protsesside teooria elemente, kuid ka nende edasiarendusi. Käeolev kursus sisaldab kaht osa – tõenäosusteooriat ja matemaatilist statistikat.

### Katse ja elementaarsündmus

Juhuslikkuse mõiste aluseks on katse, so tegevus, mille tulemus ei ole ette teada. Tõenäosusteoorias eeldatakse, et

•katse tingimused on täpselt fikseeritud;

•katse on (lõpmata palju kordi) korratav samade tingimuste püsides;

•võimalike katsetulemuste hulk on ette teada.

Katsetulemust nimetatakse ***elementaarsündmuseks.*** Kõigi katsetulemuste hulk $\Omega$ moodustab ***elementaarsündmuste ruumi***.

**Näited:**

1. Lifti astub rühm tudengeid ja sõidab neljandale korrusele. Katse korraldaja loendab, mitu tudengit on liftis. Selgub, et **tudengite arv liftis on juhuslik**.
2. Katse korraldajat huvitavad kuivaperioodide pikkused Tartus. Ta märgib iga vihmasaju alguse ja lõpu aja ning arvutab kuivaperioodi kestuse. Selgub, et ***kuivaperioodi kestus on juhuslik***.
3. Reisija, kes bussi sõidugraafikut ei tea, tuleb bussipeatusesse ja saab teda, et busside liikumise intervall on 12 minutit. Tema ***ooteaja pikkus on juhuslik***.
4. Väga lihtne katse on täringuvise. ***Täringuviskel saadud silmade arv on juhuslik***.

Neljanda katse puhul on selge, et katsel on kuus võimalikku tulemust ehk elementaarsündmust. Ka esimesel katsel on lõplik arv võimalikke tulemusi. Teisel ja kolmandal katsel on aga võimalike katsetulemuste arv lõpmata suur.

### Sündmus

Sündmused moodustuvad elementaarsündmustest, sündmused on elementaar-sündmuste hulgad.

**Näited:**

1. Esimese katse puhul võime määratleda mitmesuguseid sündmusi, näiteks *A* -- liftis sõidab üksainus tudeng; *B* – liftis sõidab vähemalt kümme tudengit; *C* – liftis

sõidab paarituarv tudengeid; *D* – liftis sõidab viis kuni kümme tudengit.

2. Teise katse puhul võime määratleda lõpmata palju erinevaid sündmusi, näiteks – kuivaperiood kestis alla ühe päeva; kuivaperiood kestis kaks kuni kolm nädalata, kuivaperiood kestis üle kuu aja jne.

3. Samuti on kolmanda katse korral võimalikke sündmusi sama palju kui erineva pikkusega kuni 15-minutilisi ajavahemikke, seega lõpmata palju.

4. Neljanda katse puhul on lihtne loetleda kõik selle katse abil määratletud sündmused. Kui palju neid on?

VASTUSED

Antud ***sündmus toimub*** katse korral siis, kui katse tulemusena esineb mõni selles sündmuses sisalduv elementaarsündmus.

Kuna sündmused on defineeritud elemantaarsündmuste hulkadena, siis on loomulik määratleda ka ***sündmus, millele vastab tühi hulk – see on võimatu sündmus***. Võimatu sündmuse tähis on $\varnothing$. **Sündmus, mis sisaldab kõiki elementaarsündmusi, on kindel sündmus** ja tema tähis on $\Omega$ – see on ühtlasi kogu elementaarsündmuste ruumi tähis.

• Võimatu sündmus ei toimu katse tulemusena kunagi.

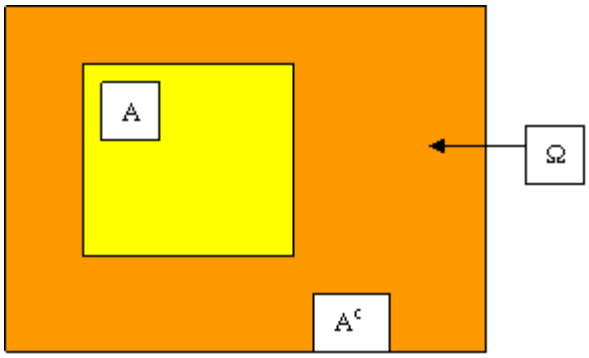
• Kui katse tehakse, toimub kindel sündmus alati.

Kui lift mahutab 10 inimest, siis on võimatu, et sellega sõidaks 100 inimest. Samuti on kindel, et sõitjate arv liftis on väiksem kui 50. Täringuviskel pole võimalik saada 8 silma ja on kindel, et saadav silmade arv on kas 1, 2, 3, 4, 5 või 6.

Sündmusi, mis pole ei kindlad ega võimatud, nimetatakse ***juhuslikeks sündmusteks***.

**Sündmuse täiendsündmus (vastandsündmus)**

Igal sündmusel on üheselt määratud ***täiendsündmus, mis toimub parajasti siis, kui sündmus ise ei toimu***. Näiteks esimese katse sündmuse *B* täiendsündmus on see, kui liftis sõidab alla 10 tudengi. Sündmuse D täiendsündmus on aga selline – liftis sõidab kas alla viie või üle kümne tudengi. Täiendsündmuse sümboliks on (sageli) kriips sümbolil: $\overline{A}$ või ka $A^C$.




**Sündmuste summa**

Olgu *A* ja *B* sündmused, mis defineeritud sama katse kaudu.

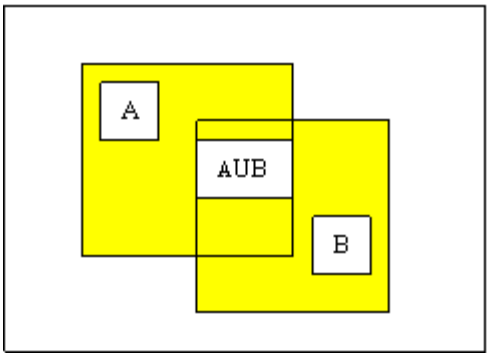
***Sündmuste summa*** $A \cup B$ ***toimub parajasti siis, kui toimub kas sündmus A, sündmus B või mõlemad.***

Näite1 korral on summaks $B \cup D$ sündmus, et liftis sõidab vähemalt viis tudengit:

• Kui sõidab üle 10 tudengi, siis toimub *B*;

• Kui sõidab 5, 6, 7, 8 või 9 tudengit, siis toimub *D*;

Kui sõidab 10 tudengit, siis toimuvad mõlemad – *B* ja *D*.

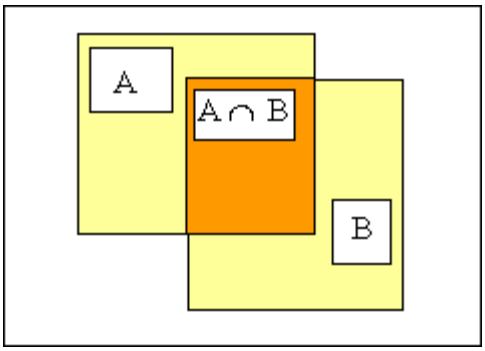



**Sündmuste korrutis**

Olgu *A* ja *B* sündmused, mis defineeritud sama katse kaudu.

**Sündmuste korrutis $A \cap B$ toimub parajasti siis, kui toimub nii sündmus *A* kui ka sündmus *B*.**

Näite 1 korral on korrutis $B \cap D$ sündmus, et liftis sõidab kümme tudengit

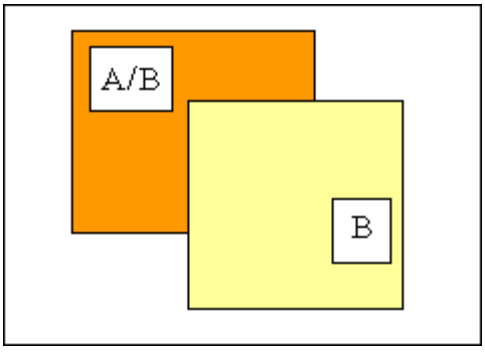



**Sündmuste vahe**

Olgu *A* ja *B* sündmused, mis defineeritud sama katse kaudu.

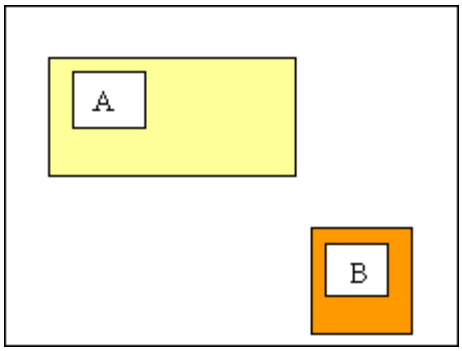
**Sündmuste vahe *A\B* toimub parajasti siis, kui toimub sündmus *A* ja sündmus *B* ei toimu.**

Näite1 korral on vahe *C\D* sündmus, et liftis sõidab kuus, kaheksa või kümme tudengit.



**Välistavad (mitteühtjad) sündmused**

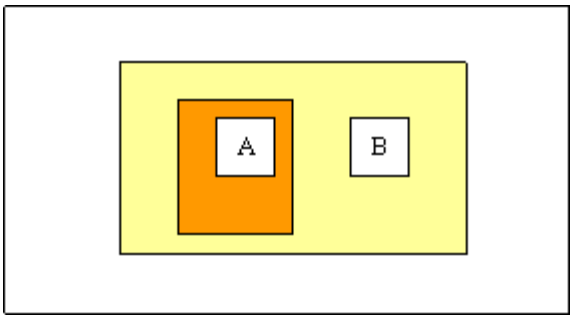
Kui kaks sündmust ei saa ühe katse tulemusena toimuda, siis on nad (üksteist) välistavad. Näite 1 korral on sündmused $A$ ja $B$ välistavad, samuti ka sündmused $A$ ja $D$.




***Välistavate sündmuste korrutis on võimatu sündmus***, sest ta ei sisalda ühtki elementaarsündmust.

**Sündmuste järeldusseos**

Kui kõik sündmuses $A$ sisalduvad elementaarsündmused sisalduvad ka sündmuses $B$, siis järeldub sündmuse $A$ toimumisest sündmuse $B$ toimumine, mida märgitakse nii: $A \longrightarrow B$ ehk $A \subset B$.

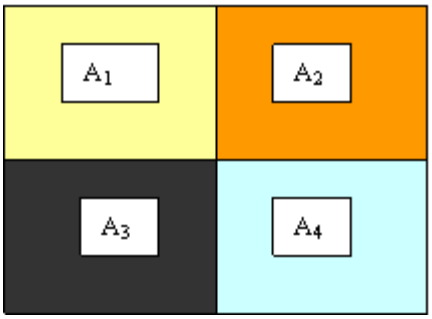



**Sündmuste täissüsteem**

Kui sündmused $A_1$, $A_2$, ..., $A_k$ rahuldavad järgmisis tingimusi:

1. Nad on (paarikaupa) välistavad;
2. Nende summa moodustab kindla sündmuse,

siis nimetatakse seda sündmuste hulka ***sündmuste täissüsteemiks***.



Iga sündmus moodustab koos oma vastandsündmusega sündmuste täissüsteemi.

**Järelemõtlemiseks**

1.Miks toimub kindal sündmus alati?

2.Kuidas saaks sündmuse vastandsündmuse defineerida sündmuste vahe kaudu?

3.Kas kahe sündmuse summa ja samade sündmuste korrutis võivad olla omavahel võrdsed?

4.Missugused on üldjuhul sisaldusseosed sündmuste ja nende summa vahel? Sündmuste ja nende korrutise vahel? Sündmuste ja nende vahe vahel?

5.Defineerige nn ***sümmeetriline vahe***, mis koosneb kahe sündmuse neist elementaar-sündmustest, mis teises sündmuses ei sisaldu.

6.Missugustest välistavatest sündmustest koosneb üldjuhul kahe sündmuse summa?

**Sündmus. Klassikaline ja geomeetriline tõenäosus**

## KLASSIKALINE TÕENÄOSUSE MÕISTE

### Klassikaline tõenäosuse mõiste

Kuigi iga juhusliku sündmuse kohta on teada, et ta võib toimuda või mitte toimuda, on siiski erinevate sündmuste toimumise võimalused erinevad. Seda iseloomustabki sündmuse tõenäosus, so ***sündmuse toimumise võimalikkuse mõõt***. Üldiselt pole aga mingi sündmuse tõenäosust sugugi lihtne määrata.

### Klassikaline tõenäosus

Kõige lihtsam on tõenäosust leida niisuguste katsete puhul, millel on lõplik arv ***võrdtõenäoseid*** katsetulemusi ehk ***elementaarsündmusi***, nagu näiteks täringuviskel. Võrdtõenäosuse omadus ei ole matemaatiliselt kontrollitav, see on eeldatav ja tuleneb katsekorralduse teatavast sümmeetriast – näiteks mündiviskel peab münt olema korrapärane, täringuviskel on kõik täringu tahud võrdse kuju ja suurusega ning täringu raskuskese asub geomeetrilises keskpunktis jne. Sellise juhu jaoks sobib alljärgnev tõenäosuse definitsioon, mis võeti kasutusele juba keskajal.

> Sündmuse tõenäosus võrdub selles sündmuses sisalduvate elementaar-sündmuste arvu $k$ ja kõigi elementaarsündmuste arvu $n$ jagatisega.

Sündmuses sisalduvaid elementaarsündmusi nimetatakse ka selle sündmuse jaoks soodsateks katsetulemusteks. Niisugusel viisil defineeritud tõenäosust nimetatakse ***klassikaliseks tõenäosuseks***.

Nii saame arvutada, et näiteks täringuviskel on iga silmade arvu saamise tõenäosus 1/6, aga näiteks paarisarvulise tulemuse saavutamise tõenäosus 3/6=0,5.

Tõenäosuse sümboliks on $P$, $P(A)$ tähistab sündmuse $A$ tõenäosust.

### Tõenäosuse omadused

Tõenäosuse definitsioonist järelduvad järgmised tõenäosuse omadused:

1. Tõenäosuse väärtus on 0 ja 1 vahel, kusjuures võimatu sündmuse tõenäosus on 0 ja kindla sündmuse tõenäosus on 1.
2. Kui sündmused $A$ ja $B$ on välistavad, siis kehtib võrdus:

$$P(AUB) = P(A) + P(B). \quad (1)$$

See omadus järeldub vahetult tõenäosuse definitsioonist, kus $k_A$ tähistab sündmuses $A$ ja $k_B$ – sündmuses $B$ sisalduvate elementaarsündmuste arvu. Et sündmused on välistavad, siis sisaldub nende summas

$$k_A + k_B$$

elementaarsündmust ja kehtib ilmne võrdus:

$$\frac{k_A}{n}+\frac{k_B}{n}=\frac{k_A+k_B}{n}. \quad (1')$$

3. Sündmuse ja tema vastandsündmuse tõenäosuste summa on 1, ehk

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

Klassikalise tõenäosuse korral kehtib ka omaduse 1 pöördväide – kui sündmuse tõenäosus on 0, siis ta on võimatu ja kui sündmuse tõenäosus on 1, siis ta on kindel. Sündmust, mille (klassikaline) tõenäosus on suurem kui null ja väiksem kui üks, nimetatakse ***juhuslikuks sündmuseks***.

**Tõenäosuste liitmise teoreem**

Teoreem väidab, et

***Kahe suvalise sündmuse summa tõenäosus avaldab järgmiselt***:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Viimane omadus vajab tõestamist. Tõestuse juures on abiks lisatud joonis.

• Sündmus *A* on esitatav kahe välistava sündmuse summana:

$$A = (A \setminus B) \cup (A \cap B)$$

• Sündmus *B* on esitatav kahe välistava sündmuse summana:

$$A = (A \setminus B) \cup (A \cap B)$$

• Sündmuste *A* ja *B* summa on

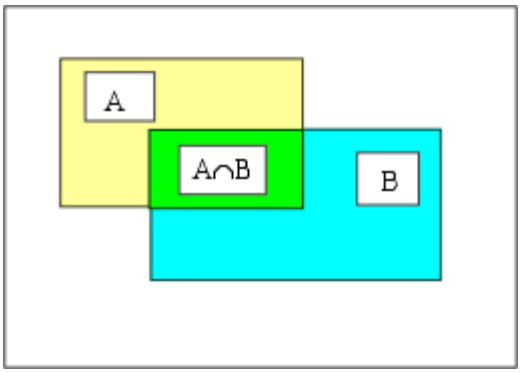
$$(A \setminus B) \cup (A \cap B) \cup (B \setminus A),$$

kusjuures kõik sündmused on taas välistavad.

• Arvutades nüüd otsitava summa tõenäosuse, saame:

$$P(A \cup B) = P(A) + P(B \setminus A) = P(A) + P(B) - P(A \cap B),$$

***mida oligi tarvis tõestada***.




Omaduste 1 ja 2 ning tõestatud teoreemi põhjal on võimalik arvutada kõigi sündmuste klassikalisi tõenäosusi, mis on saadud teatavate tuntud sündmuste summade ja vahede tulemusena (ja avalduvad samade elementaarsündmuste abil).

**Järelemõtlemiseks**

1.Kuidas on seotud sündmuse ja tema vastandsündmuse tõenäosus?

2.Kui tunnus *A* järeldub tunnusest *B*, missugune on siis nende tõenäosuste suuruse vahekord?

3.Missugused võrratused seovad sündmusi nende summa ja vahega?

4.Tuletada valem kolme sündmuse summa tõenäosuse jaoks.

5.Kas on mõtet kinnitusel, et mingi sündmus järeldub kindlast sündmusest? Võimatust sündmusest?

**Sündmus. Klassikaline ja geomeetriline tõenäosus**

## GEOMEETRILINE TÕENÄOSUS

### Geomeetriline tõenäosus lõigul

Tegelikult ei ole nõue, et elementaarsündmuste arv oleks lõplik, tõenäosuse defineerimisel oluline. Näiteks bussipeatuses ootamise näite puhul on erinevate katsetulemuste/ elementaarsündmuste arv lõpmata suur. Eeldame, et kindel sündmus $\Omega$ on bussi saabumine 15 minuti jooksul ning mingil ajavahemikul $(t_1, t_2)$ bussi saabumise tõenäosus on võrdeline selle ajavahemiku pikkusega $t_1 - t_2$. Siis on võimalik vajaliku ooteaja kestuse tõenäosust arvutada valemist

$$P(A) = \frac{t_2 - t_1}{15}.$$

Samal põhimõttel defineeritaksegi geomeetriline tõenäosus.

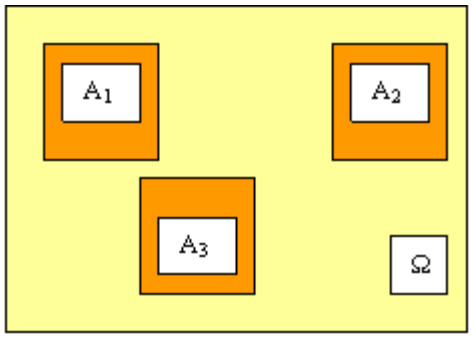
Kui tõenäosusruum $\Omega$ on lõik pikkusega $L$ ja elementaarsündmused on selle lõigu punktid, siis on kõik sellel lõigul paiknevad lõigud ja vahemikud sündmused. Iga sündmuse $A$ tõenäosus on määratud valemiga $P(A) = l(A)/L$, kus $l(A)$ tähistab lõigu $A$ pikkust.

### Geomeetriline tõenäosus tasandil

Samal viisil võib määratleda ka geomeetrilise tõenäosuse tasandil ja ruumis.

Määratlemaks geomeetrilist tõenäosust tasandil loeme teatava pinnaosa tõenäosus-ruumiks pindalaga $S$ ja kõik sellel pinnaosal paiknevad punktid elementaarsündmusteks. Mingi kujundi $A$ tõenäosus on selle kujundi pindala $s(A)$ ja tõenäosusruumi pindala $S$ suhe $P(A) = s(A)/S$.

Geomeetrilise tõenäosuse korral interpreteeritakse sageli katset (so juhusliku punkti valikut) näiteks tulistamisega märklaua pihta või punkti juhusliku viskega märklauale. Oluline on siinjuures see, et nn märklaua kõigi piirkondade tabamine toimub võrdse tõenäosusega (mis sõltub ainult piirkonna suurusest).




Joonisel on koguruumi pindala 12, sündmus *A* koosneb kolmest ruudust, neist igaüks on pindalaga 1 ja sündmuse *A* tõenäosus on seega ¼.

**Geomeetrilise tõenäosuse omadused**

Enamus klassikalise tõenäosuse omadusi kehtib ka geomeetrilise tõenäosuse korral. Tõenäosuse definitsioonist järelduvad järgmised tõenäosuse omadused:

1. Tõenäosuse väärtus on 0 ja 1 vahel, kusjuures võimatu sündmuse tõenäosus on 0 ja kindla sündmuse tõenäosus on 1.
2. Kui sündmused *A* ja *B* on välistavad, siis kehtib võrdus:

$$P(A \cup B) = P(A) + P(B). \quad (2)$$

See väide tuleneb lõigu pikkuse ja kujundite pindala aditiivsusest ning ei vaja täiendavat tõestamist.

Geomeetrilise tõenäosuse korral kehtib ka tõenäosuste liitmise teoreem, sest selle tõestamise juures me ei kasutanud klassikalise tõenäosuse määratlust ega eriomadusi.

Erinevalt klassikalisest tõenäosusest on ***geomeetrilise tõenäosuse korral elementaarsündmuse tõenäosus võrdne nulliga***. Siit tuleneb, et erinevalt klassikalisest tõenäosusest ei ole geomeetrilise tõenäosuse korra nulltõenäosusega sündmus alati võimatu ega ka ühiktõenäosusega sündmus alati kindel.

**Järelemõtlemiseks**

1.Vaatleme elementaarsündmuste ruumina lõiku pikkusega 5*a*. Vastaku sündmusele *A* lõik pikkusega *a*, sündmusele *B* lõik pikkusega 2*a*. Missuguseid väärtusi võivad omandada sündmuste $A \setminus B$ ja $B \setminus A$ tõenäosused?

2.Mis on võimatu sündmus (ruumis, kus on defineeritud geomeetriline tõenäosus)?

3.Kuidas interpreteerida olukorda, kus sündmuse tõenäosus on 0, kuid sündmus ei ole võimatu?

4.Olgu $P(A) = 0{,}4$ ja $P(B) = 0{,}8$. Kas need sündmused saavad olla välistavad?

5.Missuguses vahemikus saavad olla eelmises punktis märgitud sündmuste *A* ja *B* puhul sündmuste $A \cup B$, $A \setminus B$ ja $B \setminus A$ tõenäosused?

VASTUSED

**Statistiline tõenäosus ja suurte arvude seadus. Sündmuste sõltuvus**

## STATISTILINE TÕENÄOSUS

### Tõenäosuse üldine mõiste

Nagu selgus, pole olemas ühtset eeskirja, mille alusel saaks kõikvõimalike sündmuste jaoks tõenäosust arvutada. Küll aga on tehtud selgeks teatavad omadused, millele peab vastama elementaarsündmuste ruumil $\Omega$ defineeritud funktsioon selleks, et ta võiks olla tõenäosus. Need omadused tulenevad A. N. Kolmogorovi poolt 20.sajandi 30ndatel aastatel sõnastatud aksiomaatilisest tõenäosuse käsitlusest. Mõnevõrra lihtsustatult on need järgmised:

1.Tõenäosus on mittenegatiivne elementaarsündmuse funktsioon, seega alati $P(A) \geq 0$ .

2.Kogu tõenäosusruumi e. kindla sündmuse tõenäosus on 1

3.Tõenäosus on aditiivne, st et kui $A$ ja $B$ on üksteist välistavad sündmused, siis kehtib võrdus

$$P(A \cup B) = P(A) + P(B).$$

3*.Viimane omadus peab kehtima ka lõpmatu koonduva sündmuste jada korral:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i),$$

kus endiselt eeldatakse, et sündmused on üksteist välistavad.

Praktiliselt saab mingi katse abil defineeritud sündmustele määrata tõenäosused siis, kui on teada (1) aluseks olev elementaarsündmuste ruum, (2) eeskiri, millega igale elementaarsündmusele omistatakse tõenäosus ja (3) on selge, et omistatud tõenäosused rahuldavad tõenäosuse nõudeid 1–3 (lõpmatu elementaar- sündmuste süsteemi korral lisandub 3*).

Kuigi geomeetriline tõenäosus avardab märgatavalt nende sündmuste ringi, mille jaoks on võimalik tõenäosust arvutada, lisades lõpmatu (mitteloenduva) tõenäosusruumi, jääb siiski üle väga palju sündmusi, mille tõenäosusi seni esitatud reeglid arvutada ei võimalda. Niisugused on sündmused 1 ja 2 toodud näidetes.

### Katseseeria ja statistiline tõenäosus

Paljude reaalses elus esinevate sündmuste jaoks saab arvutada välja statistilise tõenäosuse. Selleks tuleb korraldada katseseeria, mille pikkus $n$ on korraldaja valida, kuid see tuleb enne katsete algust ette määrata. Kõik seeriasse kuuluvad katsed peavad olema korraldatud täpselt samades tingimustes, kusjuures eriti oluline on see, et need tingimused ei tohi katseseeria korral muutuda ega ka sõltuda eelnevate

katsete tulemustest. Nii saadud katseseeria korral on ***iga üksiku katse tulemus käsitletav elementaarsündmusena, kusjuures need elementaarsündmused on võrdtõenäosed***. Kogu katseseeria moodustab siis elementaarsündmuste ruumi $\Omega$.

Olgu *A* sündmus, mis vaadeldava katse tulemusena võib esineda või mitte esineda. Katseseeria puhul saame mõõta, mitu korda sündmus *A* esines. ***Sündmuse esinemiste arv k on sagedus, suhe k/n aga suhteline sagedus.***

| |
|---|
| Sündmuse *A* suhtelist sagedust *k*/*n* katseseerias pikkusega *n* nimetatakse selle sündmuse ***statistiliseks tõenäosuseks***. |

**Statistilise tõenäosuse omadused**

Fikseeritud pikkusega katseseeria puhul on statistilise tõenäosuse omadused peaaegu samad, mis klassikaliselgi tõenäosusel. Statistilise tõenäosuse definitsioonist järeldub, et:

1.Tõenäosuse väärtus on 0 ja 1 vahel, kusjuures võimatu sündmuse tõenäosus on 0 ja kindla sündmuse tõenäosus on 1.

2.Kui sündmused *A* ja *B* on üksteist välistavad, siis kehtib võrdusreale (vt valem (1) loengust 1):

$$P(A \cup B) = P(A) + P(B).$$

See väide tuleneb statistilise tõenäosuse definitsioonist üksikute katsetulemuste kui elementaarsündmuste kaudu täpselt sarnaselt klassikalise tõenäosusega.

Statistilise tõenäosuse korral kehtib ka tõenäosuste liitmise teoreem, kusjuures selle tõestus langeb täpselt ühte sama teoreemi tõestusega klassikalise tõenäosuse jaoks.

Statistilisel tõenäosusel on üks oluline erinevus võrreldes klassikalise tõenäosusega.

• Sellest, et mingi sündmuse statistiline tõenäosus võrdub nulliga, ei järeldu see, et ta on võimatu.

• Iga sündmus, mille statistiline tõenäosus on 1, ei ole kindel sündmus.

Nende omaduste poolest sarnaneb statistiline tõenäosuse geomeetrilise tõenäosusega.

**Statistilise tõenäosuse juhuslikkus**

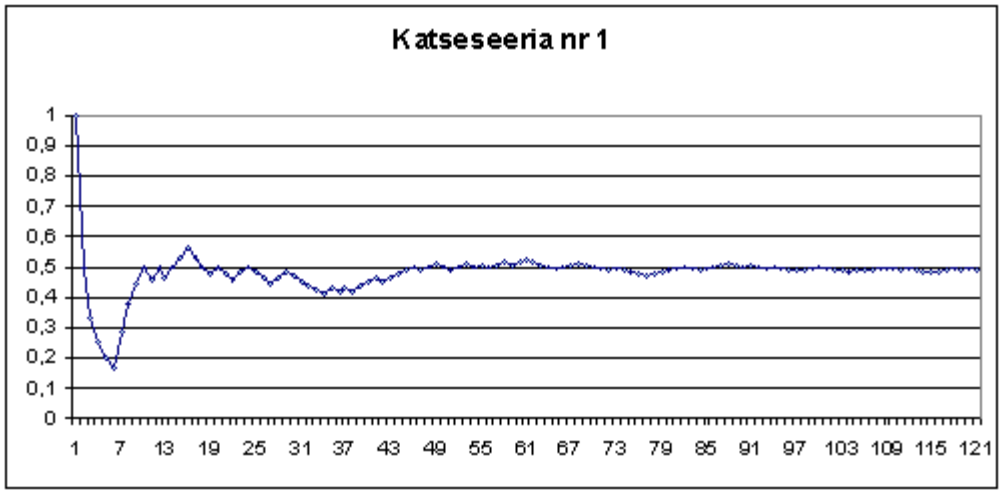
Statistilise tõenäosuse puhul tuleb arvestada seda, et katseid korrates saadakse üldiselt erinevad katseseeriad, ning enamasti on erineva katseseeria korral arvutatud sama sündmuse statistilised tõenäosused erinevad. Üldiselt erinev tulemus saadakse ka siis, kui sama katseseeriat jätkatakse. Tekib küsimus – kas nii muutlikust tõenäosuse näitajast üldse on mingit kasu?

Selgub siiski, et on. Kogemuslikult on teada, et kui katseseeriad on küllalt pikad, siis tavaliselt hakkavad suhtelise sageduse väärtused mingile konstandile lähenema. Kuid see lähenemine ei toimu nii, nagu toimuvad piirprotsessid nn mittejuhuslikus matemaatikas, kus tavaliselt jada elemendi ja piirväärtuse erinevus iga sammuga aina väheneb.

**Tõenäosuse järgi koondumise graafiline pilt**

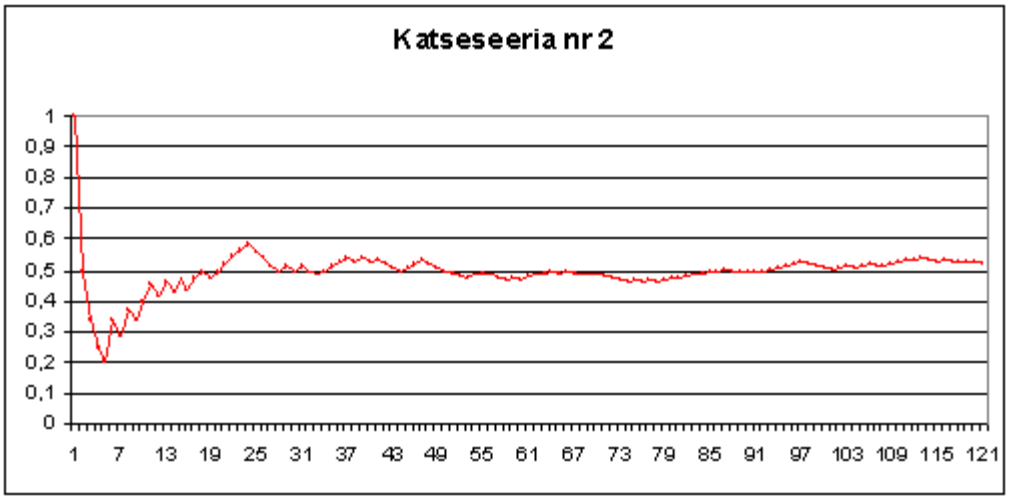
Juhuslikkude sündmuste matemaatikas toimub ka koondumine teisiti, nagu võime

veenduda lisatud graafikutelt. Neil on kirjeldatud mündiviskel kirjapoole pealelangemise (sündmuse *A*) suhtelise sageduse graafikud kolme 120-viskese seeria puhul.



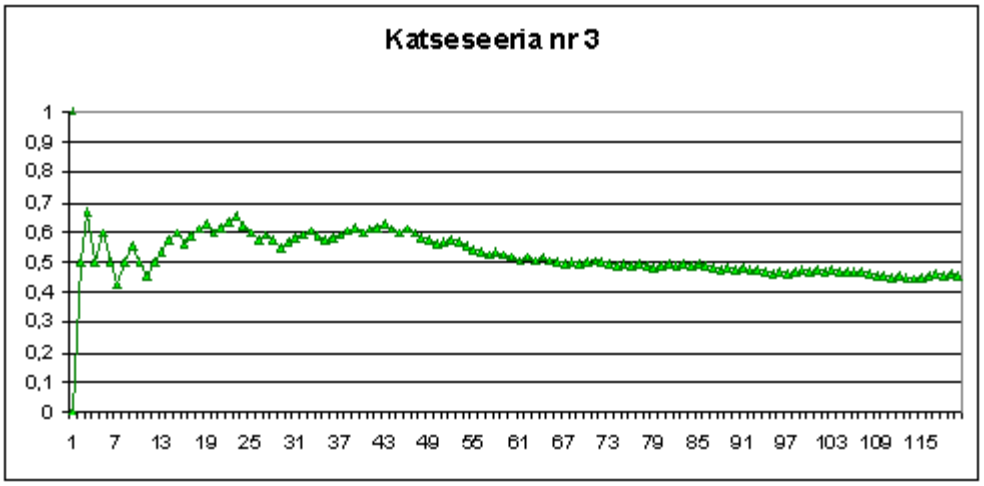

Graafikut jälgides paneme tähele, et:

•Esialgu on graafikul saehambuline kuju vastavalt sellele, kas järjestikusel katsel toimus sündmus




*A* või mitte.

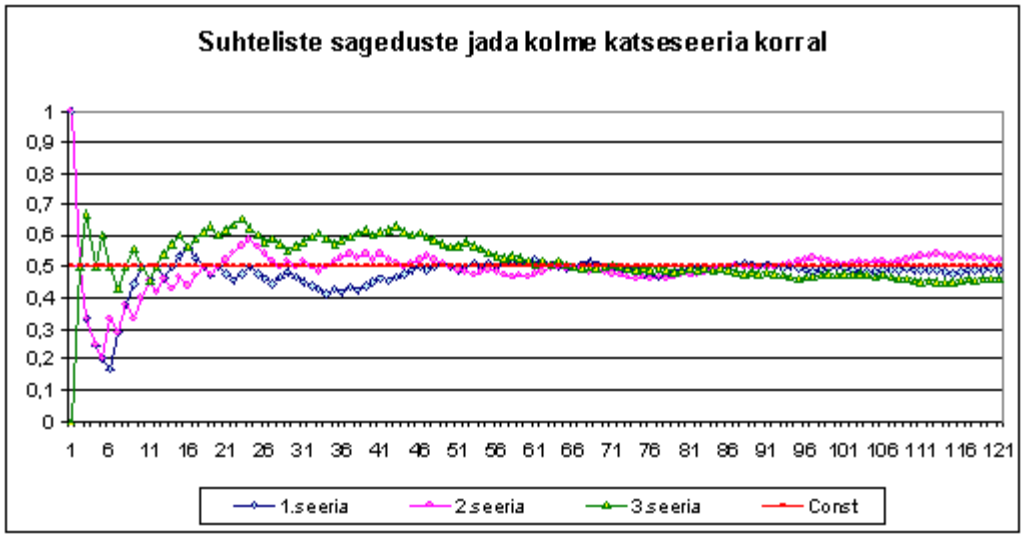
•Hammaste suurus aegapidi väheneb, kuid murdjoon moodustab ebakorrapäraseid laineid. Üks



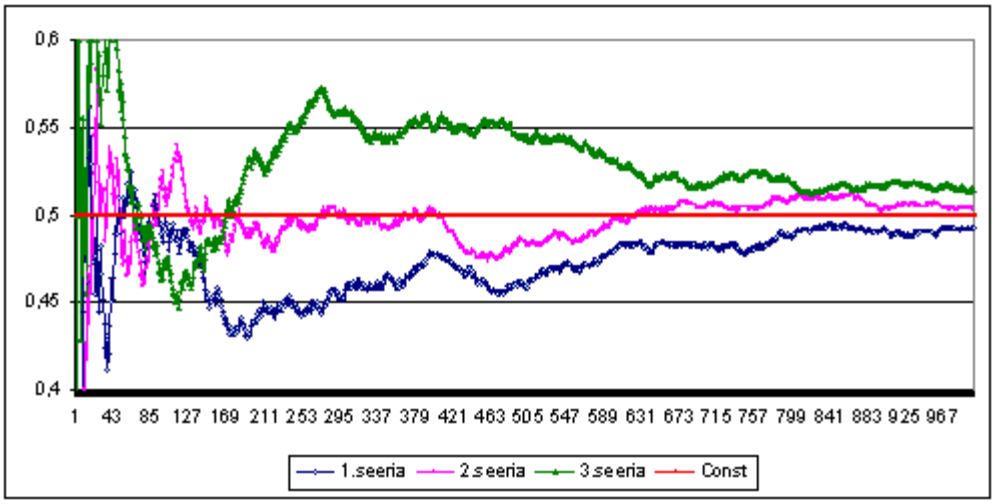

lainetuse miinimumpunkte on 35. katsel, järgmine, väiksema sügavusega aga 78. katsel.

•Suhtelise sageduse graafiku kuju muutub katsete arvu suurenemise tulemusena järjest siledamaks.

Samasugust üldist tendentsi näitavad ka kahe sarnase katseseeria protokollid. Joonis, millel on kolme katseseeria protokollid koos, näitab, et üldiselt käituvad katseseeriad erinevalt, kuigi näiteks esimese ja teise katseseeria viis esimest punkti ühtivad. Ühtivad ka esimese ja kolmanda punktid järjekorranumbritega 10—12. Niisugune olukord leiab aset veel vaatluste 60, 62 , 63 ja 70 korral. Sama kehtib ka esimese ja teise vaatlusseeria puhul, kus samad väärtused on punktides 18—20 ja 90 ning 92. Tähelepanuväärne on vaatlus nr 66 – siis on kõigi kolme vaatlusseeria korral suhtelise sageduse väärtuseks 0,5.




Viimasel joonisel on esitatud samad katseseeriad kuni 1000 katseni; horisontaalteljel on tehtud skaalateisendus, nii et graafikul paikneb neli korda kitsam sageduste riba kui eelmisel graafikul. Sellel graafikul ilmneb, hoolimata vahepealsetest võngetest, suhteliste sageduste lähenemine üksteisele alates 600ndast vaatlusest. Niisugune graafiku kuju iseloomustab koonduvust, mida nimetatakse ***koonduvuseks tõenäosuse järgi***.




**Suurte arvude seadus**

Ühtlasi illustreerivad kõik esitatud joonised ***suurte arvude seadust***, mille sõnastame käesolevas loengus ilma tõestuseta.

•Olgu määratud katse, mille tulemusena on võimalik sündmuse *A* esinemine.

•Eeldame, et sündmuse *A* tõenäosus $p = P(A)$ on teada (näiteks arvutatud klassikalise või geomeetrilise tõenäosusena).

•Olgu sündmuse *A* suhteline sagedus n katse korral $k_n/n$, kus $k_n$ on sündmuse esinemiste arv (sagedus) *n* katse korral.

•Siis koondub katseseeria piiramatul pikenemisel sündmuse *A* suhteline sagedus tõenäosuse järgi selle sündmuse tõenäosuseks:

$$\frac{k_n}{n} \xrightarrow{p} p.$$

**Suurte arvude seaduse seos statistilise tõenäosusega**

Suurte arvude seaduse oluline järeldus on see, et küllalt pika katseseeria korral erineb statistiline tõenäosus küllalt vähe oma piirväärtusest, mida võib lugeda tõenäosuse "õigeks" väärtuseks. Praktika seisukohast on aga kõige tähtsam see, ***et fikseeritud katseseeria korral arvutatud statistiline tõenäosus rahuldab tõenäosuse põhiomadusi 1—3,*** ning seega on selle kasutamine igati korrektne. Fikseeritud pikkusega katseseeria korral ei ole tarvis kontrollida tõenäosuse omaduse 3* täidetust. Küll aga on omadus 3* oluline suurte arvude seaduse puhul, kui käsitletakse katseseeriate jada.

Statistilist tõenäosust võib nimetada ka õige, ent mitte teadaoleva tõenäosuse ***hinnanguks***. Hinnangu mõistega kohtume kursuse teises, statistikale pühendatud osas. Kuna statistiline tõenäosus sisaldab teatavat juhuslikku viga, võib tema väärtust soovi korral ka ümmardada. Siiski pole alust arvata, et kõik "õiged" tõenäosused on naturaalarvude jagamisel saadud lihtsa struktuuriga ratsionaalarvud, nagu me saame klassikalise tõenäosuse arvutamisel.

Praktiliste ülesannete lahendamisel kasutatakse kõige sagedamini statistilisi tõenäosusi. Siinjuures on aga üldiseks nõudeks, et tegemist on ühe katseseeria

alusel määratud tõenäosustega.

**Järelemõtlemiseks**

1.Olgu katseseerias *n* katset, ning olgu leitud sündmustele statistilised tõenäosused. Tehakse üks täiendav katse ja arvutatakse uued statistilised tõenäosused. Missuguste sündmuste puhul erinevad uued tõenäosused eelmistest?

2.Kui suur on küsimuse 1 ülesandes kõige suurem erinevus endise ja uue tõenäosuse vahel?

3.Kas on võimalik selline katseseeria, mille korral statistiline tõenäosus kogu katseseeria korral ainult kasvab? Kui on, siis millal?

4.Kas on võimalik niisugune katseseeria, mille korral suhtelise sageduse ja tõenäosuse erinevuse absoluutväärtus iga katse korral kahaneb?

5.Kas tõenäosuste liitmise teoreem kehtib siis, kui sündmused *A* ja *B* on defineeritud küll sama katse tulemuste põhjal, kuid neist ühe tõenäosus on määratud *n* katsest koosneva seeria põhjal, teise tõenäosuse määramisel kasutati katseseeriat, milles esialgsele *n* katsele lisandus veel *m* täiendavat katset?

**Statistiline tõenäosus ja suurte arvude seadus. Sündmuste sõltuvus**

## TINGLIK TÕENÄOSUS JA SÜNDMUSTE SÕLTUVUS

**Tingimus**

Katse korraldamise juures oli nõudeks katsetingimuste fikseeritus ja püsivus. Mõnikord aga võib lisaks katsega määratud tingimustele defineerida täiendavaid tingimusi ka sündmuste abil. Näiteks on teada, et mingi sündmus *B* toimub/ toimus, ja pakub huvi leida teiste sündmuste tõenäosused ***seda tingimust arvestades***. Sisuliselt tähendab tingimus *B* ***teatava lisainformatsiooni*** olemasolu katsetulemuste kohta, ning tavaliselt on otstarbekas seda kasutada.

***Näide***

1.Vaatleme tudengite hulka liftis, kui on teada, et lift mahutab vaid 10 inimest. Siis võime tingimuseks *B* lugeda sündmuse, et liftis on kuni 10 inimest, ning kõigi teiste sündmuste tõenäosusi arvutame seda tingimust arvesse võttes.

2.Vaatame täringuviske ülesannet, kui on teada, et täringul ei langenud peale maksimaalne silmade arv 6.

Tingliku tõenäosuse puhul on võimalike elementaarsündmuste hulk piiratud ja kindla sündmuse $\Omega$ asemele asub nüüd tingimust määrav sündmus *B* – võimalikud on ainult selles sündmuses sisalduvad elementaarsündmused.

***Tinglik tõenäosus***

Leiame sündmuse *A* tingliku tõenäosuse tingimusel *B*, ning tähistame seda sümboliga $P(A/B)$.

Selleks arutleme järgmiselt:

• Sündmuse *A* toimumine tingimusel *B* tähendab tegelikult sündmuste *A* ja *B* koos toimumist, st sündmuste korrutise $A \cap B$ toimumist.

• Kui kehtib tingimus *B*, siis moodustavad sündmuses *B* sisalduvad elementaarsündmused kindla sündmuse.

Seda arvestades saame sündmuse *A* tingliku tõenäosuse avaldiseks tingimusel *B* järgmise murru:

$$P(A/B)=\frac{P(A\cap B)}{P(B)}. \quad (3)$$

Iga sündmusega *B* määratud ***tinglikud tõenäosused on tõenäosused***, st nad rahuldavad tingimusi 1—3. Erinevuseks võrreldes tingimatute tõenäosustega on see, et tingliku tõenäosuse puhul on $\Omega = B$, kus *B* tähistab tingimust määravat sündmust.

**Sündmuste korrutise tõenäosus**

Tingliku tõenäosuse valemist (3) tuleneb teine praktikas väga sageli vajalik valem – see on sündmuste korrutise tõenäosuse valem:

$$P(A \cap B) = P(B)\,P(A/B) = P(A)\,P(B/A). \quad (4)$$

Kuna kahe sündmuse korrutis on kommutatiivne (st ei olene tegurite järjestusest), siis on võrduse keskmine ja parempoolne avaldis samaväärsed. Tõenäosuste korrutamise tõenäosuse lause saame sõnastada järgmiselt.

| Kahe sündmuse korrutise tõenäosus võrdub ühe sündmuse tõenäosuse ja teise sündmuse tingliku tõenäosuse korrutisega, kusjuures tingimuse määrab esimene sündmus. |
|---|

**Sündmuste sõltumatus**

Mõnikord ei lisa ühe sündmuse *B* toimumine ja sellega määratud tingimus mõne teise sündmuse *A* toimumise kohta, mingit täiendavat informatsiooni ja ei muuda selle tõenäost, st et kehtib seos

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

Sel korral öeldakse, et **sündmus *A* ei sõltu sündmusest *B***.

Asetades saadud võrduse valemisse (4), saame seose

$$P(A \cap B) = P(A)\,P(B). \quad (5)$$

Seos (5) kehtib parajasti siis, kui sündmus *A* ei sõltu sündmusest *B*, kuid tema sümmeetrilisusest *A* ja *B* suhtes järeldub, et siis ei sõltu ka sündmus *B* sündmusest *A*. Seega:

- Sündmused *A* ja *B* on sõltumatud, kui kehtib võrdus (5).
- Sündmuste sõltumatus on vastastikune.

**Sündmuste sõltuvus**

Sündmuste sõltuvus defineeritakse sõltumatuse eitusena – ***kui sündmused ei ole sõltumatud, on nad sõltuvad***. Ka sündmuste sõltuvus on vastastikune, siin ei tehta vahet, kumb sündmustest teist mõjutab. Küll aga võib vaadelda mõjutuse suunda selles mõttes, et peale sündmuse *B* toimumist võib sündmuse *A* tõenäosus kas suureneda või väheneda. Vahel öeldakse selle kohta, et sündmusel *B* on sündmusele *A* kas positiivne (tõenäosust suurendav) või negatiivne (tõenäosust vähendav) mõju.

***Näide***

Olgu ülesandeks leida, kuidas mõjutab leibkonna vaesusesse sattumise tõenäosust (1) leibkonnapea kõrgem haridus; (2) see, kui leibkonna moodustab üksikvanem alaealise lapse või lastega.

Vaeseks loetakse 2001. aastal Eestis leibkonda, mille netosissetulek tarbimisühiku kohta on väiksem kui 1488 krooni kuus. Tarbimisühikute arv leibkonnas määratakse Eestis vastavalt valemile

$$t = 1 + 0{,}8(p - 1),$$

kus *p* on leibkonnaliikmete arv. Seega 3-liikmeline pere kuusissetulekuga 3900

krooni kuus ei ole vaene, kuid 2-liikmeline pere kuusissetulekuga 2600 krooni on vaene.

Ülesande lahendamiseks saame kasutada sotsioloogiliste uuringute põhjal arvutatud statistilisi tõenäosusi.

•Defineerime sündmuse $A$ – leibkonna sissetulek on väiksem kui kehtestatud vaesuspiir. Sündmuse $A$ tõenäosus leibkonna eelarve uuringute andmetel on 0,24.

•Defineerime sündmused $B$ – leibkonnapea on kõrgharidusega ja $C$ – tegemist on lapse või lastega üksikvanema perega. Nende sündmuste tõenäosused on vastavalt 0,20 ja 0,04.

•Statistiliselt saab leida ka sündmuste korrutise tõenäosusi, sest ka need on sündmused, mille suhtelisi sagedusi saab arvutada.

•Sündmus $A \cap B$ tähendab seda, et tegemist on vaese leibkonnaga, kus perekonnapea on kõrgharidusega. Selle sündmuse tõenäosus on 0,02.

•Sündmuse $A \cap C$ tõenäosus on 0,03. See näitab, kui suur on tõenäosus selleks, et juhuslikult valitud leibkond on üksikvanema leibkond ja tema sissetulekud on allpool vaesuspiiri.

•Nüüd saame leida tinglikud tõenäosused $P(A/B)$ ja $P(A/C)$.

•Kui leibkonnapea on kõrgharidusega, siis on leibkonna vaesusesse sattumise tõenäosus on $P(A/B) = 0{,}02/0{,}20 = 0{,}10$.

•Kui leibkonnas on üksik vanem ja laps või lapsed, siis on leibkonna vaesusesse sattumise tõenäosus $P(A/C) = 0{,}03/0{,}04 = 0{,}75$.

Näeme, et $0{,}10 < 0{,}24$, seega kõrgharidusega perekonnapeaga leibkonna vaesusesse sattumise tõenäosus on üle kahe korra väiksem kui keskmise leibkonna vaesusesse sattumise tõenäosus.

Teiselt poolt, $0{,}75 > 0{,}24$, seega lapse või lastega üksikvanema perel on vaesusesse sattumise tõenäosus on üle kolme korra suurem kui keskmisel perel.

Toodud näited selgitavad ühtlasi seda, kuidas tinglikud tõenäosused võimaldavad kasutada lisateavet.

**Järelemõtlemiseks**

1.Kas üksteist välistavad sündmused on sõltuvad või sõltumatud?

2.Kui üks sündmus järeldub teisest, mida võib siis öelda tinglike tõenäosuste kohta?

3.Kas siis, kui üks sündmus järeldub teisest, on tegemist sõltuvate või sõltumatute sündmustega?

4.Kas katsetulemused (elementaarsündmused) on omavahel sõltuvad?

5.Millega võrdub üksteist välistavate sündmuste korrutis?

6.Kas sündmus ja tema vastandsündmus on omavahel sõltuvad?

VASTUSED

**Statistiline tõenäosus ja suurte arvude seadus. Sündmuste sõltuvus**

**BAYESI TEOREEM**

**Bayesi teoreem**

Kaheksateistkümnenda sajandi keskel tõestas inglise teadlane T. Bayes teoreemi, millele on kaasajal üles ehitatud väga oluline suund matemaatilises statistikas. Selle teoreemi sisuks on anda eeskiri katse toimumise järgselt saadud informatsiooni arvestamiseks selle katsega seotud sündmuste tõenäosuste hindamisel, üldisemalt – lisainformatsiooni kasutamine tõenäosuste arvutamisel. Kõigepealt tõestame üldkasuliku valemi, mida tuntakse täistõenäosuse valemina.

**Täistõenäosuse valem**

Tinglikke tõenäosusi on mõnikord sobiv kasutada ka tingimatu tõenäosuse arvutamiseks. Ühe võimaluse selleks pakub täistõenäosuse valem, mille eeldused on järgmised.

Moodustagu sündmused $H_1$, $H_2$, ..., $H_k$ sündmuste täissüsteemi, st et

- Nad on üksteist välistavad;
- Nende summa moodustab kindla sündmuse.
- Eeldame, et sündmuste $H_i$ tõenäosused $P(H_1)$, ..., $P(H_k)$ on teada.
- Olgu sündmus $A$ sama katse abil defineeritud, ning olgu teada tema tinglikud tõenäosused sündmuste $H_i$ suhtes $P(A/H_1)$, ..., $P(A/H_k)$.
- Siis avaldub sündmuse $A$ tõenäosus järgmise valemiga:

$$P(A)=\sum_{i=1}^{k} P(A \mid H_i)\, P(H_i), \quad (6)$$

mida nimetatakse ***täistõenäosuse valemiks***.

Täistõenäosuse valemi tõestamiseks paneme tähele, et tehtud eelduse tõttu kehtib võrdus

$$\Omega=\sum_{i=1}^{k} H_i,$$

järelikult ka

$$A=\sum_{i=1}^{k} A \cap H_i$$

ja eelduste tõttu ka

$$s_a^2=s^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_x^2}\right),\ s_b^2=\frac{s^2}{S_x^2},\ S_x^2=\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

Kasutades iga liidetava puhul tõenäosuste korrutamise eeskirja (4) saamegi valemi (6), mida oligi tarvis tõestada.

***Näide 2.1.***

Lisatud tabelis on näidatud kõigi Eesti koolilaste jaotus maakondade järgi ja igas maakonnas venekeelses koolis õppivate õpilaste arv. On tarvis leida, kui suur on tõenäosus selleks, et juhuslikult valitud Eesti õpilane käib venekeelses koolis.

| Maa-kond | Harju | Hiiu | Ida-Viru | Jõgeva | Järva | Lääne | Lääne-Viru | Põlva | Pärnu | Rapla | Saare | Tartu | Valga | Viljandi | Võru |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P(Hi) | 0,358 | 0,009 | 0,118 | 0,030 | 0,031 | 0,024 | 0,055 | 0,026 | 0,072 | 0,030 | 0,030 | 0,111 | 0,028 | 0,046 | 0,032 |
| P(A/Hi) | 0,369 | 0 | 0,804 | 0,050 | 0,020 | 0,053 | 0,086 | 0,018 | 0,092 | 0 | 0 | 0,125 | 0,148 | 0,013 | 0,025 |

Tabeli teine rida annab tõenäosuse selleks, et juhuslikult valitud õpilane kuulub mingisse maakonda, kolmas rida aga tingliku tõenäosuse selleks, et vastava maakonna juhuslikult valitud õpilane käib venekeelses koolis. Esitatud küsimusele vastuse saamiseks kasutame valemit (6):

$$P(A) = 0{,}358\times0{,}369+0{,}118\times0{,}804+\ldots+0{,}032\times0{,}025 = 0{,}262.$$

Saadud arv – 0,262 – ongi otsitav tõenäosus, et juhuslikult valitud õpilane käib venekeelses koolis ehk venekeelsete koolide õpilaste suhtelist sagedust Eesti õpilaste seas.

**Bayesi valem**

Eeldame taas, et vaadeldava katse tulemuste kaudu on määratud sündmuste täissüsteem $H_1$, $H_2$, ..., $H_k$ , mille tõenäosused $P(H_1)$, ..., $P(H_k)$ on teada. Lisaks sellele on sama katse abil defineeritud sündmus $A$, mille kohta on teada tema tinglikud tõenäosused sündmuste $H_i$ suhtes $P(A/H_1)$, ..., $P(A/H_k)$.

Oletame, et katse tulemusena sündmus $A$ toimus. Nüüd on võimalik sündmuste $H_i$ tõenäosusi täpsustada, arvutades nende tinglikud tõenäosused $P(H_i/A)$ valemist.

$$P(H_i/A)=\frac{P(H_i)\cdot P(A/H_i)}{\sum_{i=1}^{k} P(H_i)\,P(A/H_i)},\ i=1,2,\ldots,k. \quad (7)$$

Valemit (7) nimetatakse Bayesi valemiks. Bayesi teoreemi puhul kasutatakse järgmisi termineid:

•Sündmusi $H_i$ nimetatakse hüpoteesideks;

•Tingimatuid tõenäosusi $P(H_i)$ nimetatakse aprioorseteks ehk eeltõenäosusteks.

•Tinglikke tõenäosusi

$$P(H_i/A)$$

nimetatakse aposterioorseteks ehk järeltõenäosusteks.

Paneme tähele, et Bayesi valem kehtib iga hüpoteesi korral. Kuna hüpoteesid

moodustavad sündmuste täissüsteemi, siis peab ka järeltõenäosuste summa olema võrdne ühega.

**Bayesi teoreemi tõestus**

Valemi (7) vasakul poolel asuv tinglik tõenäosus avaldub vastavalt definitsioonile (3) järgmiselt:

$$P(H_i / A) = \frac{P(A \cap H_i)}{P(A)}.$$

Lugejas paiknev sündmuste korrutise tõenäosus avaldub vastavalt valemile (4) tõenäosuste korrutisena:

$$P(A \cap H_i) = P(H_i)\, P(A / H_i).$$

Nimetajas oleva sündmuse A tõenäosuse avaldamiseks kasutame täistõenäosuse valemit (6).

**Sellega on teoreem tõestatud.**

***Näide 2.2.***

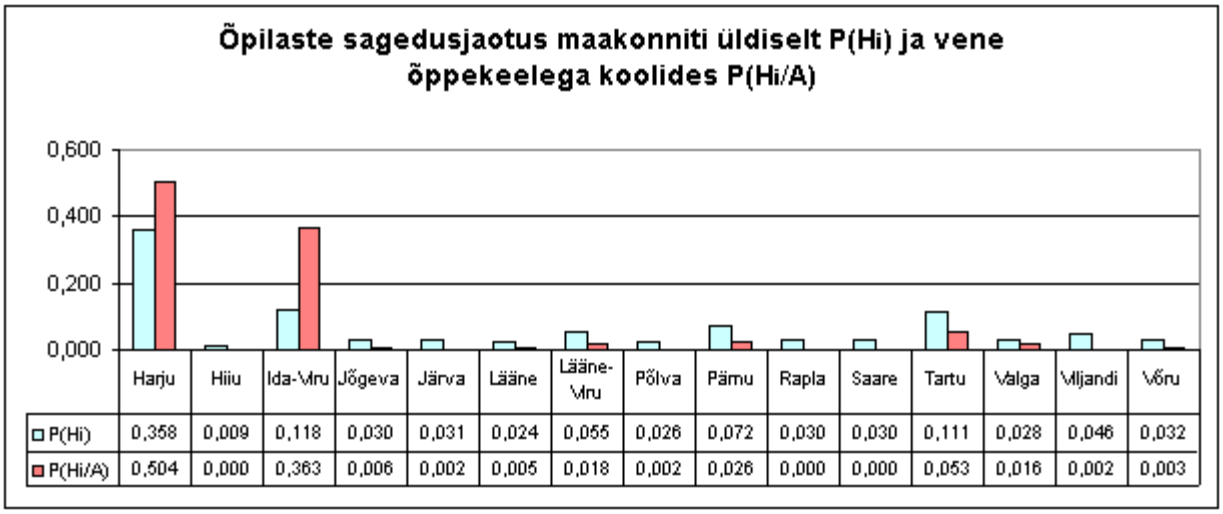
Jätkame näites 2.1 vaadeldud ülesannet. Eeldame, et juhuslikult valitud õpilane käib venekeelses koolis. Kui suur on tõenäosus, et ta on pärit Harjumaalt? Ida-Virumaalt? Viljandimaalt?

Selle ülesande lahendamiseks tuleb meil leida sündmuse *A* (õpilane käib venekeelses koolis) ja hüpoteesi $H_i$ (õpilane käib koolis Harjumaal) korrutise tõenäosus, mida saame teha, kasutades korrutise tõenäosuse valemit (4) ja tabelis antud andmeid:

$$P(A \cap H_i) = 0{,}358 \times 0{,}369 = 0{,}132.$$

Et sündmuse A (õpilane käib venekeelses koolis) tõenäosus on eelmise näite põhjal 0,26, siis on otsitav tinglik tõenäosus 0,132/0,26= 0,51.

Sarnaselt on lihtne leida, et tõenäosus, et venekeelse kooli õpilane õpib Ida-Virumaal, on 0,36, seevastu aga õppimine Järva-, Põlva- Viljandi- ja Võrumaal on arvutustäpsuse piires tõenäosusega 0, kuid pole võimatud sündmused. Seevastu aga Hiiumaal venekeelses koolis õppimine on võimatu sündmus, sest selles maakonnas ei ole üldse vene õppekeelega koole. Kontrollimise tulemusena ilmneb, et tinglike tõenäosuste summa on 1, mis kinnitab arvutuste õigsust. On näha, et järeltõenäosus (täpsustatud tõenäosus) suureneb osas maakondades eeltõenäosusega võrreldes, osas aga väheneb. Tõenäosuste muutumise suurus ilmestab seda, kui tõhusalt sündmuse A toimumine täpsustas hüpoteeside tõenäosuse hinnanguid – osutus võimalikuks osa hüpoteese kui võimatud (Hiiumaa, Raplamaa, Saaremaa) hoopis kõrvale jätta ja keskenduda kõige tõenäosemate hüpoteeside (Harjumaa, Ida-Virumaa) kontrollimisele.




| | Harju | Hiiu | Ida-Viru | Jõgeva | Järva | Lääne | Lääne-Viru | Põlva | Pärnu | Rapla | Saare | Tartu | Valga | Viljandi | Võru |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P(Hi) | 0,358 | 0,009 | 0,118 | 0,030 | 0,031 | 0,024 | 0,055 | 0,026 | 0,072 | 0,030 | 0,030 | 0,111 | 0,028 | 0,046 | 0,032 |
| P(Hi/A) | 0,504 | 0,000 | 0,363 | 0,006 | 0,002 | 0,005 | 0,018 | 0,002 | 0,026 | 0,000 | 0,000 | 0,053 | 0,016 | 0,002 | 0,003 |

**Järelemõtlemiseks**

1.Millega võrdub Bayesi valemi järgi arvutatud järeltõenäosuste (aposterioorsete tõenäosuste summa)? Miks?

2.Kas võib juhtuda nii, et mingi sündmuse $H_i$ korral on eeltõenäosus positiivne, kuid järeltõenäosus võrdub nulliga? Millal see nii juhtub?

3.Mida võib öelda sündmuse *A* kohta, kui kõik järeltõenäosused on vastavate eeltõenäosustega võrdsed?

VASTUSED

**Juhuslik suurus ja vektor. Jaotus ja tema esitused**

## JUHUSLIK SUURUS JA VEKTOR

### Juhuslik suurus

Juhusest ei sõltu mitte ainult sündmuste toimumine või mittetoimumine, vaid juhusest võib sõltuda ka mõni arvuliselt väljenduva suuruse väärtus. Lihtne näide selle kohta on lauamäng, kus kõigepealt veeretatakse täringut ja seejärel astutakse nii mitu sammu, nagu täring näitab. Sammude arv sõltub katse tulemusest, so juhusest.

### Juhusliku suuruse määratlus

Juhuslikuks suuruseks nimetatakse elementaarsündmuse arvuliste väärtustega funktsiooni. Võime öelda ka nii – juhuslik suurus omandab iga katsetulemuse korral mingi arvväärtuse. Siinjuures on oluline, et kui mingil järgmisel katsel sama tulemus kordub, on ka juhusliku suuruse väärtus sama. Seega ei saa juhusliku suuruse erinevate väärtuste arv olla suurem kui antud katse korral esineda võivate erinevate katsetulemuste arv.

Kui juhuslikul suurusel on lõplik või loenduv hulk väärtusi, siis nimetatakse teda ***diskreetseks***. Kui juhusliku suuruse väärtuste hulk on mitteloenduv, siis on tegemist ***pideva juhusliku suurusega***. Juhusliku suuruse tähiseks on sageli tähed $X$, $Y$, $Z$ jne.

### Juhusliku suuruse jaotus ja tõenäosusfunktsioon

Juhuslikku suurust iseloomustab lisaks tema väärtuste hulgale veel tema ***jaotus***, mis näitab erinevate väärtuste/ väärtushulkade esinemise tõenäosust. Jaotuse aluseks on vastavate elementaarsündmuste tõenäosused. Kui mingi katsetulemuste hulga $A$ korral on juhusliku suuruse $X$ väärtuseks $x$, siis tähendab see ühtlasi, et juhuslik suurus $X$ omandab väärtuse $x$ tõenäosusega $p = P(A)$. Seda märgitakse ka nii: $P(X = x) = p$.

Jaotusel on mitu võimalikku esitust. Diskreetse juhusliku suuruse esitusena kasutatakse peamiselt ***tõenäosusfunktsiooni***, mis näitab iga juhusliku suuruse väärtuse puhul tema esinemise tõenäosust:

$$P(X = x_i) = p_i,\ i = 1, \ldots, m,$$

$m$ tähistab juhusliku suuruse $X$ erinevate väärtuste arvu, $m \leq n$ ja $n$ on erinevate katsetulemuste arv.

Tõenäosusfunktsiooni puhul on alati täidetud järgmised tingimused:

1. Kõik tõenäosused $p_i$ on mittenegatiivsed;

2. Tõenäosusfunktsiooni kõigi tõenäosuste summa on 1,

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

Viimane tingimus tähendab sisuliselt seda, et kui katse teostatakse, siis omandab selle katse tulemuste abil defineeritud juhuslik suurus kindlasti mingi väärtuse, kuid mitte kunagi rohkem kui ühe väärtuse. Seda nimetatakse ka tõenäosusfunktsiooni ***põhiomaduseks***.

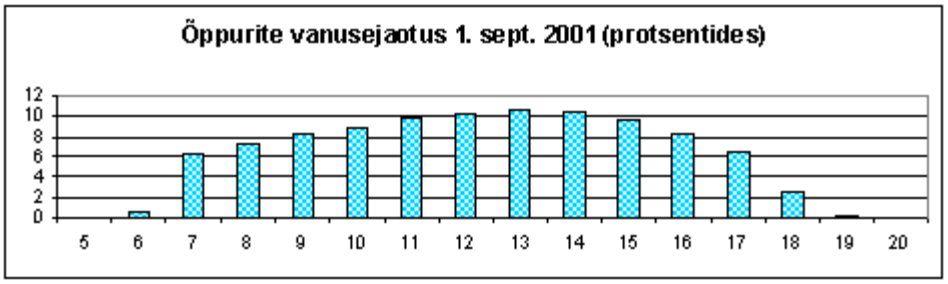
***Juhusliku suuruse abil defineeritud sündmused***

Juhusliku suuruse abil on võimalik defineerida mitmesuguseid sündmusi, näiteks:
$X = a$, $X > c$, $X < b$, $d < X < e$.
Kõigi nende sündmuste tõenäosusi saab arvutada, teades juhusliku suuruse jaotust.

**Empiiriline jaotus**

***Näide 3.1***

Vaatleme juhusliku suurusena Eesti üldhariduskooli õppurite vanusejaotust 1. septembril 2001.




| Vanus | Õpilaste arv | Suhteline sagedus |
|---|---|---|
| 5 | 2 | 0 |
| 6 | 1461 | 0,7 |
| 7 | 13130 | 6,32 |
| 8 | 15210 | 7,33 |
| 9 | 17015 | 8,2 |
| 10 | 18500 | 8,91 |
| 11 | 20649 | 9,95 |
| 12 | 21327 | 10,27 |
| 13 | 22302 | 10,74 |
| 14 | 21732 | 10,47 |
| 15 | 20195 | 9,73 |
| 16 | 17022 | 8,2 |
| 17 | 13399 | 6,45 |

Üks võimalus juhusliku suuruse üksikväärtuste tõenäosuste määramiseks on suhteliste sageduste kasutamine, nagu tehtud ka lisatud näites. Niisugusel viisil saadud juhusliku suuruse jaotust nimetatakse empiiriliseks jaotuseks.

Empiirilisi jaotusi esitatakse tavaliselt tabelina (vt lisatud tabel), kus väga sageli suhtelised sagedused avaldatakse protsentidena ja lisaks illustreeritakse tihti graafiku (nt tulpdiagrammi) abil.

Empiiriline jaotus võib olla täpne sel juhul, kui tegemist on lõpliku üldkogumiga, kus kõik objektid on samaväärsed ja selle üldkogumi kõigi objektide puhul on tunnuse väärtus määratud. Niisugune on olukord käesolevas näites.

Teise võimalusena saadakse empiiriline

| | | |
|---|---|---|
| 18 | 5300 | 2,55 |
| 19 | 335 | 0,16 |
| 20 | 33 | 0,02 |
| Kokku | 207612 | 100 |

jaotus siis, kui juhusliku suuruse väärtuste esinemise sagedusi hinnatakse katseseeria põhjal, st rakendatakse statistilist tõenäosuse mõistet.

**Juhuslik vektor ja selle jaotus**

Kui sama katse abil on defineeritud mitu juhuslikku suurust, siis nad moodustavad juhusliku vektori, mille iga komponent on juhuslik suurus. Vektorit iseloomustab tema komponentide ühisjaotus, mis näitab iga komponentide väärtuste kombinatsiooni jaoks selle esinemise tõenäosust. Kui juhusliku vektori ühel komponendil on *m* ja teisel *h* väärtust, siis on juhuslikul vektoril kokku *mh* erinevat väärtust. Tavaliselt tähistatakse ühisjaotuse tõenäosusfunktsiooni tõenäosusi tähisega $p_{ij}$. Ka ühisjaotuse tõenäosused rahuldavad tõenäosusfunktsiooni tingimusi 1 ja 2, neist viimane on esitatav kujul:

$$\sum_{i=1}^{m}\sum_{j=1}^{h} p_{ij} = 1.$$

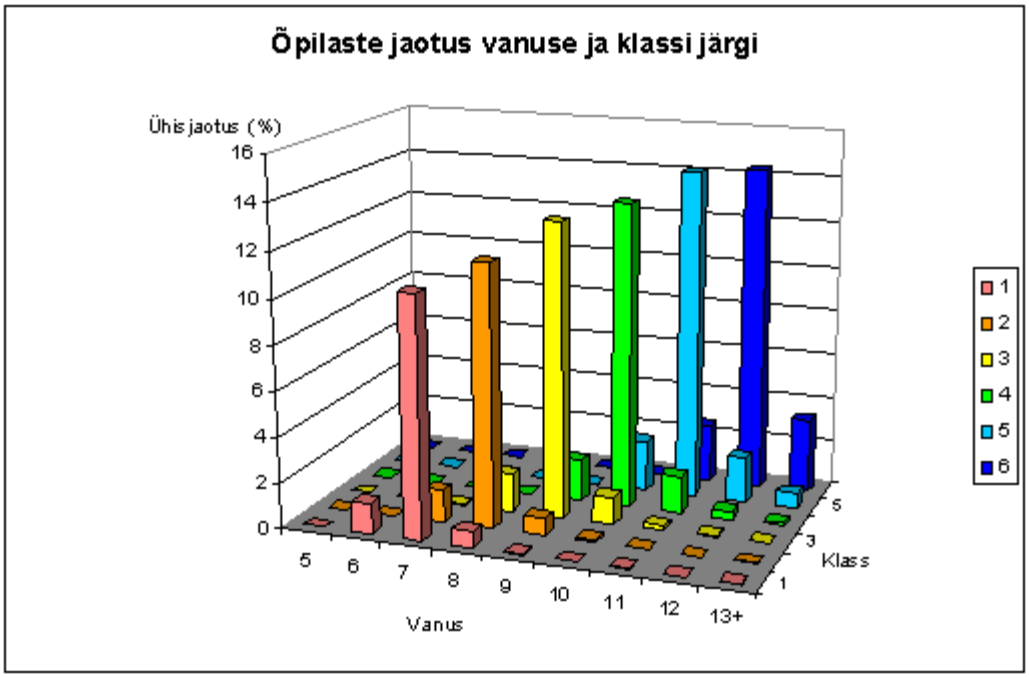
***Näide 3.2***

Alljärgnevas tabelis on esitatud Eesti kuni 6. klassi õpilaste ühisjaotus vanuse ja klassi järgi 1.09.2001..

Tõenäosused on esitatud, nagu empiiriliste jaotuste korral tavaks, protsentides.

| Vanus/ klass | 1 | 2 | 3 | 4 | 5 | 6 | Kokku |
|---|---|---|---|---|---|---|---|
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 1,34 | 0.01 | 0 | 0 | 0 | 0 | 1,35 |
| 7 | 10,63 | 1,44 | 0,02 | 0 | 0 | 0 | 12,09 |
| 8 | 0,71 | 11,57 | 1,71 | 0,02 | 0 | 0 | 14,01 |
| 9 | 0,04 | 0,79 | 12,95 | 1,83 | 0,05 | 0 | 15,66 |
| 10 | 0,01 | 0,11 | 1,17 | 13,4 | 2,31 | 0,04 | 17,04 |
| 11 | 0,01 | 0,02 | 0,23 | 1,66 | 14,49 | 2,56 | 18,97 |
| 12 | 0 | 0,01 | 0,06 | 0,36 | 2 | 14,33 | 16,76 |
| 13+ | 0,01 | 0,02 | 0,05 | 0,12 | 0,68 | 3,27 | 4,15 |
| Kokku | 12,75 | 13,97 | 16,19 | 17,39 | 19,53 | 20,2 | 100 |

Tabeli alumises ja parempoolses reas on esitatud vastavalt õpilaste jaotus vanuse järgi ja klassi järgi, need on juhusliku vektori komponentide jaotused ehk ühisjaotuse ***marginaaljaotused***. Paneme tähele, et ühisjaotusest on võimalik marginaaljaotusi leida vastavalt ridu või veerge pidi summeerimisel. Vastupidine – marginaaljaotuste järgi ühisjaotuse määramine – ei ole põhimõtteliselt võimalik. Juhusliku vektori jaotust saab näitlikustada mitmemõõtmelise tulpdiagrammi abil.




**Juhusliku vektori komponentide sõltumatus**

Juhusliku vektori (*X, Y*) komponendid on sõltumatud, kui iga *i* ja *j* korral kehtib tingimus:

$$p_{ij} = P(X = x_i) \cdot P(Y = y_j),$$

st et ühisjaotus avaldub täies ulatuses marginaaljaotuste korrutisena.

**Juhusliku suuruse funktsioon**

Juhuslikule suurusele võib rakendada mitmesuguseid funktsioone – teda logaritmida, korrutada konstandiga, astendada – tulemuseks on üldiselt kõneldes ikka juhuslik suurus, mille väärtused saab arvutada, rakendades vastavat funktsiooni algsele juhuslikule suurusele ning omistades nii saadud väärtustele endiste väärtuste tõenäosused. Seega on ***juhusliku suuruse funktsioon juhuslik suurus***. Kui mingi katse tulemuste abil on defineeritud mitu juhuslikku suurust, st juhuslik vektor, siis on võimalik defineerida juhusliku vektori komponentide summa, vahe, korrutise jne. ***Ka juhusliku vektori komponentide funktsioon on juhuslik suurus.*** Juhusliku suuruse või vektori funktsioonina defineeritud juhuslikul suurusel ei saa olla rohkem erinevaid väärtusi kui vastaval juhuslikul suurusel/ vektoril. Saadava juhusliku suuruse väärtuste tõenäosused arvutatakse nende jaoks soodsate katsetulemuste tõenäosuste põhjal.

**Järelemõtlemiseks**

1.Olgu antud kolm ühesugust kahe väärtusega juhuslikku suurust. Mitu erinevat väärtust on nende summal?

2.Olgu antud kolm kahe väärtusega juhuslikku suurust, kusjuures nende väärtused on erinevad. Mitu erinevat väärtust on nende summal? Kas see arv sõltub ka sellest, missugused on nende juhuslike suuruste väärtused?

3.Defineerige juhuslik suurus 1-sendise, 5-sendise ja 10-sendise mündi abil nii, et

juhusliku suuruse väärtuseks on peale jäänud (kirjapoolel paiknevate) arvude summa.

**Juhuslik suurus ja vektor. Jaotus ja tema esitused**

**DISKREETSED JAOTUSSEADUSED**

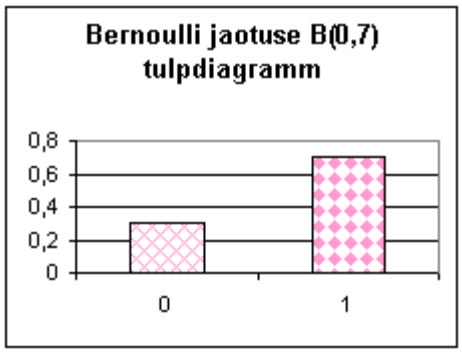
### Teoreetiline diskreetne jaotus

Teine võimalus diskreetse jaotuse defineerimiseks on määrata see mingi eeskirja/ valemi abil. Niisuguseid jaotusi nimetatakse ka ***jaotusseadusteks***. Kuigi põhimõtteliselt määrab jaotuse iga arvude [$p_i$ , $i$ = 1, ..., $m$] komplekt, mis rahuldab tõenäosusfunktsiooni tingimusi 1 ja 2, tuletatakse tavaliselt jaotused, lähtudes teatavatest sisulistest kaalutlustest vaadeldava juhusliku suuruse kohta. Järgnevalt vaatleme rida sagedamini kasutatavaid teoreetilisi diskreetseid jaotusi.

### Bernoulli jaotus

Bernoulli jaotusega juhuslikul suurusel on ainult kaks väärtust, need on 0 ja 1. Väärtuse 1 esinemise tõenäosus on tähistatud tähega $p$, $0 < p < 1$. Bernoulli jaotuse tavaliseks tähiseks on B($p$).

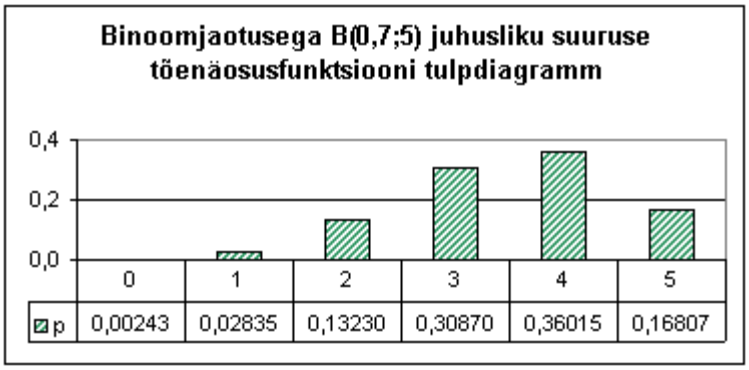
Tõenäosus $p$ on selle jaotuste pere parameetriks, erineva $p$ väärtuse korral saadakse üldiselt erinev jaotus. Bernoulli jaotusega juhuslik suurus $X$ on käsitletav teatava juhusliku sündmuse ***indikaator-funktsioonina***, mille väärtus 1 näitab, et sündmus toimus, väärtus 0 – et ei toimunud.




### Binoomjaotus

Vaatleme taas katset, mille tulemusena sündmus $A$ võib toimuda tõenäosusega $p$. Kordame seda katset $n$ korda ja loendame, mitu korda sündmus $A$ toimus. Sündmuse $A$ toimumise arv katseseeria jooksul on juhuslik suurus, mille väärtuseks on mingi täisarv $k$,

$$0 \le k \le n.$$




Seda juhuslikku suurust nimetatakse **binoomjaotusega** juhuslikuks suuruseks $B(p,n)$, kus jaotuse parameetriteks on tõenäosus $p$, $0 \le p \le 1$ ja katseseeria pikkus . Sündmuse $A$ mittetoimumise tõenäosuse tähiseks on $q$, $q = 1 - p$.

Seda juhuslikku suurust nimetatakse **binoomjaotusega** juhuslikuks suuruseks $B(p,n)$, kus jaotuse parameetriteks on tõenäosus $p$, ja katseseeria pikkus $n \le 1$. Sündmuse $A$ mittetoimumise tõenäosuse tähiseks on $q$, $q = 1 - p$.

Binoomjaotuse puhul on üsna lihtne leida tõenäosusfunktsiooni väärtusi $P(X = k)$,

$$P(X = k) = C_n^k p^k q^{n-k},$$

kus

$$C_n^k = \frac{n!}{k!\,(n-k)!}.$$

Tõepoolest, alati, kui katseseerias esineb katsetulemus $A$ täpselt $k$ korda, peab $A$ vastandsündmus esinema täpselt $n - k$ korda. Et katsed on eelduse kohaselt sõltumatud, siis on iga sellise katseseeria tõenäosus $p^k q^{n-k}$. Nüüd on vaja tehe veel selgeks, kui palju kõigi katseseeriate hulgas on niisuguseid seeriaid. Et $C_n^k$ on defineeritud, kui $k$-elemendiste kombinatsioonide arv $n$ elemendi hulgast, siis ongi selge, et sobivate jadade arv on $C_n^k$.

Ühtlasi on ka selge, et kõik tõenäosused on mittenegatiivsed, sest nad on saadud kolme positiivse suuruse korrutisena. Jääb veel üle kontrollida, kas kehtib tõenäosusfunktsiooni teine omadus, so võrdus

$$\sum_{i=0}^{n} C_n^i p^i q^{n-1} = 1.$$

Selle võrduse tõestamiseks kasutatakse ***binoomvalemina*** tuntud võrdust, mille kohaselt vasakul esitatud summa võrdub binoomi $(p + q)$ $n$-nda astmega. Et definitsiooni kohaselt $p + q = 1$, siis järelikult kehtib ka tõenäosusfunktsiooni teine omadus.

Binoomjaotus on üks olulisi jaotusi, mille abil kirjeldatakse paljusid reaalses elus toimuvaid nähtusi ja protsesse.

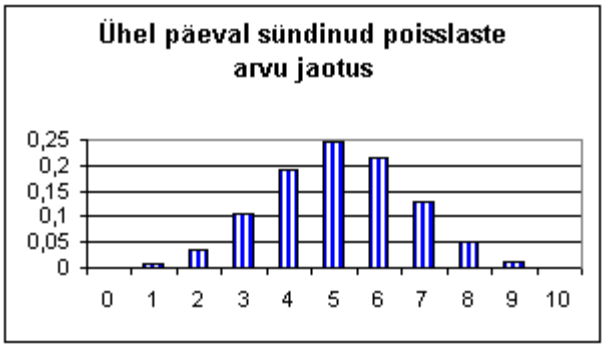
***Näide 3.3.***

Sünnitusmajas sündis ühel päeval 10 last Statistiliselt on kindlaks tehtud, et poisi

sünni tõenäosus on 0,514. Kui suur on tõenäosus selleks, et sündis vähemalt 7 poissi?

Lahendus: Poiste arv vastsündinute seas on binoomjaotusega $B(0{,}514;\ 10)$. Laiame vastavad tõenäosusfunktsiooni väärtused:

| k | P(k) |
|---|---|
| 7 | 0,01217 |
| 8 | 0,130567 |
| 9 | 0,001287 |
| 10 | 0,051783 |
| Kokku | 0,195807 |




Järelikult on vaadeldava sündmuse tõenäosus ligi 1/5.

**Diskreetne ühtlane jaotus**

Täringu abil saame samuti defineerida juhusliku suuruse, lugedes viskel pealelangenud silmade arvu juhusliku suuruse väärtuseks. Korrapärase täringu puhul on kõigi väärtuste saamise tõenäosus sama. Niiviisi on määratud diskreetne ühtlane jaotus, mille tõenäosusfunktsioon on:

$$\mathrm{P}(i) = 1/k,\ i = 1, \ldots, k,$$

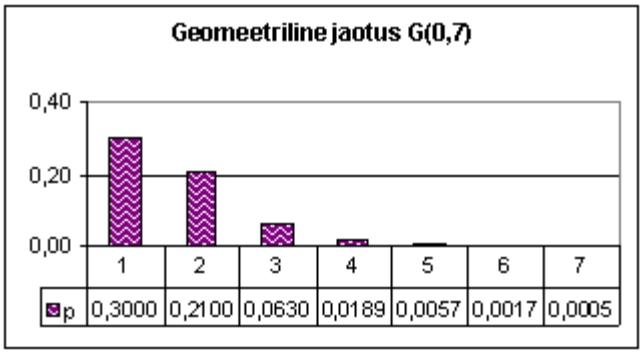
kus $k$ on juhusliku suuruse väärtuste arv; täringuviske puhul $k = 6$.

**Geomeetriline jaotus**

Olgu katsel sündmuse $A$ tõenäosus $p$. Oletame, et katset korratakse nii kaua, kuni sündmus $A$ toimub. Juhusliku suuruse väärtuseks on selle katse järjekorranumber, millal $A$ toimus. Sel juhul on tõenäosusfunktsioon lihtsalt leitav:

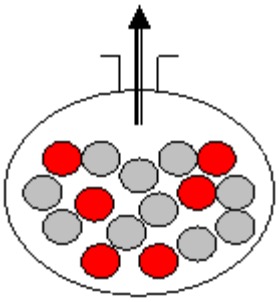
$$P(X = k) = pq^{k-1}.$$

Võrreldes eelmiste juhuslike suurustega on erinevuseks see, et geomeetrilise jaotuse puhul võib juhuslik suurus omandada kuitahes suuri väärtusi, seega on selle juhusliku suuruse **väärtuste hulk lõpmatu**. Geomeetrilise jaotuse parameetriks on tõenäosus $p$.




**Hüpergeomeetriline jaotus**

Paljud jaotused defineeritakse urniskeemi abil. Kasutame seda hüpergeomeetrilise jaotuse defineerimiseks. Sisaldagu urn $n$ kuuli, neist olgu $m$ punased ja $n - m$ mustad. Urnist võetakse juhuslikult $r$ kuuli. Juhusliku suuruse väärtuseks $k$ on selles komplektis saadud punaste kuulide arv. Pidades silmas kombinatsioonide arvu definitsiooni, on arusaadav, et



$$P(X = k) = \frac{C_m^k C_{n-m}^{r-k}}{C_n^r}.$$

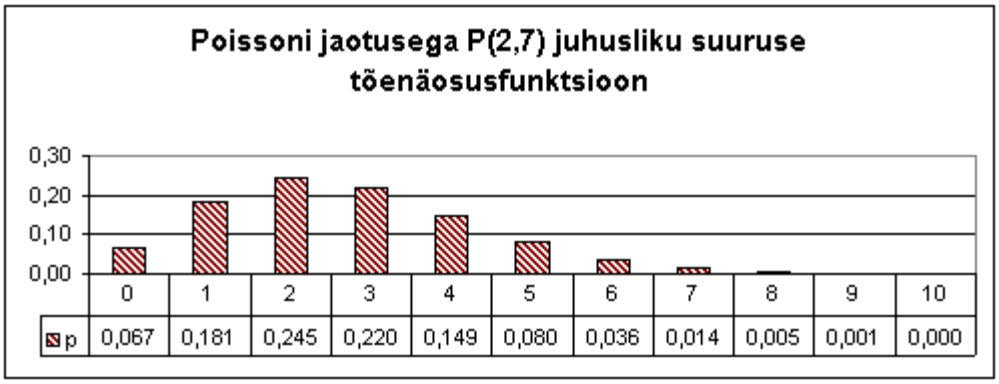
Hüpergeomeetrilise jaotuse parameetriteks on täisarvud $n$, $m$ ja $r$.

**Poissoni jaotus**

Vaatleme abonendile saabunud telefonikõnede arvu tunni aja jooksul. Oletame, et keskmiselt on nende kõnede arv $\lambda$. Kõned saabuvad sõltumatult, ning nende arv ei ole põhimõtteliselt piiratud. Sellistel eeldustel on igas konkreetses tunnis laekunud kõnede arv juhuslik suurus täisarvuliste väärtustega. Sellisel juhul sobib kõnede arvu kirjeldamiseks hästi Poissoni jaotus $P(\lambda)$, mille tõenäosusfunktsioon on järgmine:

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

Poissoni jaotusega juhuslikul suurusel on samuti kui geomeetrilise jaotusega juhuslikul suuruselgi lõpmata palju väärtusi, kuid arusaadavalt on suurel osal neist tõenäosused kaduvväikesed.

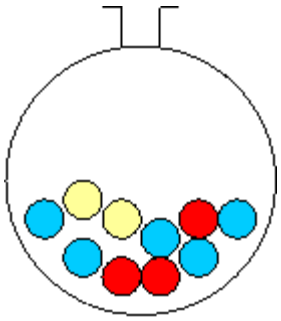



Poisssoni jaotus kuulub kõige olulisemate teoreetiliste jaotuste hulka, sest ta sobib kirjeldama väga paljusid reaalses elus toimuvaid nähtusi (radioaktiivset lagunemist, trükivigade paiknemist tekstis, klientide saabumist teenindusele jne). Poissoni jaotuse parameetriks on vaadeldavate nähtuste keskmine arv $\lambda$ ühikus.

**Multinomiaaljaotus**

Binoomjaotuse korral määrab juhusliku suuruse $X$ jaotuse ühe võimaliku sündmuse $A$ esinemise sagedus. Sel korral võiksime samaaegselt kõnelda ka teisest juhuslikust suurusest $Y$, mille väärtuse määrab sündmuse $A$ täiendsündmuse $A^c$ esinemissagedus. Niiviisi defineeritud juhuslik suurus $Y$ on $X$-i poolt üheselt määratud seosega $Y = n - X$.

Hoopis huvitavam olukord tekib aga siis, kui katsel on kolm võimalikku tulemust. Niisuguse katse võime taas defineerida urni abil, milles on kolme värvi kuule.



Sündmus $A$ on punase kuuli saamine, selle tõenäosus olgu $p_1$, käesoleva näite puhul $p_1 = 0{,}3$. Sündmus $B$ on sinise kuuli saamine; selle sündmuse tõenäosus on $p_2 = 0{,}5$. Sündmus $C$ on kollase kuuli saamine, selle sündmuse tõenäosus on $p_3 = 0{,}2$. Et alati võttel üks kolmest värvist ilmneb, peab kehtima võrdus $p_1 + p_2 + p_3 = 1$.

Selleks, et järgmistel võtetel oleks sündmustel sama tõenäosus, pannakse kuul iga võtte järel urni tagasi.
Nüüd on võimalik defineerida kolm juhuslikku suurust:

- $X_1$ näitab sündmuse $A$ esinemiste arvu $n$ katsel;
- $X_2$ näitab sündmuse $B$ esinemiste arvu $n$ katsel;
- $X_3$ näitab sündmuse $C$ esinemiste arvu $n$ katsel.

Juhuslikud suurused moodustavad kolmemõõtmelise juhusliku vektori, mille jaotust nimetatakse ***multinomiaaljaotuseks***. Multinomiaaljaotusega vektori komponendid omandavad täisarvulisi väärtusi $k_1$, $k_2$, $k_3$, mis rahuldavad tingimusi:

$$k_i \geq 0,$$

$$k_1 + k_2 + k_3 = n.$$

Selle vektori jaotust iseloomustab tõenäosusfunktsioon

$$P(k_1, k_2, k_3) = \frac{n!}{k_1! k_2! k_3!} p_1^{k_1} p_2^{k_2} p_3^{k_3}$$

***Näide 3.4.***

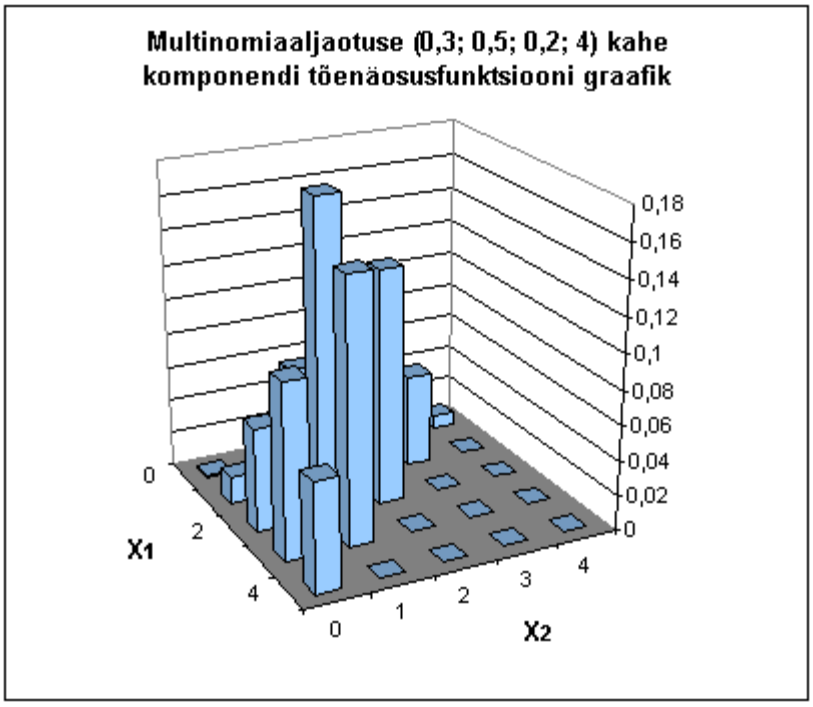
Olgu ülalkirjeldatud urnist võetud 4 kuuli (iga tulemus registreeritud ja tagasi pandud). Sellega on määratud juhuslik vektor jaotusega $M(0{,}3;\ 0{,}5;\ 0{,}2;4)$.

Vaatame kõigepealt, missugused on selle vektori võimalikud väärtused. On selge, et nii $X_1$ kui ka $X_2$ väärtused muutuvad 0 ja 4 vahel, kuid sealjuures on nad omavahel sõltuvad, ning ühtlasi määravad üheselt $X_3$ väärtuse, vt lisatud tabel, kus esimene veerg näitab $X_1$, esimene rida – $X_2$ väärtust ja tabeli vastavas lahtris on $X_3$ väärtus. Võimatutele väärtuskombinatsioonidele vastavad tühjad lahtrid.

| $X_1$ / $X_2$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 0 | 4 | 3 | 2 | 1 | 0 |
| 1 | 3 | 2 | 1 | 0 | |
| 2 | 2 | 1 | 0 | | |
| 3 | 1 | 0 | | | |
| 4 | 0 | | | | |

Järgmises tabelis on leitud iga väärtuskombinatsiooni tõenäosus, st on antud juhusliku vektori ühisjaotust esitav tõenäosusfunktsioon, ning seda illustreerib lisatud graafik.

| | 0 | 1 | 2 | 3 | 4 | Kokku |
|---|---|---|---|---|---|---|
| 0 | 0,0016 | 0,016 | 0,06 | 0,1 | 0,0625 | 0,2401 |
| 1 | 0,0096 | 0,072 | 0,18 | 0,15 | 0 | 0,4116 |
| 2 | 0,0216 | 0,108 | 0,135 | 0 | 0 | 0,2646 |
| 3 | 0,0216 | 0,054 | 0 | 0 | 0 | 0,0756 |
| 4 | 0,0081 | 0 | 0 | 0 | 0 | 0,0081 |
| Kokku | 0,0625 | 0,25 | 0,375 | 0,25 | 0,0625 | 1 |




Näeme, et kõige suurema tõenäosusega saadakse tulemus 1,2,1, st sündmus *A* esineb 1 kord, sündmus *B* – kaks korda ja sündmus *C* 1 kord.

**Järelemõtlemiseks**

1.Missuguse jaotusega juhuslik suurus saadakse, liites kolm sõltumatut ühesuguse Bernoulli jaotusega juhuslikku suurust?

2.Tõestada, et geomeetrilise jaotuse puhul on täidetud tõenäosusfunktsiooni põhiomadus.

3.Kas multinomiaaljaotusega juhusliku vektori komponendid saavad olla sõltumatud?

4.Missugused väärtused saavad olla suurima tõenäosusega geomeetrilise jaotuse puhul?

VASTUSED

**Juhuslik suurus ja vektor. Jaotus ja tema esitused**

**PIDEVAD JAOTUSSEADUSED**

Kui elementaarsündmuste hulk on lõpmatu ja mitteloenduv, siis on võimalik defineerida ka sellisesid juhuslikke suuruseid, mille väärtuste hulk on lõpmatu ja mitteloenduv. Selleks kasutame varasemaga sarnast määratlust – ***juhuslik suurus on elementaarsündmuse funktsioon***.

Kui juhusliku suuruse väärtused moodustavad reaalarvude hulgas intervalli (so lõigu koos otspunktidega, ühe otspunktiga või ilma) või katavad kogu reaaltelje või poole sellest, on tegemist ***pideva juhusliku suurusega***. Pideva juhusliku suuruse puhul on iga üksikväärtuse tõenäosus 0, seetõttu ei saa pideva juhusliku suuruse jaotust esitada tõenäosusfunktsiooni abil, vaid tuleb kasutada teisi funktsioone.

**Jaotusfunktsioon**

Olgu $X$ juhuslik suurus ja $x$ suvaline reaalarv. Funktsiooni

$$F(x)=P(X < x)$$

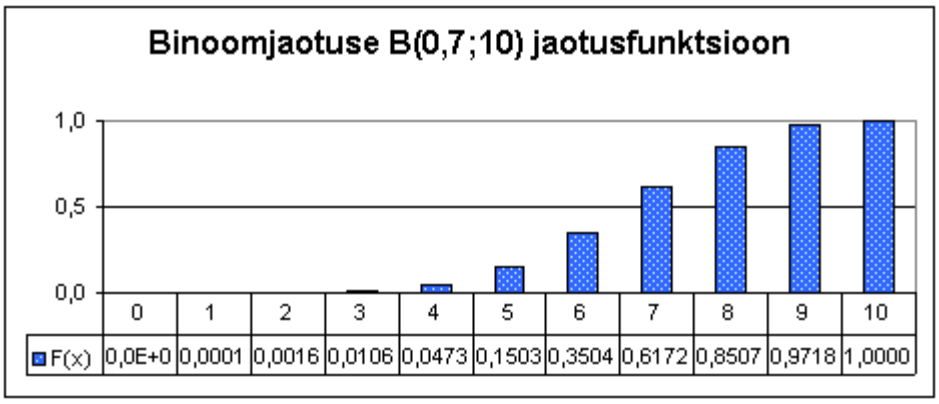
nimetatakse juhusliku suuruse **jaotusfunktsiooniks** ja see on üks juhusliku suuruse jaotuse esitusi. Jaotusfunktsioonil on rida olulisi omadusi:

1. $F(-oo) = 0$;
2. $F(oo) = 1$;
3. $F(x)$ on monotoonselt mittekahanev, st et kui $x < y$, siis

$$P(a \le X < b)= F(b)- F(a);$$

4. $F(x)$ on vasakult pidev.

Diskreetse juhusliku suuruse jaotusfunktsioon on treppfunktsioon, pideva juhusliku suuruse korral aga pidev funktsioon.



Jaotusfunktsiooni abil on võimalik leida kõigi selle juhusliku suuruse abil defineeritud sündmuste tõenäosusi. Võtme selleks annab võrdus:

$$P(a < X < b) = F(b) - F(a).$$

**Tihedusfunktsioon**

Pideva juhusliku suuruse korral on võimalik leida jaotusfunktsioonist tuletis. Jaotusfunktsiooni tuletist nimetatakse ***juhusliku suuruse tihedusfunktsiooniks***. Tihedusfunktsiooni tähistatakse tähega $f(x)$. Tihedusfunktsioonil on järgmised omadused, mis vahetult tulenevad jaotusfunktsiooni omadustest:

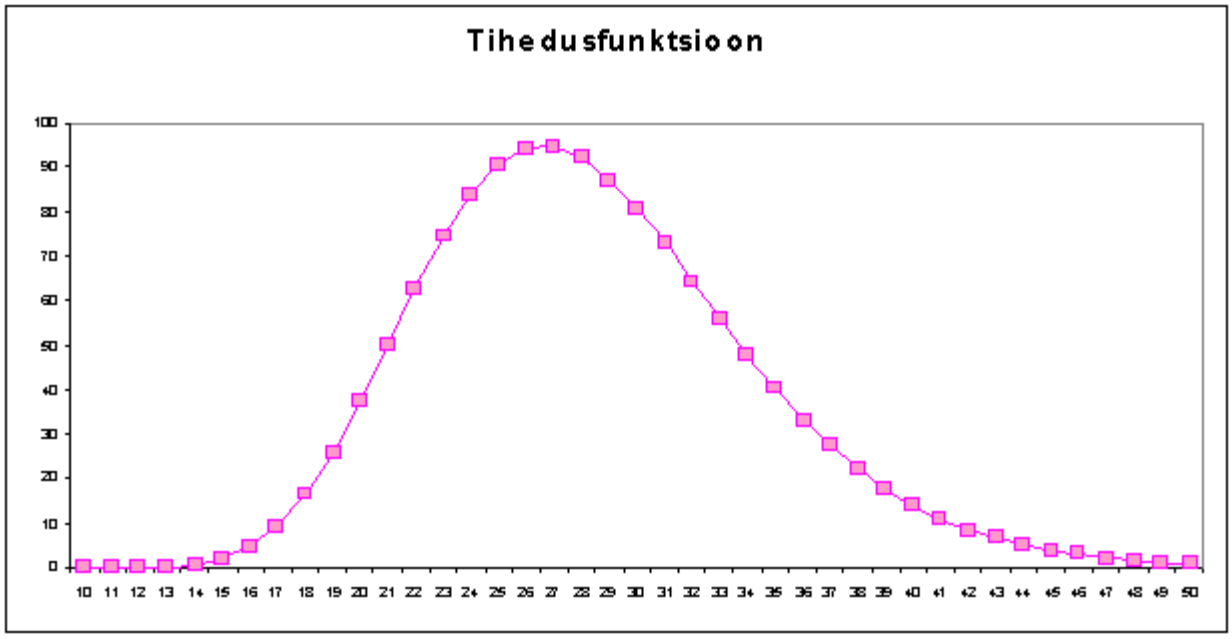
$$f(x) \geq 0,$$

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

Tihedusfunktsiooni abil on võimalik määrata juhusliku suuruse $X$ abil defineeritud sündmuste tõenäosusi:

$$P(a \leq X < b) = \int_{a}^{b} f(x)\,dx.$$

Kuna pidevate juhuslike suuruste korral on iga üksikpunkti tõenäosus 0, kehtib sama võrdus ka siis, kui tingimuses vahetada ranged ja mitteranged võrratused.




Lisatud joonisel on esitatud ühe teoreetilise jaotuse (nn Hadwigeri jaotuse) tihedusfunktsioon. Seda jaotust kasutatakse sünnitajate vanusejaotuse kirjeldamiseks. Järgnevalt tutvume mõne sageli kasutatava pideva jaotusega.

**Ühtlane jaotus**

Ühtlast jaotust on kõige lihtsam defineerida tihedusfunktsiooni järgi: juhuslik suurus $X$ on ühtlase jaotusega lõigul $[a, b]$, kui tema tihedusfunktsioon sellel lõigul on konstantne. Ühtlase jaotuse tähiseks on $U(a, b)$, kus parameetrid $a$ ja $b$ tähistavad vaadeldava lõigu otspunkte. Siis avaldub tihedusfunktsioon järgmiselt:

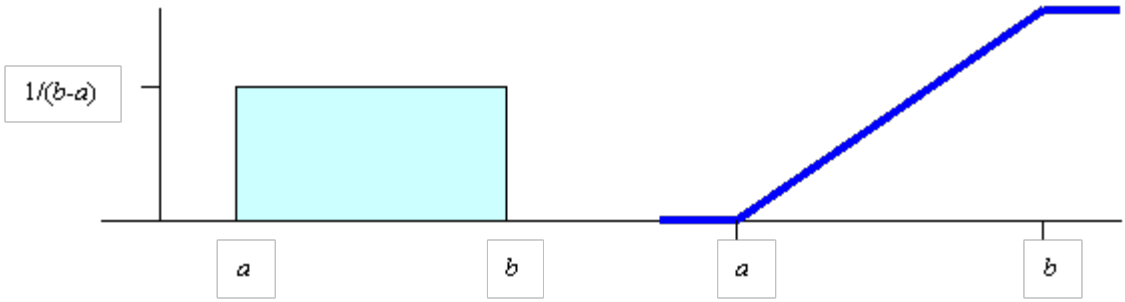
$$f(x)=\begin{cases}\frac{1}{b-a}, & \text{kui } a\le x\le b,\\ 0 & \text{muidu.}\end{cases}$$

Et jaotusfunktsioon avaldub tihedusfunktsiooni integraalina,

$$F(x)=\int_{-\infty}^{x} f(t)\,dt,$$

siis on ühtlase jaotuse jaotusfunktsiooni avaldis alljärgnev:

$$F(x)=\begin{cases}0, & \text{kui } x<a,\\ \frac{x-a}{b-a}, & \text{kui } a\le x\le b,\\ 1, & \text{kui } x>b.\end{cases}$$



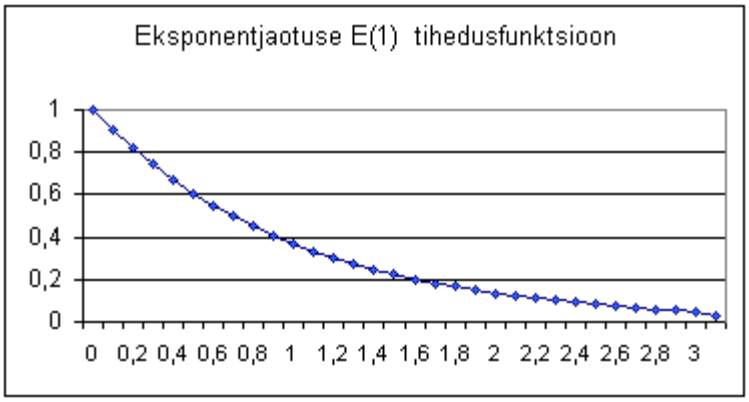

Paneme tähele, et kasutades geomeetrilist tõenäosust eeldatakse, et tegemist on ühtlase jaotusega.

**Eksponentjaotus**

Teine pidev jaotus, millega me tutvume, on eksponentjaotus. Selle tihedusfunktsioon on

$$f(x)=\begin{cases}0, & \text{kui } x\le 0\\ \lambda e^{-\lambda x}, & \text{kui } x>0.\end{cases}$$

Eksponentjaotust kasutatakse ooteaja kirjeldamiseks siis, kui sündmused toimuvad juhuslike ajavahemike järgi. Eksponentjaotuse tähiseks on $E(\lambda)$, kus $\lambda$ on jaotust iseloomustav parameeter, $\lambda>0$.




**Järelemõtlemiseks**

1.Tuletada eksponentjaotuse jaotusfunktsioon.

2.Kuidas muutub ühtlase jaotusega juhusliku suuruse tihedusfunktsioon lõigu [$a$, $b$] pikenemisel?

3.Missuguste argumendi väärtuste korral on eksponentjaotuse tihedusfunktsioon võrdne nulliga ja millal ta läheneb nullile?

4.Missugused sarnased ja erinevad jooned on diskreetse juhusliku suuruse tõenäosusfunktsiooni ja pideva juhusliku suuruse tihedusfunktsiooni vahel?

**Juhusliku suuruse jaotusparameetrid**

## JUHUSLIKU SUURUSE ASENDIKARAKTERISTIKUD

### Juhusliku suuruse asendikarakteristikute otstarve

Juhuslikku suurust iseloomustavad tema väärtused ja nende tõenäosused (esitatuna kas tõenäosus-, jaotus- või tihedusfunktsiooni kaudu), kuid sageli on kasulik juhusliku suuruse väärtuste paiknemist iseloomustada ***üheainsa arvuga***. Loomulikult annab üks arv edasi vaid osalise info, kuid see peaks olema mingis mõttes kõige iseloomulikum juhusliku suuruse paiknemist iseloomustav arv. Paljudel juhtudel sobib selliseks arvuks ***juhusliku suuruse keskväärtus.***

### Juhusliku suuruse keskväärtus

Juhusliku suuruse $X$ keskväärtust tähistatakse tihti sümboliga $EX$, vahel ka $\mu$. Diskreetse juhusliku suuruse keskväärtus on defineeritud valemiga

$$EX = \sum p_i x_i, \quad (1)$$

kus $x_i$ on juhusliku suuruse erinevad väärtused, $p_i$ nende tõenäosused ja summeeritakse üle kõigi juhusliku suuruse väärtuste – nende hulk on kas lõplik ($m$) või lõpmatu.

Erijuhul, kui on tegemist empiirilise jaotusega, st kui juhusliku suuruse jaotus on kindlaks tehtud $n$ katsetulemuse korral , siis kasutatakse tihti ka keskväärtuse avaldist:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i . \quad (2)$$

Valemid (1) ja (2) on samaväärsed; valemist (2) saadakse (1), rühmitades vaatlused vastavalt juhusliku suuruse erinevatele väärtustele ja kasutades empiirilise tõenäosuse määratlust $p_i = k_i/n$ , kus $k_i$ tähistab väärtuse $x_i$ esinemiste arvu katseseerias. Siinjuures võib empiiriline jaotus olla määratud niihästi statistilise kui ka klassikalise tõenäosusena. Sümbolit $\overline{x}$ kasutatakse peamiselt sel korral, kui jaotus on määratud statistiliselt.

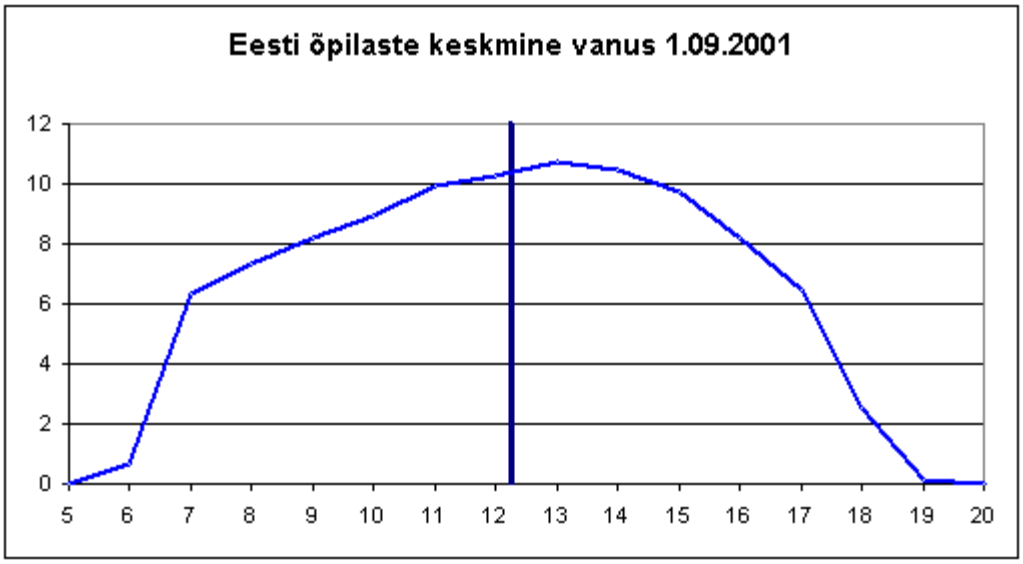
Pideva juhusliku suuruse keskväärtus avaldub integraalina

$$EX = \int_{-\infty}^{\infty} xf(x)dx,$$

kus $f(x)$ tähistab tihedusfunktsiooni ja integreerimine toimub üle juhusliku suuruse väärtuste piirkonna, st nende $x$ väärtuste korral, kus tihedusfunktsioon on positiivne. Tihedusfunktsiooni $f(x)$ määratakse empiiriliselt suhteliselt harva, enamasti eeldatakse, et tegemist mõne tuntud jaotusseadusega.

Juhusliku suuruse keskväärtust aitab interpreteerida järgmine arutelu: vaatleme juhusliku suuruse jaotuse mudelit mingist raskest materjalist (näiteks metallplaati, mis imiteerib tihedusfunktsiooni kuju, või varrast, kus paiknevad kuulid, mille kaal on võrdeline diskreetse jaotuse tõenäosustega). Sel juhul määrab kujundi raskuskese selle juhusliku suuruse keskväärtuse.

***Näide 4.1.***




Arvutades eesti kooliõpilase keskmise vanuse (vt andmed eelmisest loengust) valemi (1) järgi, saame selle väärtuseks 12,25 aastat. Märgime aga, et siin kasutati õpilaste vanuseid täisaastates. Eeldades sünnipäevade ühtlast jaotust aasta jooksul saaksime tulemuseks umber pooleaastase nihke (keskmiselt on 1. septembril 8-aastane laps 8,5-aastane jne). Seda arvestades võime kinnitada, et keskmine koolimineja oli 12,75-aastane ehk 12 aastat ja 9 kuud vana.

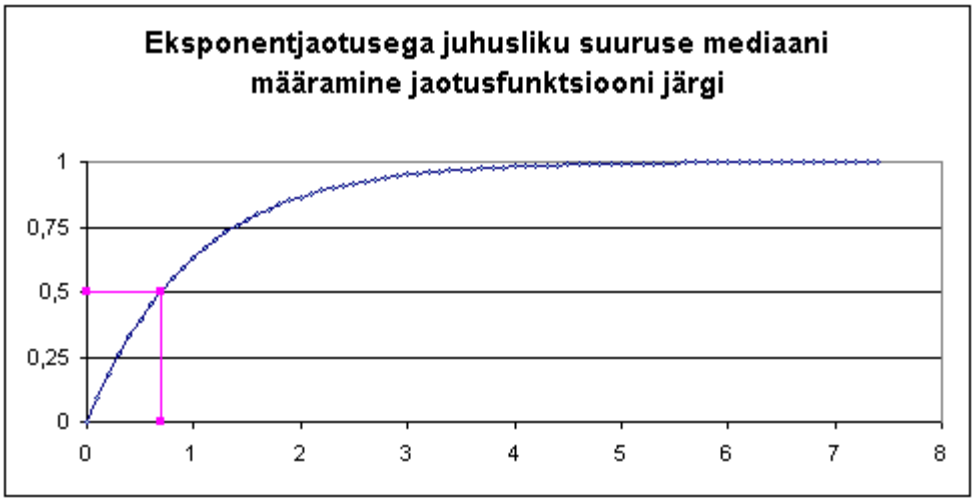
**Pideva juhusliku suuruse mediaan**

Teine väga sageli kasutatav asendikarakteristik on mediaan, mis on ühtlasi jaotuse keskpaik, so niisugune väärtus, millest vastava juhusliku suuruse väärtused on võrdse tõenäosusega nii suuremad kui ka väiksemad. Pideva juhusliku suuruse korral leidub alati punkt, mille tähistame sümboliga *med*, mille puhul kehtivad võrratused:

$$P(X < med) = P(X > med) = 0{,}5. \quad (3)$$

Kuna pideva juhusliku suuruse iga üksikpunkti tõenäosus on 0, siis kehtib täpselt sama võrdus sama punkti *med* korral ka mitterangete võrratuste korral.

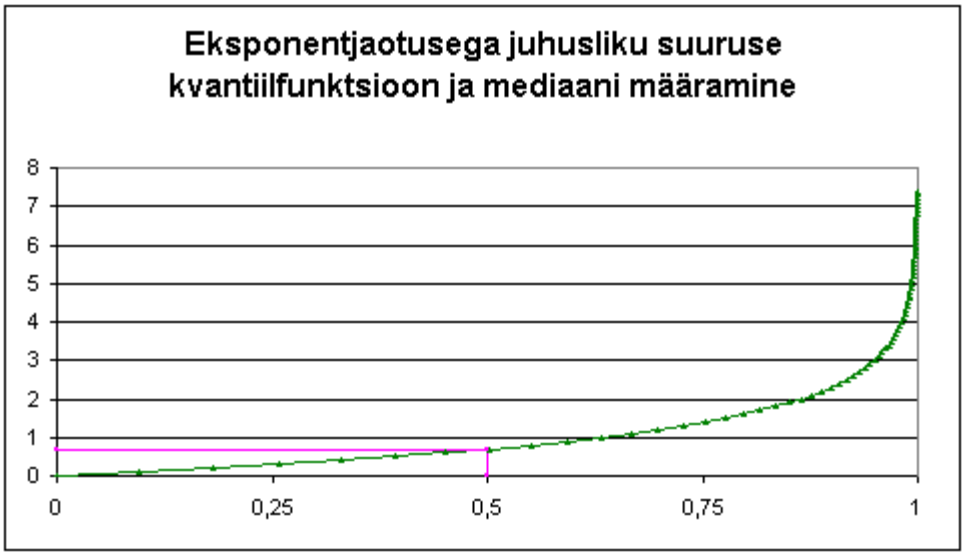
Mediaani leidmiseks on sobiv kasutada juhusliku suuruse jaotusfunktsiooni, sest mediaan on võrrandi $F(x) = 0{,}5$ lahendiks.




**Kvantiilfunktsioon**

Kui pideva juhusliku suuruse jaotusfunktsioon on kasvav, on tal olemas üheselt määratud pöördfunktsioon, mida me nimetame **kvantiilfunktsiooniks** $Q(p)$,

$$Q(p) = x \Leftrightarrow F(x)=p \quad (4)$$




Jaotusfunktsiooni argument $x$ on juhusliku suuruse väärtus, mis võib omandada põhimõtteliselt suvalisi väärtusi, jaotusfunktsiooni väärtus $p$ on tõenäosus, mis muutub 0 ja 1 vahel. Kvantiilfunktsioonil on vastupidi argumentideks tõenäosused ja funktsiooni enese väärtused võivad põhimõtteliselt omandada suvalisi reaalarvulisi väärtusi.

***Pideva juhusliku suuruse mediaaniks on tema kvantiilfunktsiooni väärtus kohal 0,5.***

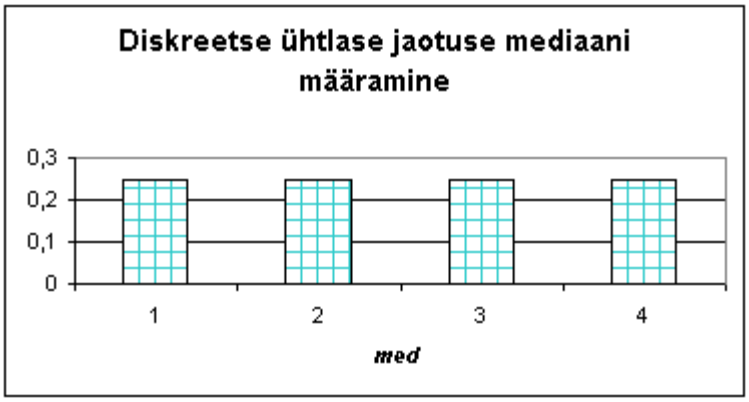
**Diskreetse juhusliku suuruse mediaan**

Diskreetse juhusliku suuruse **mediaani** defineerimiseks on mitu võimalust.

1. Diskreetse juhusliku suuruse mediaaniks loetakse seda juhusliku suuruse väärtust, millest väiksemaid ja suuremaid väärtusi omandab juhuslik suurus tõenäosusega, mis on väiksem või võrdne kui 0,5, st et võrduste (3) asemel nõutakse võrratustepaari (5) täidetust.

$$P(X < med) \leq 1{,}5;\ P(X > med) \leq 0{,}5 \quad (5)$$

On selge, et mõlemad tingimused ei saa olla võrdusena täidetud siis, kui nõuda, et mediaan on diskreetse juhusliku suuruse väärtus. Kuid selgub, et mõnikord võib niisuguseid juhusliku suuruse väärtusi, mis mõlemat võrratust (5) rahuldavad, olla mitu, vt näiteks alljärgnev joonis:




Siin sobivad definitsiooni järgi mediaaniks nii punkt 2 (millest on $X$ väiksem tõenäosusega 0,25 ja suurem tõenäosusega 0,5) ja punkt 3 (millest on $X$ väiksem tõenäosusega 0,5 ja suurem tõenäosusega 0,25). Kui tingimusi (5) rahuldab mitu juhusliku suuruse väärtust, siis tehakse üks alljärgnevatest otsustest:

• Valitakse mediaaniks juhusliku suuruse väärtus (üks neist väärtustest), mis tingimust (5) mingis mõttes kõige täpsemini rahuldab;

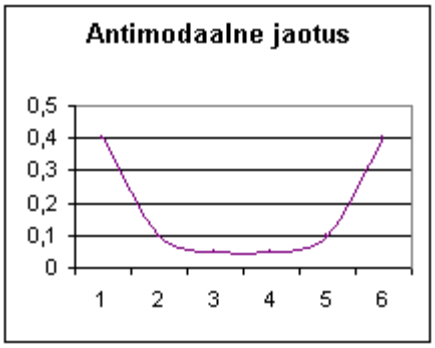
• Loetakse mediaaniks punkt, mis ei ole juhusliku suuruse väärtus, vaid paikneb nende vahel. See punkt kas määratakse lihtsalt tingimust (5) täitvate punktide aritmeetilise keskmisena või kasutatakse mingit interpolatsiooniskeemi.

Joonisel esitatud juhusliku suuruse mediaaniks loetakse tavaliselt punkt 2,5.

**Mood**

Diskreetse juhusliku suuruse mood on tema ***suurima tõenäosusega väärtus***. Pideva juhusliku suuruse mood on see tema väärtus, millele vastab ***tihedusfunktsiooni maksimum***. Näites 4.1 esitatud õpilaste vanusejaotuse mood on punktis 13, eelmises loengus näitena vaadeldud binoomjaotuse mood on punktis 7.

Igal juhuslikul suurusel ei tarvitse moodi üldse eksisteerida, näiteks selle kohta on pideva ja diskreetse ühtlase jaotusega juhuslikud suurused. On ka selliseid juhuslikke suurusi, millel on kaks moodi (tihedus- või tõenäosusfunktsiooni lokaalset maksimumi) – neid nimetatakse bimodaalseteks. Mitme moodiga juhuslikke suurusi nimetatakse multimodaalseteks. Kui mood paikneb juhusliku suuruse väärtustehulga otspunktides, räägitakse antimodaalsetest jaotusest.




Antimodaalsete jaotuste seas eristatakse vastavalt kujule U-, L- ja J- jaotusi, näiteks eksponentjaotus on L-jaotus.

**Juhusliku suuruse keskväärtuse omadusi**

***Keskväärtuse monotoonsus:***

Keskväärtus paikneb juhusliku suuruse minimaalse ja maksimaalse väärtuse vahel

$$min(x_i) \leq EX \leq max(x_i)$$

See omadus koosneb kahest võrratusest. Tõestame neist esimese diskreetse juhusliku suuruse puhul:

$$EX = \sum_i p_i x_i \geq \sum_i p_i \, min(x_i) = min\,(x_i) \sum_i p_i = min(x_i).$$

Teise võrratuse tõestus on samasugune, pideva juhusliku suuruse puhul tuleb aga summa omaduste asemel kasutada sarnaseid integraali omadusi.

***Konstandi keskväärtus:***

Konstandi keskväärtus võrdub sama konstandiga

$$Ec = c.$$

Selle omaduse tõestamisel saab kasutada sama mõttekäiku, mis eelmisegi valemi puhul.

***Keskväärtuse lineaarsus***

Selle omaduse tõestamiseks paneme kõigepealt tähele, et juhusliku suuruse lineaarfunktsiooni jaotus on esialgse juhusliku suurusega üheselt määratud.:

$$P(a + bX = a + bx) = P(X = x).$$

Edasi rakendatakse keskväärtuse definitsiooni juhuslikule suurusele $a + bX$.

$$E(a + bX) = a + bEX.$$

***Keskväärtuse aditiivsus***

Kahe juhusliku suuruse summa keskväärtus võrdub nende juhuslike suuruste keskväärtuste summaga.

$$E(X + Y) = EX + EY$$

***Sõltumatute juhuslike suuruste keskväärtuse multiplikatiivsus***

Kui juhuslikud suurused $X$ ja $Y$ on sõltumatud, siis võrdub nende korrutise

keskväärtus nende juhuslike suuruste keskväärtuste korrutisega:

$$E(XY) = EX \cdot EY.$$

Kaks viimast omadust vajavad korrektseks tõestamiseks juhusliku vektori funktsiooni jaotuse kasutamist, mistõttu need käesolevasse kursusesse ei mahu.
Põhimõtteliselt võib juhuslikul suurusel keskväärtus ka puududa, nimelt siis, kui keskväärtust määrav summa (1) või integraal ei koondu.

**Mediaani ja moodi omadused**

Mediaanil on keskväärtusega ühised monotoonsuse ja lineaarsuse omadused. Konstandi mediaan ei ole sisuliselt määratud, kuid tinglikult võib konstandi puhul sama konstanti lugeda ka mediaaniks, samuti nagu keskväärtusekski.
Moodi puhul kehtivad samuti monotoonsuse ja lineaarsuse omadused, tinglikult võib ka konstantse jaotuse korral seda konstanti lugeda jaotuse moodiks.
Üldiselt aga ei kehti mediaani ja moodi puhul aditiivsuse ja multiplikatiivsuse omadused.

Keskväärtus, mediaan ja mood on mittejuhuslikud suurused, nende väärtus ei sõltu katsetulemustest.

**Järelemõtlemiseks**

1.Kas täisarvuliste väärtustega juhusliku suuruse keskväärtus on alati täisarv?
2.Kas täisarvuliste väärtustega juhusliku suuruse mood on alati täisarv?
3.Tõestada, et mittenegatiivsete väärtustega juhusliku suuruse keskväärtus, mediaan ja mood on mittenegatiivsed.
4.Missugune on keskväärtuse ja mediaani vahekord siis, kui juhuslik suurus on sümmeetriline (st, et kas tõenäosus- või tihedusfunktsioon on sümmeetriline).
5.Mida saab öelda $X$ ja $Y$ keskväärtuse kohta, kui on teada, et $X < Y$ selles mõttes, et iga katse tulemusel on $X$ väärtus väiksem kui $Y$ väärtus?
6.Kas lõpliku hulga väärtustega juhuslikul suurusel on alati keskväärtus olemas?
7.Kas tõkestatud väärtushulgaga juhuslikul suurusel on keskväärtus alati olemas?
8.Millega võrdub nullpunkti suhtes sümmeetrilise tihedusfunktsiooniga/ tõenäosusfunktsiooniga juhusliku suuruse keskväärtus?
9.Arvutada järgmiste jaotuste keskväärtus:
• Bernoulli jaotus,
• Binoomjaotus,
• Poissoni jaotus,
• Ühtlane jaotus,
• Eksponentjaotus.
10.Leida ühtlase jaotuse ja eksponentjaotuse mediaan.
11.Missugustel ülesandes 9 nimetatud jaotustest eksisteerib mood ja kus see paikneb?

VASTUSED

**Juhusliku suuruse jaotusparameetrid**

**JUHUSLIKU SUURUSE HAJUVUSKARAKTERISTIKUD**

**Juhusliku suuruse hajuvus**

Juhuslikku suurust eristab mittejuhuslikust e determineeritud suurusest e. konstandist nimelt hajuvus – kui mittejuhusliku suuruse väärtus on alati ühesugune, siis juhusliku suuruse väärtus sõltub katsetulemusest, st ***varieerub*** e. ***hajub***. Juhusliku suuruse hajuvust iseloomustavad hajuvuskarakteristikud. Võiks öelda nii, et juhuslik suurus, mille hajuvus on väike, erineb suhteliselt vähe determineeritud suurusest, seevastu suure hajuvusega juhuslik suurus on väga erinev konstandist.

**Dispersioon**

Juhuliku suuruse *X* dispersioon *DX* on üks tuntumaid hajuvuse karakteristikuid, millel on oluline osa eriti teoreetilistes arutlustes. Dispersiooni käsitlemisel lähtutakse tõsiasjast, et juhusliku suuruse hajuvust iseloomustab juhusliku suuruse väärtuste ***hälbimine keskväärtuse*** ümber. Juhusliku suuruse dispersioon defineeritaksegi kui hälvete ruutude keskväärtus, kusjuures ruutu võtmise üks põhjuseid on see, et muuta negatiivsed ja positiivsed hälbed samaväärseteks

$$DX = E(X-EX)^2 \quad (6).$$

Dispersiooni arvutusvalem on diskreetse juhusliku suuruse korral:

$$DX = \sum_i (x_i - EX)^2 p_i$$

ja pideva juhusliku suuruse korral:

$$DX = \int (x - EX)^2 f(x) dx.$$

**Dispersiooni omadused**

Dispersioonil on rida omadusi, mis tulenevad peaaegu vahetult keskväärtuse omadustest.

***Dispersiooni mittenegatiivsus***

Kehtib võrratus

$$DX \geq 0,$$

mis järeldub dispersiooni definitsioonist ja keskväärtuse monotoonsuse omadusest.

***Konstandi dispersioon***

Konstantse (so mittejuhusliku) suuruse dispersioon on null.

See tuleneb vahetult keskväärtuse teisest omadusest, mille kohaselt konstandi keskväärtus võrdub selle konstandiga, järelikult on sel juhul hälve võrdne nullliga, ja ka hälbe ruudu keskväärtus on siis null.

***Dispersiooni invariantsus nihke suhtes***

$$D(X + c) = D(X).$$

See omadus tuleb sellest, et $E(X + c) = EX + c$, mistõttu dispersiooni avaldis konstandi liitmisel juhuslikule suurusele ei muutu.

***Dispersiooni ruuthomogeensus***

Kui juhuslikku suurust $X$ korrutada konstandiga $c$, siis muutub tema dispersioon $c^2$ korda:

$$D(cX) = c^2DX.$$

Ka see omadus järeldub vahetult definitsioonist ja keskväärtuse homogeensusest.

***Dispersiooni aditiivsus***

Kui juhuslikud suurused $X$ ja $Y$ on sõltumatud, siis kehtib seos:

$$D(X + Y) = DX + DY.$$

Tõestuseks arvutame

$$E(X + Y - E(X + Y))^2 = E((X-EX)+(Y-EY))^2 =$$
$$= E(X-EX)^2+E(Y-EY)^2+2E((X-EX)(Y-EY)) = DX + DY,$$

sest viimane liidetav võrdub keskväärtuse omaduste tõttu nulliga:

$$E((X-EX)(Y-EY)) = E(X-EX)\cdot E(Y-EY) = (EX - EX) \cdot (EY - EY) = 0.$$

**Dispersiooni avaldis momentide kaudu**

Juhusliku suuruse $k$-järku momendiks $m_k$ nimetatakse tema $k$-nda astme keskväärtust

$$m_k = E(X^k),$$

mille arvutuseeskiri on vastavalt diskreetse ja pideva juhusliku suuruse korral alljärgnev:

$$m_k=\sum x_i^k p_i, \quad m_k=\int x^k f(x)dx.$$

Definitsioonist järeldub, et esimest järku momendiks on keskväärtus. Dispersiooni saab arvutada ka esimest ja teist järku momentide kaudu:

$$DX=m_2-(m_1)^2.$$

ÜLESANNE

**Standardhälve**

Praktiliste ülesannete lahendamisel kasutatakse juhusliku suuruse $X$ hajuvuse

karakteristikuna tavaliselt standardhälvet, mille tähiseks on $\sigma$ ja mis mis määratletakse kui ruutjuur dispersioonist $DX = \sigma^2$. Ka standardhälve (samuti kui dispersioongi) loetakse positiivseks suuruseks, ning tema eeliseks võrreldes dispersiooniga on see, et ta on sama dimensiooniga kui juhuslik suurus $X$.

**Variatsioonikordaja**

Variatsioonikordaja $V$ mõõdab juhusliku suuruse suhtelist hajuvust tema keskväärtuse suhtes,

$$V = \sigma / EX.$$

Variatsioonikordaja väljendatakse sageli protsentides. Variatsioonikordajat on korrektne arvutada vaid selliste juhuslike suuruste jaoks, mille kõik väärtused on positiivsed.

**Variatsiooniulatus e. haare**

Kui juhusliku suuruse väärtustehulk on tõkestatud, siis on sel juhuslikul suurusel ka vähim väärtus $\min(x)$ ja suurim väärtus $\max(x)$. Juhusliku suuruse suurima ja vähima väärtuse vahet nimetatakse selle juhusliku suuruse haardeks. Haare on üks juhusliku suuruse jaotuse hajuvust iseloomustavaid näitajaid, suurema hajuvusega juhusliku suuruse (jaotuse) korral on ka tema haare suurem, väiksema hajuvuse puhul on väiksem ka haare.

**Kvantiilid**

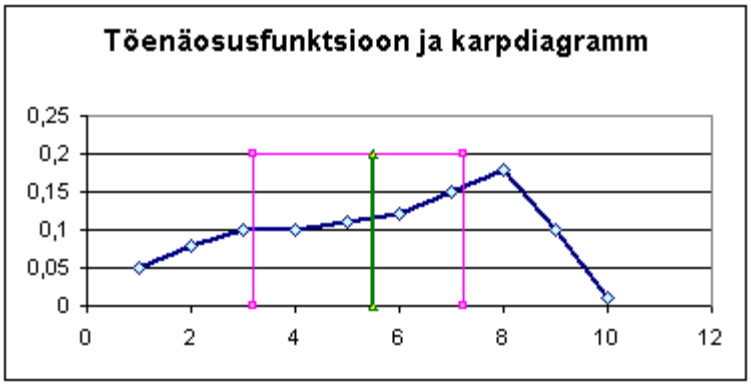
Olgu $q$ mingi arv 0 ja 1 vahel. Pideva juhusliku suuruse $q$-kvantiil defineeritakse jaotusfunktsiooni abil kui võrrandi $F(x) = q$ lahend ehk kui kvantiilfunktsiooni väärtus $Q(q)$. Diskreetse juhusliku suuruse korral otsitakse juhusliku suuruse väärtust $x$, mis rahuldab tingimusi

$$P(X \leq x) \geq q, \; P(X \geq x) \geq 1-q.$$

Kui niisuguseid väärtusi on rohkem kui üks, valitakse välja üks, mis tingimusi kõige paremini täidab või rakendatakse interpolatsiooni. Kvantiili määratlusest järeldub, et mediaan on 0,5-kvantiil.

**Kvartiilid ja kvartiilhaare**

Alumiseks kvartiiliks nimetatakse 0,25-kvantiili, ülemiseks kvartiiliks 0,75-kvantiili. Ülemise ja alumise kvartiili vahe moodustab kvartiilhaarde, mis samuti on üks hajuvuse karakteristikuid – mida suurem on juhusliku suuruse kvartiilhaare, seda suurema hajuvusega ta on.




Andmeanalüüsis kujutatakse kvartiile ja mediaani sageli graafiliselt nn karpdiagrammi abil, kus karbi otsi tähistavad lõigud märgivad vastavalt alumist ja ülemist kvartiili ning lisaks on märgitud ka mediaan. Sageli lisatakse karbile veel "vuntsid", mis näitavad väikseimat ja suurimat mõõdetud väärtust.

**Teised kvantiilid**

Samuti nagu kvartiile vaadeldakse sageli komplektina – alumine kvartiil ehk ¼-kvantiil, mediaan ehk 2/4-kvantiil ja ülemine kvartiil ehk ¾-kvantiil, defineeritakse ka rida teisi kvantiilikomplekte:

***Kvintiilid***

Kvintiilide, so 1/5-, 2/5, 3/5 ja 4/5-kvartiilide abil jaotatud väärtushulka vaadeldakse tihti sotsiaal- ja majandusuuringute puhul.

***Detsiilid***

Detsiilide, so 0,1-, 0,2- ... ja 0,9-kvartiilide abil jaotatud väärtushulka käsitletakse samuti sageli sotsiaal- ja majandusuuringute puhul, kus näiteks isikud või leibkonnad jaotatakse sissetulekute järgi detsiilideks.

***Protsentiilid e. sentiilid***

Protsentiile, so 0,01-, 0,02, ..., 0,99-kvartiile kasutatakse sagedamini biostatistikas.

***Järelemõtlemiseks***

1.Kuidas avaldub kahe sõltumatu juhusliku suuruse *X* ja *Y* summa standardhälve? Vahe standardhälve?

2.Arvutada järgmiste jaotuste dispersioonid, standardhälbed ja variatsioonikordajad.

- Bernoulli jaotus
- Binoomjaotus
- Poissoni jaotus
- Ühtlane jaotus.

3.Missugusel ülalnimetatud jaotustest on olemas lõplik haare ja kui suur see on?

VASTUSED

## Juhusliku suuruse jaotusparameetrid

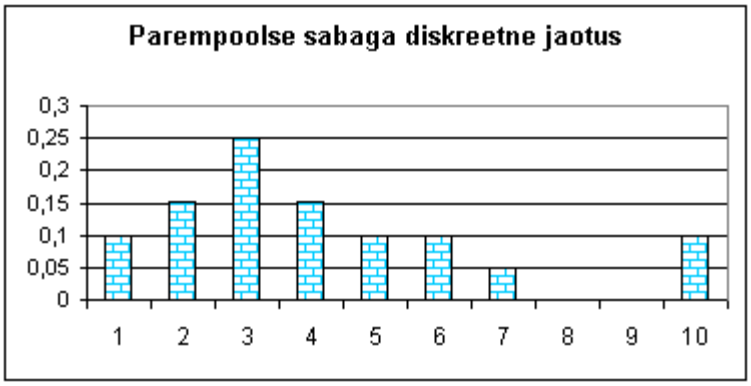
## JAOTUSE KUJU ISELOOMUSTAVAD KARAKTERISTIKUD

### Jaotuse sümmeetrilisus

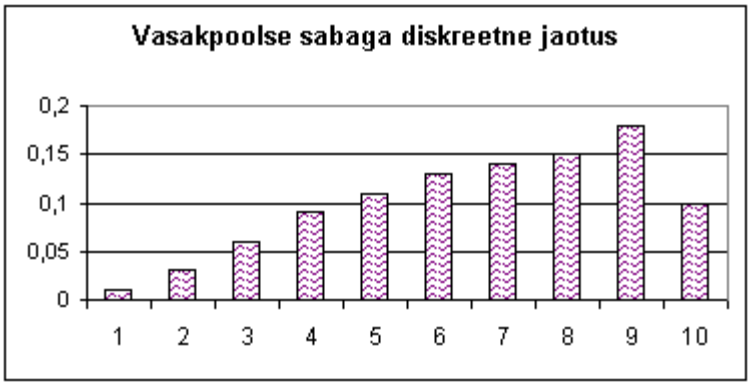
Kui juhusliku suuruse tõenäosus- või tihedusfunktsioon on (mingi väärtuse) suhtes sümmeetriline, siis öeldakse, et juhuslik suurus ja ühtlasi tema jaotus on sümmeetriline. Jaotuse sümmeetrilisust mõõdab asümmeetriakordaja

$$a = \frac{E(X-EX)^3}{DX^{3/2}}.$$

Sümmeetrilise juhusliku suuruse korral on asümmeetriakordaja võrdne nulliga. Kui jaotus on suurte väärtuste suhtes välja venitatud, tal on "raske saba" paremal, siis on asümmeetriakordaja positiivne, vt ülemine joonis ($EX$ = 4,2, $DX$ = 2,5, $a$ = 1,02).




Kui aga "raske saba" on vasakul, siis on asümmeetriakordaja negatiivne, vt alumine joonis ($EX$ = 6,74, $DX$ = 2,27, $a$ = –0,42).

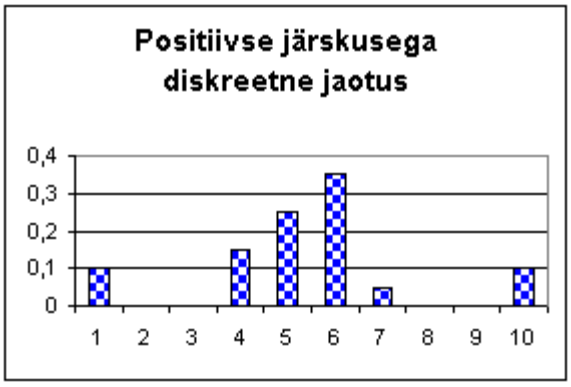


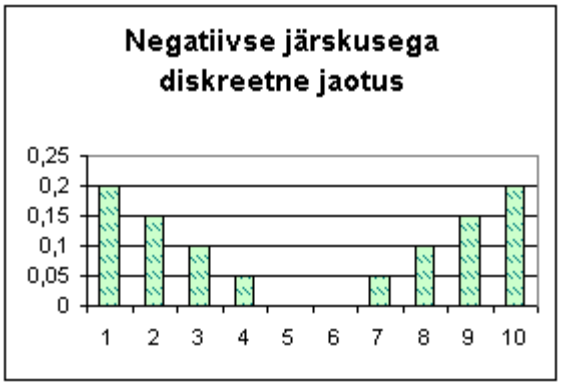

Asümmeetriakordaja võib omandada suvalisi väärtusi.

**Juhusliku suuruse kuju järskus/ lamedus**

Juhusliku suuruse jaotuse "raskete" ja kaugele ulatuvate sabadega käib tihti kaasas suhteliselt terav, kontsentreeritud väärtustega tipp. Niisugust jaotust nimetatakse järsuks, tal on positiivne järskuse kordaja ehk ekstsess $e$ (vt vasakpoolne joonis, $e=0{,}86$). Seevastu jaotust, mille väärtused on tõkestatud ja millel puudub selgelt eristuv tipp (või on neid mitu), nimetatakse lamedaks, sellise jaotuse ekstsess on negatiivne, vt parempoolne joonis, $e=-1{,}76$). Ekstsessi avaldis on:

$$e=\frac{E(X-EX)^4}{DX^2}-3$$




**Järelemõtlemiseks**

1.Kui juhusliku suuruse dispersioon võrdub nulliga, mida saab sel juhul öelda kvantiilide kohta?

2.Kuidas paiknevad omavahel kvartiilid ja kvintiilid (näidata suurusvahekorrad?)

3.Variatsioonikordaja arvutamisel eeldatakse, et juhusliku suuruse väärtused on mittenegatiivsed. Miks?

4.Missugune jaotuseparameeter on sümmeetrilise juhusliku suuruse sümmeetriakeskmeks? Mida saab öelda sümmeetrilise juhusliku suuruse kvantiilide kohta?

5.Missugune on ekstsessi minimaalne võimalik väärtus? Miks?

6.$X$ on Bernoulli jaotusega $B(p)$. Leida tema asümmeetria kordaja ja ekstsess. Missuguse $p$ väärtuse korral on Bernoulli jaotus sümmeetriline?

7.Leida ühtlase jaotuse kvartiilid, asümmeetria kordaja ja ekstsess.

**Juhusliku suuruse jaotusparameetrid**

**TŠEBÕŠEVI VÕRRATUS JA SUURTE ARVUDE SEADUSE TÕESTUS**

**Juhusliku suuruse standardiseerimine**

Selleks, et muuta erinevatel skaaladel mõõdetud juhuslikud suurused omavahel võrreldavateks, kasutatakse standardiseerimist. Kui juhusliku suuruse $X$ keskväärtus on $EX = \mu$ ja standardhälve on $\sigma$, siis vastav standardiseeritud juhuslik suurus $X_0$ on

$$X_0 = \frac{X-\mu}{\sigma}.$$

Standardiseeritud juhusliku suuruse keskväärtus on 0 ja dispersioon ning standardhälve on 1.

**Tšebõševi võrratus**

Tšebõševi võrratus näitab, et standardiseeritud juhusliku suuruse puhul esineb suuri hälbeid suhteliselt väikese tõenäosusega – siit järeldub, et sama kehtib suvalise juhusliku suuruse korral, kui me võrdleme tema hälbeid tema standardhälbega.

***Teoreem***

Iga juhusliku suuruse $X$ ja positiivse konstandi $c$ korral kehtib võrratus:

$$P(|X - EX| \geq c) \leq \frac{DX}{c^2}.$$

***Tõestus***

Esitame tõestuse diskreetse juhusliku suuruse jaoks; see on analoogia põhjal üle kantav ka pideva juhusliku suuruse juhule. Valime vabalt suuruse $c$, kirjutame välja diskreetse juhusliku suuruse dispersiooni avaldise ja teisendame seda alljärgnevalt:

$$DX = \sum_i p_i (x_i - EX)^2 = \sum_{|x-EX| \geq c} p_i (x_i - EX)^2 + \sum_{|x-EX| < c} p_i (x_i - EX)^2.$$

Asendades esimeses liidetavas avaldise $|X - EX|$ konstandiga $c$ avaldis kindlasti väheneb. Samuti väheneb avaldis siis, kui jätta ära teine, alati mittenegatiivne liidetav. Seega saame võrratuse:

$$DX \geq c^2 \sum_{|x-EX| \geq c} p_i = c^2 P(|X - EX| \geq c),$$

mis avaldise esimese ja viimase liikme jagamisel positiivse konstandiga $c^2$ annabki soovitava tulemuse.

**Suurte hälvete tõenäosused**

Võttes Tšebõševi võrratuses $c = a\sigma$, kus $a > 1$, saame võrratuse:

$$P(|X - EX| \geq a\sigma) \leq \frac{1}{a^2}.$$

Siit järeldub, et standardhälbest kaks korda suuremate hälvete tõenäosus ei saa olla suurem kui ¼ ja 10 korda standardhälvet ületavate hälvete tõenäosus ei ületa üht sajandikku. Enamuse juhuslike suuruste korral on aga suurte hälvete tõenäosused veel palju väiksemad kui seda lubaks Tšebõševi võrratus.

**Tõenäosuse järgi koondumine**

Juhuslike suuruste jada koondumisel on oluline see, et jadas edasi minnes (indeksi suurenedes) järjest väheneb suurte hälvete tõenäosus. Veelgi enam, igasuguse suurusega hälbe tõenäosus muutub jadas küllalt kaugele minnes järjest väiksemaks. Nii on defineeritud tõenäosuse järgi koondumine.

Juhuslike suuruste jada $X_n$ koondub tõenäosuse järgi juhuslikuks suuruseks $X$ siis, kui iga positiivsete suuruste paari $\varepsilon$ ja $\delta$ korral leidub selline indeks $n$, et kehtib võrratus:

$$P(|X_n - X| > \varepsilon) < \delta,$$

kui $n > n^*$. Tõenäosuse järgi koondumist tähistatakse sümboliga:

$$X_n \xrightarrow{p} X.$$

Juhuslike suuruste jada piirväärtuseks võib olla nii juhuslik suurus $X$ kui ka konstant.

**Suurte arvude seadus**

***Teoreem***

Olgu katse tulemusena esineva sündmuse $A$ tõenäosus $P(A) = p$. Kui seda katset on sooritatud $n$ korda, siis olgu sündmuse $A$ esinemise suhteline sagedus $S_{n/n}$. Katseseeria lõpmatul pikendamisel koondub suhteline sagedus tõenäosuse järgi selle sündmuse tõenäosuseks:

$$\frac{S_n}{n} \xrightarrow{p} p.$$

***Tõestus***

Tõenäosuse järgi koondumise definitsiooni kohaselt on meil tarvis näidata,
et iga $\varepsilon$ ja $\delta$ korral leidub selline indeks $N$, et kui $n > N$, siis kehtib võrratus

$$P\left(\left|\frac{S_n}{n} - p\right| \geq \varepsilon\right) < \delta.$$

Arvutame juhusliku suuruse $S_n/n$ keskväärtuse ja dispersiooni:

• Et $S_n$ on $n$ Bernoulli jaotusega juhusliku suuruse summa ja iga Bernoulli jaotusega juhusliku suuruse puhul $EX = p$, siis $E(S_n/n) = np/n = p$.

• Bernoulli jaotusega juhusliku suuruse dispersioon on $p(1–p)$. See suurus ei ületa kunagi murdu 0,25. Arvestades, et katsed on sõltumatud, saame $D(S_n) = 0{,}25n$ ning $D(S_n/n) = 0{,}25/n$.

• Kasutame nüüd suhtelise sageduse $S_n/n$ jaoks Tšebõševi võrratust, mille kohaselt iga puhul kehtib võrratus:

$$P\left(\left|\frac{S_n}{n}-p\right|\geq\varepsilon\right)<\frac{0{,}25}{n\varepsilon^2}.$$

Kui ja $\delta$ on antud, tuleb *N* määrata nii, et kehtiks võrratus

$$\frac{0{,}25}{n\varepsilon^2}<\delta\Leftrightarrow N>\frac{1}{4\delta\varepsilon^2}.$$

Sellega on teoreem tõestatud.

**Suhtelised sagedused pika katseseeria vältel**

Suurte arvude seaduse tähtsus on kahesuunaline.

• Ühelt poolt annab see loogilise ja teoreetilise aluse statistilise tõenäosuse kasutamiseks.

• Teiselt poolt näitab ka seda, milleks üldse on tõenäosuse teadmine kasulik. Teades sündmuste tõenäosusi, on võimalik hinnata, kui palju on lootust nende toimumiseks, kui pikki katseseeriaid tuleb teha, et sündmus toimuks.

• Suurte arvude seadusest järeldub, et siis, kui sündmuse tõenäosus on väga väike, juhtub see väga harva (näiteks loterii peavõidu saamine või meteoori langemine Maale).

**Järelemõtlemiseks**

1.Kas niisugune sündmus, et kogu mündiviske seeria jooksul langeb peale vapipool, on võimalik?

2.Juhusliku suuruse väärtuse määramiseks tehti 10 katset. Kui suur on suurim võimalik normeeritud hälve keskväärtuse suhtes?

VASTUSED

**Normaaljaotus ja tsentraalne piirteoreem. Lineaarne korrelatsioonikordaja**

## NORMAALJAOTUS

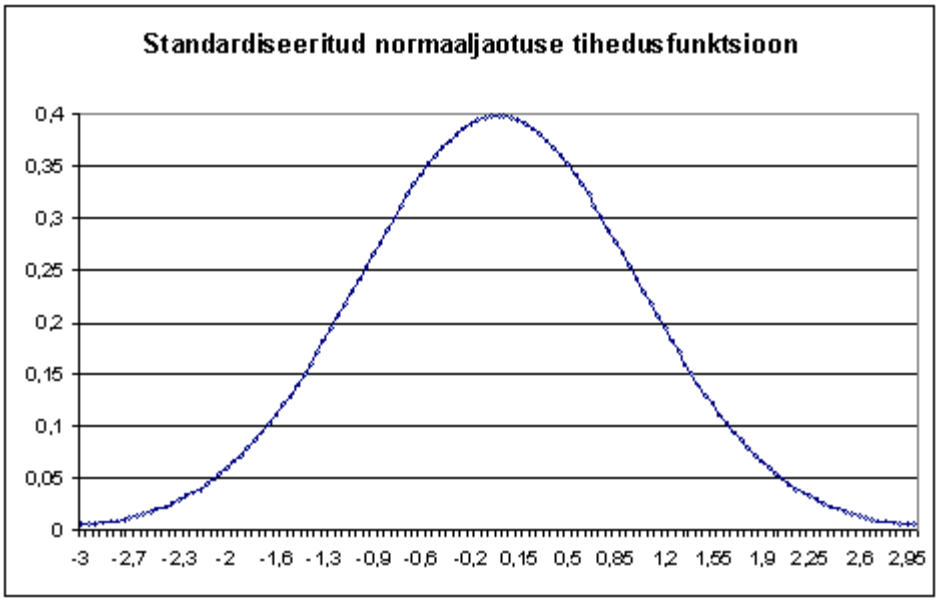
### Normaaljaotuse tihedusfunktsioon

Üks kõige sagedamini kasutatavaid jaotusseadusi on ***normaaljaotus***. Normaaljaotus sobib mudeliks nii mõõtmisvigade jaotusele, laskude hajumisele märklaua keskme ümber, kuid samuti paljude bioloogiliste objektide mõõtude muutlikkusele. Näiteks kirjeldab normaaljaotus üsna hästi nii meeste kui ka naiste pikkuste jaotust. Normaaljaotus sobib ka inimeste võimekuse testide jaotuse kirjeldamiseks – need on kõik sellised ligilähedaselt sümmeetrilised jaotused, kus suur hulk väärtusi on koondunud keskväärtuse lähemasse ümbrusesse, moodustamata eriti teravat tippu, seevastu suuremaid hälbeid esineb suhteliselt harvem. Samas ei ole hälbe suurusel piiri – kuigi üliväikese tõenäosusega võib esineda kuitahes suuri hälbeid (tuleb siiski tõdeda, et "kuitahes" on pigem abstraktsioon, praktilises elus on hälvetel enamasti sisulised piirid).

Normaaljaotusega juhuslik suurus on pidev ja seda iseloomustab alljärgneva kujuga tihedusfunktsioon

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

kus $\mu$ ja $\sigma$ on jaotuse parameetrid, argument $x$ võib omandada suvalise arvväärtus ja $\pi$ = 3,14… ning e = 2,718… on konstandid. Normaaljaotuse tähiseks on $N(\mu, \sigma)$. Normaaljaotuse tihedusfunktsiooni illustreerib lisatud graafik.




**Normaaljaotuse keskväärtus**

Normaaljaotuse keskväärtuse arvutamiseks kasutame pideva juhusliku suuruse keskväärtuse arvutamise eeskirja ja teisendust $t = x - \mu$

$$EX=\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} xe^{-\frac{(x-\mu)^2}{2\pi\sigma}}dx=\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}(t+\mu)e^{-\frac{t^2}{2\sigma^2}}dt=\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}te^{-\frac{t^2}{2\sigma^2}}dt+\mu.$$

Teisel sammul saadud avaldis laguneb kahe integraali summaks, teisest liidetavast saab konstandi $\mu$ integraalimärgi ette tuua, mille järel integraal esitab tihedusfunktsiooni integraali lõpmatutes rajades, seega võrdub ühega. Esimene liidetav on aga ilmselt paaritu funktsioon sümmeetrilistes rajades, seega on selle väärtus 0. Kokkuvõttes oleme tõestanud, et $EX = \mu$, seega on **normaaljaotuse esimene parameeter tema keskväärtus**.

**Normaaljaotuse dispersioon**

Et juhusliku suuruse dispersioon on nihke suhtes invariantne, siis ***arvutame standardiseeritud normaaljaotuse dispersiooni,*** st et võtame $\mu = 0$. Saame siis arvutusvalemi, kus kasutame teisendust $t = x/\sigma$, millest järeldub $dt = dx/\sigma$.

$$DX=\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}x^2e^{-\frac{x^2}{2\sigma^2}}dx=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\sigma^2t^2e^{-\frac{t^2}{2\sigma^2}}dt.$$

Edasisel arvutamisel rakendame ositi integreerimise valemit,

$$\int u\,dv = uv - \int v\,du$$

võttes

$$u = t,\ dv = e^{-\frac{t^2}{2}}tdt,\ du = dt,\ v = -e^{-\frac{t^2}{2}}.$$

Siis me saame:

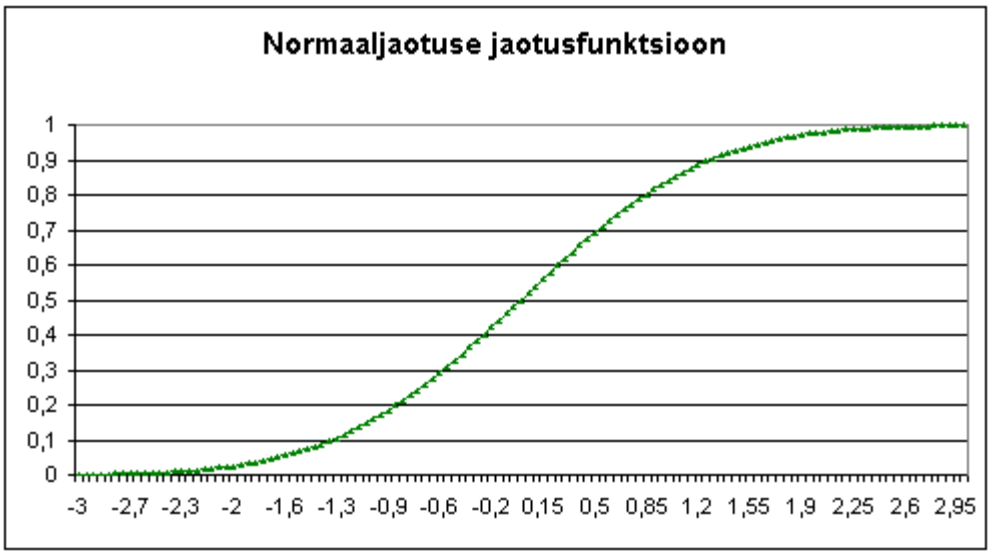
$$DX = \sigma^2 \left\{ \frac{1}{\sqrt{2\pi}} \left( te^{-\frac{t^2}{2}} \right) \Big|_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt \right\} = \sigma^2,$$

sest esimene liidetav sulgudes võrdub nulliga, nagu selgub piirväärtuste arvutamisest.

**Seega on normaaljaotusega juhusliku suuruse dispersioon *DX* = $\sigma^2$ ja standardhälve on $\sigma$.**

*Standardiseeritud normaaljaotus N*(0, 1) on vastavalt parameetritega 0 ja 1. Selle jaotuse kohta on koostatud rida tabeleid, mis varem moodustasid väga olulise vahendi praktiliste tõenäosuseülesannete lahendamisel.

**Normaaljaotuse jaotusfunktsioon**




Normaaljaotuse jaotusfunktsioon avaldub tihedusfunktsiooni kaudu:

$$F(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt.$$

Lihtsamat avaldist (elementaarfunktsioonides) selle funktsiooni jaoks ei leidu.

**Normaaljaotuse teised arvkarakteristikud**

Normaaljaotuse tihedusfunktsiooni avaldisest järeldub, et ta on oma keskpunkti suhtes sümmeetriline jaotus, seetõttu ***ühtivad mediaan ja keskväärtus***.

• Sümmeetrilisuse tõttu ***on asümmeetriakordaja võrdne nulliga***.

• Arvutamise tulemus näitab, et ***normaaljaotuse järskus on samuti võrdne nulliga***.

Seega võib kinnitada, et järskuse kordaja võrdleb jaotusi normaaljaotusega – positiivse järskusega jaotustel on normaaljaotusega võrreldes teravam tipp ja/ või raskemad sabad; negatiivse järskusega jaotused on normaaljaotusest lamedamad ja

nende sabad on kergemad/ lühemad või puuduvad hoopis.

• Normaaljaotusel ***on olemas kõik kvantiilid, need on üheselt määratud ja neid saab leida tabelist*** (puudub avaldis elementaarfunktsioonide kaudu).

• Normaaljaotusel ***puudub minimaalne ja maksimaalne väärtus***, põhimõtteliselt võib normaaljaotus omandada kuitahes suuri väärtusi.

**Normaaljaotuse lineaarfunktsiooni jaotus**

Kui $X$ on normaaljaotusega juhuslik suurus, siis on ka $Y = a + bX$ normaaljaotusega juhuslik suurus. See omadus ei kehti kaugeltki kõigi jaotuste puhul, olles pigem normaaljaotuse erandlik omadus. Juhusliku suuruse $Y$ parameetrid arvutatakse $X$ parameetrite järgi:

• $EY = a + bEX$;

• $DY = b^2\, DX$.

• Ka juhusliku suuruse $Y$ jaotusfunktsioon on $X$ jaotusfunktsiooni järgi lihtsalt arvutatav. Eeldame, et $b>0$. Siis kehtivad alljärgnevad võrratused:

$$F_Y(y) = P(Y < y) = P(a + bX < y) = P(X < (y-a)/b)) = F_X((y - a)/b).$$

Samasugune seos on tuletatav ka negatiivse kordaja $b$ korral, ning kokkuvõttes kehtib võrdus:

$$F_Y(y) = F_X\left(\frac{y - a}{|b|}\right).$$

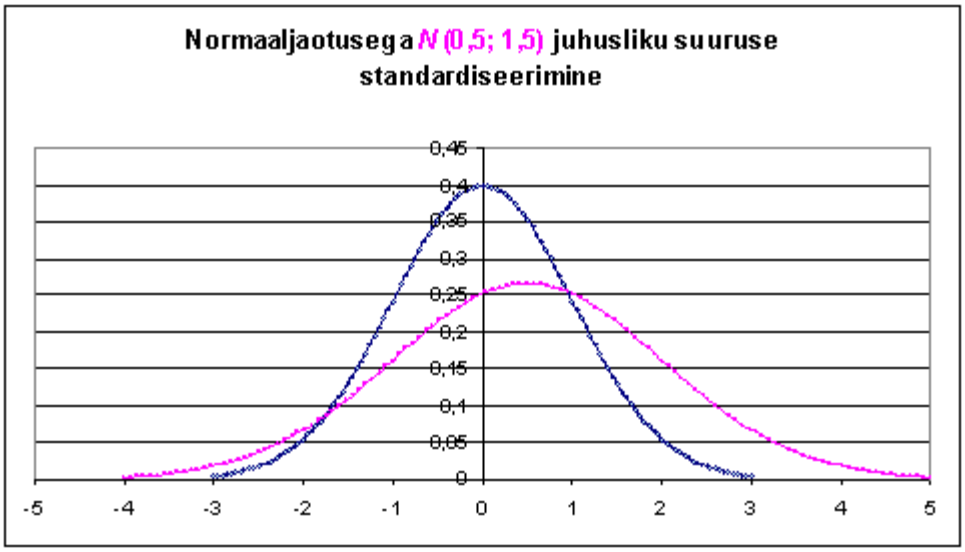
Viimasest võrratusest järeldub, et ka juhusliku suuruse $Y$ kvantiilfunktsioon on normaaljaotusega juhusliku suuruse $X$ kvantiilfunktsiooni kaudu avaldatav. Leiame $Y$ kvantiilfunktsiooni $Q_Y$ väärtuse kohal $Q_Y(p)$, kasutades vahetult definitsioonist:

$$Q_Y(p) = a + bQ_X(p).$$

Seosest jaotusfunktsioonide vahel saame diferentseerimise teel ka seose tihedusfunktsioonide vahel:

$$F_Y(y) = \frac{1}{|b|}\, f_X\left(\frac{y - a}{|b|}\right).$$

**Normaaljaotuse standardiseerimine**




Kui $X \sim N(\mu,\sigma)$, siis vastav standardiseeritud juhuslik suurus on $X_0$,

$$X_0 = \frac{X-\mu}{\sigma}.$$

$X_0$ on standardiseeritud normaaljaotusega.

| |
|---|
| Iga normaaljaotusega juhuslik suurus on saadav standardiseeritud normaaljaotusega juhuslikust suurusest lineaarteisenduse abil. |

See annab võimaluse kasutada normaaljaotuse tabeleid, mis sisaldavad standardiseeritud normaaljaotuse jaotusfunktsiooni ja kvantiilfunktsiooni väärtusi, suvalise normaaljaotuse abil defineeritud sündmuste tõenäosuste arvutamiseks.

**Normaaljaotuse tabel**

| $x$ | $F(x)$ | -1,25 | 0,106 | 0,05 | 0,520 | 1,35 | 0,911 |
|---|---|---|---|---|---|---|---|
| -2,5 | 0,006 | -1,2 | 0,115 | 0,1 | 0,540 | 1,4 | 0,919 |
| -2,45 | 0,007 | -1,15 | 0,125 | 0,15 | 0,560 | 1,45 | 0,926 |
| -2,4 | 0,008 | -1,1 | 0,136 | 0,2 | 0,579 | 1,5 | 0,933 |
| -2,35 | 0,009 | -1,05 | 0,147 | 0,25 | 0,599 | 1,55 | 0,939 |
| -2,3 | 0,011 | -1 | 0,159 | 0,3 | 0,618 | 1,6 | 0,945 |
| -2,25 | 0,012 | -0,95 | 0,171 | 0,35 | 0,637 | 1,65 | 0,951 |
| -2,2 | 0,014 | -0,9 | 0,184 | 0,4 | 0,655 | 1,7 | 0,955 |
| -2,15 | 0,016 | -0,85 | 0,198 | 0,45 | 0,674 | 1,75 | 0,960 |
| -2,1 | 0,018 | -0,8 | 0,212 | 0,5 | 0,691 | 1,8 | 0,964 |
| -2,05 | 0,020 | -0,75 | 0,227 | 0,55 | 0,709 | 1,85 | 0,968 |
| -2 | 0,023 | -0,7 | 0,242 | 0,6 | 0,726 | 1,9 | 0,971 |
| -1,95 | 0,026 | -0,65 | 0,258 | 0,65 | 0,742 | 1,95 | 0,974 |
| -1,9 | 0,029 | -0,6 | 0,274 | 0,7 | 0,758 | 2 | 0,977 |
| -1,85 | 0,032 | -0,55 | 0,291 | 0,75 | 0,773 | 2,05 | 0,980 |
| -1,8 | 0,036 | -0,5 | 0,309 | 0,8 | 0,788 | 2,1 | 0,982 |
| -1,75 | 0,040 | -0,45 | 0,326 | 0,85 | 0,802 | 2,15 | 0,984 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -1,7 | 0,045 | -0,4 | 0,345 | 0,9 | 0,816 | 2,2 | 0,986 |
| -1,65 | 0,049 | -0,35 | 0,363 | 0,95 | 0,829 | 2,25 | 0,988 |
| -1,6 | 0,055 | -0,3 | 0,382 | 1 | 0,841 | 2,3 | 0,989 |
| -1,55 | 0,061 | -0,25 | 0,401 | 1,05 | 0,853 | 2,35 | 0,991 |
| -1,5 | 0,067 | -0,2 | 0,421 | 1,1 | 0,864 | 2,4 | 0,992 |
| -1,45 | 0,074 | -0,15 | 0,440 | 1,15 | 0,875 | 2,45 | 0,993 |
| -1,4 | 0,081 | -0,1 | 0,460 | 1,2 | 0,885 | 2,5 | 0,994 |
| -1,35 | 0,089 | -0,05 | 0,480 | 1,25 | 0,894 | | |
| -1,3 | 0,097 | 0 | 0,500 | 1,3 | 0,903 | | |

Esitatud tabel sisaldab arvupaare, kus esimeses veerus on argument $x$, teises vastav jaotusfunktsiooni väärtus $F(x)$. Ruumi kokkuhoiu mõttes on neli niisugust veerupaari asetatud kõrvuti.

**Normaaljaotuse kaudu defineeritud sündmuste tõenäosuste leidmine**

Normaaljaotuse tabelit on sobiv kasutada mitmesuguste sündmuste tõenäosuste leidmiseks; selleks avaldatakse sündmuse tõenäosus jaotusfunktsiooni abil ning tõenäosus leitakse jaotusfunktsiooni $F(x)$ väärtusena.

Teiselt poolt on tabelite abil võimalik lahendada ka vastupidiseid ülesandeid, leides vastavalt sündmuse tõenäosusele juhusliku suuruse väärtusi; selleks kasutatakse tabelit vastupidi, leides $F(x)$ väärtuste järgi $x$ väärtusi.

***Näide 5.1.*** $X \sim N(0,1)$. Leida $P(X > 1,3)$. Lahendus: $P(X > 1,3) = 1 - P(X < 1,3) = 1 - 0,903 = 0,097$.

***Näide 5.2.*** $X \sim N(0,1)$. Leida niisugused väärtused $a$ ja $-a$, et $P(-a < X < a) = 0,6$. Lahendus: Leiame $-a$ nii, et $P(X < -a) = 0,2$, st $-a = -0,85$, siis $a = 0,85$. Vajadusel saab $a$ väärtust täpsustada interpoleerimise teel või täpsemaid tabeleid kasutades.

***Näide 5.3.*** $X \sim N(2,5)$. Leida $P(X > 1,3)$. Lahendus: $P(X > 1,3) = P(X_0 > (1,3-2)/5) = 1 - P(X_0 < 0,14) = 1 - 0,444 = 0,556$.

**Järelemõtlemiseks**

1.Leida normaaljaotuse tabelist (vastavast arvutiprogrammist) standardiseeritud normaaljaotuse kvartiilide väärtused.

2.Leida standardiseeritud normaaljaotuse tabelist 0,025-kvantiil ja 0,975-kvantiil. Kui suure tõenäosusega jääb juhusliku suuruse väärtus nendest piiridest väljapoole?

3.Missuguse tõenäosusega toimuvad sündmused $X > 1$ ja $X < -1$, kui $X \sim N(0, 1)$?

**Normaaljaotus ja tsentraalne piirteoreem. Lineaarne korrelatsioonikordaja**

## TSENTRAALSED PIIRTEOREEMID

### Piirteoreemi mõiste

Suurte arvude seaduse puhul nägime, et juhuslike suuruste jada võib koonduda konstandiks. Kuid juhuslike suuruste jada koondumise tulemuseks ei tarvitse tingimata olla konstant, on ka selliseid protsesse, mille korral juhuslike suuruste jada koondub juhuslikuks suuruseks. Esitame järgnevas kõige olulisemad tsentraalsed piirteoreemid. Oma nimetuse on nad saanud sellest, et käsitlevad tsentreeritud (standardiseeritud) juhuslike suuruste jadade käitumist.

### De Moivre'-Laplace'i piirteoreem

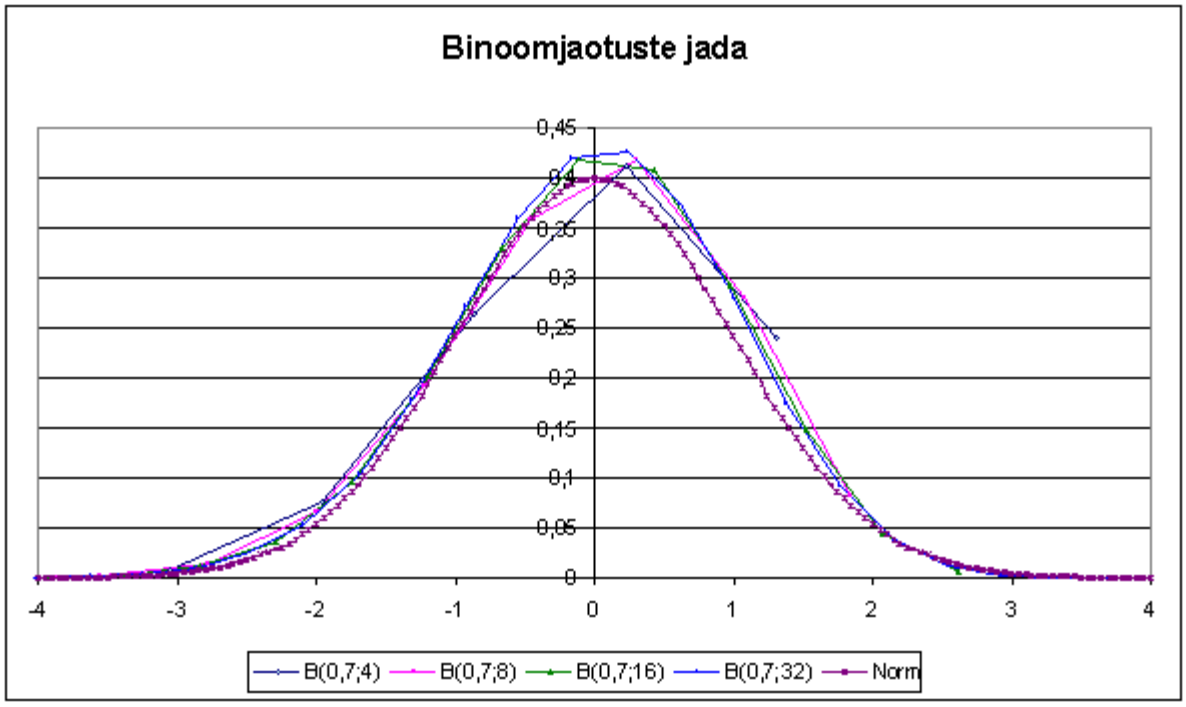
Tähistagu $X_n$ binoomjaotusega $B(n, p)$ juhuslike suuruste jada, $n \longrightarrow \infty$.

Arvutame vastava standardiseeritud juhuslike suuruste jada

$$Y_n = \frac{X_n - pn}{\sqrt{np(1-p)}},$$

Siis koondub juhuslike suuruste $Y_n$ jada juhuslikuks suuruseks $Y$, kus $Y$ on standardiseeritud normaaljaotusega.

Sellel teoreemil on mitu võimalikku tõestust. Klassikaline, de Moivre' ja Laplace'i tõestus on küll elementaarne, kuid tehniliselt tülikas ja töömahukas. Hiljem on antud sellele ja tervele reale sarnastele teoreemidele juhusliku suuruse karakteristlike funktsioonide kaudu elegantne ja lühike tõestus, ent selle jaoks vajaliku aparatuuri omandamine vajab täiendavat aega. Selletõttu jääb klassikalise piirteoreemi tõestus käesolevast kursusest välja.




**Üldine klassikaline tsentraalne piirteoreem**

Tõestatud piirteoreemi saab üldistada.

Olgu $X_n$ sõltumatute juhuslike suuruste jada, millel on sama keskväärtus $\mu$ ja sama standardhälve $\sigma$. Siis koondub jada

$$\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \to Y,$$

kus $Y$ on standardiseeritud normaaljaotusega juhuslik suurus.

**Poissoni piirteoreem**

Binoomjaotusega juhusliku suuruse piirväärtuseks pole ainult normaaljaotus, sõltuvalt jada konstruktsioonist võib see koonduda ka teisteks piirväärtusteks. Tuntud on Poissoni piirteoreem:

Olgu korduvate katsete jada määratud nii, et sündmuse $A$ tõenäosus $n$-dal katsel $p_n$ rahuldab tingimust

$$np_n \to \lambda.$$

Siis läheneb binoomjaotusega $B(p_n, n)$ juhuslike suuruste jada Poissoni jaotusega juhuslikule suurusele. See tähendab, et iga $k$ korral kehtib seos

$$P(X=k) \to e^{-k} \frac{\lambda^k}{k!}.$$

**Järelemõtlemiseks**

1.Kas De Moivre-Laplace'i teoreem on üldise klassikalise tsentraalse piirteoreemi erijuht?

2.Kuidas seostub viimase piirteoreemiga see, et normaaljaotus on levinud mudel vigade teoorias?

3.Kas binoomjaotuste normaaljaotuseks koondumise kiirus sõltub parameetrist $p$? Miks?

4.Kas Poissoni piirteoreem on ka tsentraalne? Miks mitte?

5.Millal sobib binoomjaotusega juhuslike suuruste jada lähendiks normaaljaotus, millal Poissoni jaotus?

**Normaaljaotus ja tsentraalne piirteoreem. Lineaarne korrelatsioonikordaja**

## NORMAALJAOTUSEGA JUHUSLIK VEKTOR. KORRELATSIOONIKORDAJA

### Normaaljaotusega juhusliku vektori mõiste ja omadused

Kui sama katse/ katseseeria abil on defineeritud mitu normaaljaotusega juhuslikku suurust, siis moodustavad nad normaaljaotusega juhusliku vektori. Normaaljaotusega juhuslik vektor on pidev, ning tema jaotuse määrab tihedusfunktsioon, mis sõltub sama suurest arvust argumentidest nagu on juhuslikul vektoril komponente. Kui vektori komponendid on sõltumatud, siis avaldub vektori tihedusfunktsioon komponentide tihedusfunktsioonide korrutisena.

Normaaljaotusega juhusliku vektoril on rida rakenduste seisukohast väga olulisi omadusi.

- ***Normaaljaotusega juhusliku vektori iga komponentide lineaarkombinatsioon on normaaljaotusega***, kusjuures selle jaotusparameetrid on lineaarkombinatsiooni kordajate abil vahetult arvutatavad.
- See tähendab, et normaaljaotusega juhusliku vektori iga projektsioon on normaaljaotusega.
- Samuti on ***normaaljaotusega juhusliku vektori kõik tinglikud jaotused normaaljaotusega.***

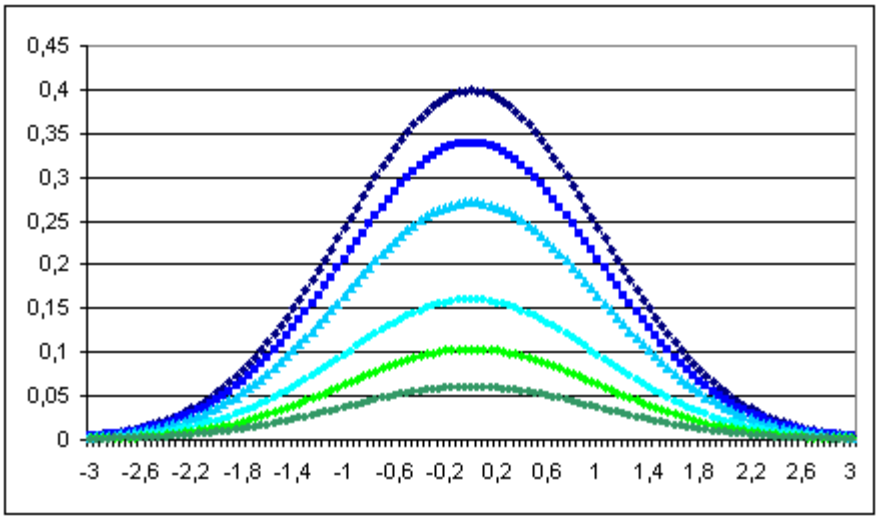
### Kahemõõtmeline normaaljaotusega juhuslik vektor

Kahemõõtmelise standardiseeritud normaaljaotusega juhusliku vektori tihedusfunktsioon on:

$$f(x,y)=\frac{1}{2\pi(1-r^2)}e^{\frac{-(x^2-2rxy+y^2)}{2\pi(1-r^2)}}$$

Siin $x$ ja $y$ on argumendid, mis näitavad kummagi komponendi väärtusi ja $r$ on komponentide $X$ ja $Y$ vaheline korrelatsioonikordaja.

Esitatud joonisel on kahemõõtmelise normaaljaotuse tihedusfunktsiooni (mis on kupli- või kellukesekujuline) projektsioonid erinevatele joonise pinnaga paralleelsetele tasanditele. Sõltumatute standardiseeritud koordinaatidega kahemõõtmeline normaaljaotus on sümmeetriline kõigi tasandite suhtes, mis on risti koordinaattelgede tasandiga ja läbivad nullpunkti.




**Korrelatsioonikordaja**

Korrelatsioonikordaja arvutatakse valemist

$$r = \frac{E(X - EX)(Y - EY)}{\sqrt{DX \cdot DY}}.$$

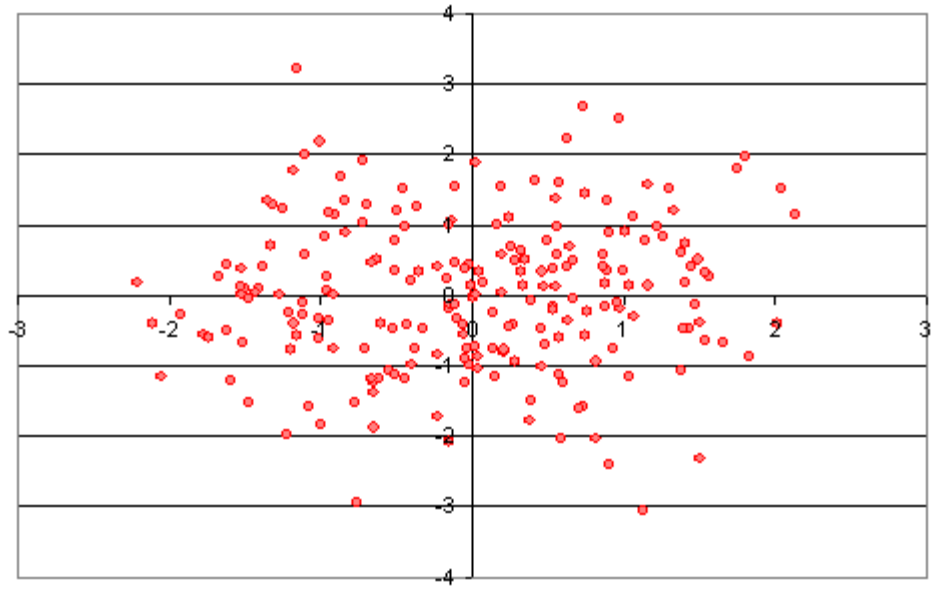
Korrelatsioonikordaja iseloomustab kahe juhusliku suuruse vahelise lineaarse seose tugevust. Kui korrelatsioonikordaja väärtus on 0, siis öeldakse, et juhuslikud suurused on mittekorreleeritud.

•Kui juhuslike suuruste vahel on täielik lineaarne sõltuvus, siis on korrelatsioonikordaja absoluutväärtus 1,

$$Y = a + bX \Leftrightarrow r(X,Y) = \operatorname{sgn}(b), |r| = 1.$$

•Korrelatsioonikordaja väärtused muutuvad piirkonnas [–1, 1].

**Kahemõõtmelise normaaljaotusega punktiparv, $r$=0,06**



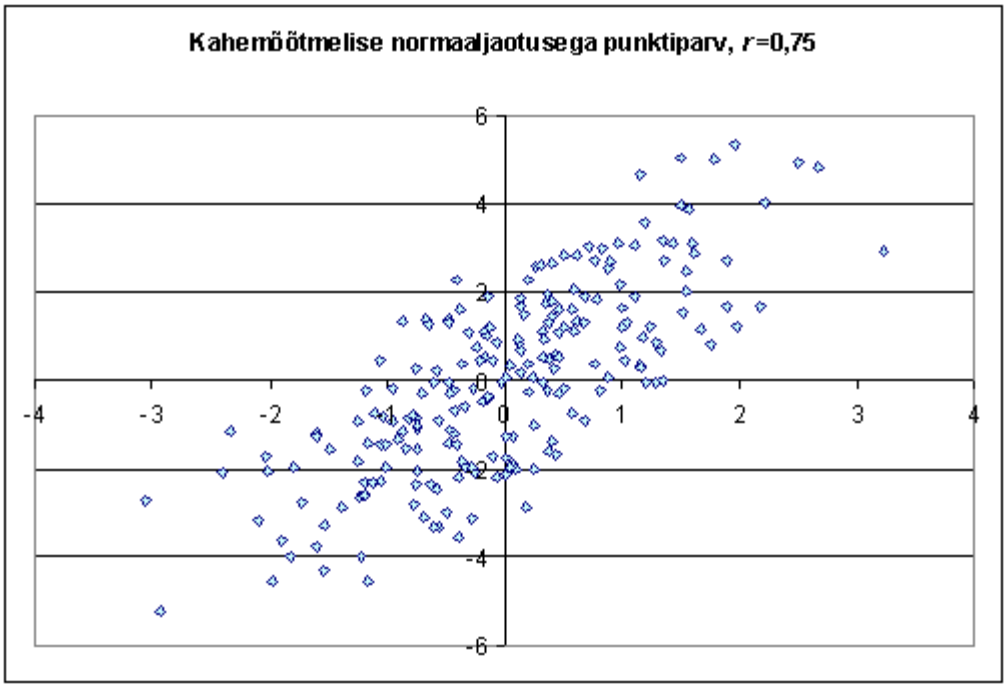
• Mida tugevam on tunnustevaheline lineaarne seos (st, mida lähedasem see on determineeritud seosele), seda suurem on korrelatsioonikordaja absoluutväärtus.

• Kui korrelatsioonikordaja väärtus on positiivne, siis ühe juhusliku suuruse suurenedes suureneb keskmiselt ka teine (ja vastupidi).

• Kui korrelatsioonikordaja väärtus on negatiivne, siis ühe juhusliku suuruse suurenedes keskmiselt teine juhuslik suurus väheneb ja vastupidi.

• Kui juhuslikud suurused on sõltumatud, siis on korrelatsioonikordaja võrdne nulliga.

• Kui kahemõõtmelise normaaljaotuse komponendid on mittekorreleeritud, siis on nad ka sõltumatud. Muude jaotuste korral see üldiselt nii ei ole.




**Järelemõtlemiseks**

1.Tõestada, et mittekorreleeritud komponentidega kahemõõtmeline normaaljaotusega vektor on ühtlasi sõltumatute komponentidega.

2.Kuidas muutub kahe juhusliku suuruse vaheline korrelatsioonikordaja siis, kui ühele neist rakendada lineaarteisendust?

3.Juhuslik suurus $Y$ on defineeritud eeskirjaga $Y = -X$. Missugune on suuruste $X$ ja $Y$ korrelatsioonikordaja?

4.Juhuslik suurus $Y$ on defineeritud eeskirjaga $Y = -X$ ja $X$ on standardiseeritud normaaljaotusega. Missugune on $Y$ jaotus?

5.Juhuslik suurus $Y$ on defineeritud eeskirjaga $Y = -X$ ja $X$ on normaaljaotusega $N(\mu, s)$. Missugune on $Y$ jaotus?

6.Juhuslikud suurused $X \sim N(\mu_1, \sigma_1)$ ja $Y \sim N(\mu_2, \sigma_2)$. on sõltumatud. Leida juhuslike suuruste $2X$, $X+Y$, $X-Y$ *ja* $2X-3Y$ jaotus.

VASTUSED

**Üldkogum ja valim. Hindamine**

## ÜLDKOGUM JA VALIM

### Matemaatilise statistika põhiülesanne

Matemaatilise statistika põhiliseks ülesandeks on järelduste tegemine statistiliste andmete põhjal. Statistilisteks andmeteks on lõplik hulk mõõtmistulemusi. Mõõdetavad suurused on juhuslikud suurused e. ***tunnused***, ning nende juhuslikkus seisneb selles, et erinevate objektide puhul omandavad nad erinevad väärtuse (näiteks inimese pikkus). Põhimõtteliselt võib mõõtmistulemuse juhuslikkust põhjustada ka mõõtmis- või vaatlusviga - matemaatilise statistika seisukohast ei ole juhuslikkuse aluseks olev põhjuslik mehhanism olulise tähtsusega. Matemaatilise statistika eesmärgiks on teha järeldusi mõõdetud juhuslike suuruste ***jaotuste*** (sh jaotusparameetrite omavaheliste seoste jne) kohta. Matemaatiline statistika tugineb sealjuures tõenäosusteooriale kui meetodile, tehtavad järeldused on oma olemuselt tõenäosuslikud.

### Üldkogum ja valim

Matemaatilises statistikas kasutatava andmestiku kohta tehakse tavaliselt rida eeldusi. Vaatleme neid lihtsaimal erijuhul, kui andmestik on ühemõõtmeline, st et mõõdetud on ainult üht tunnust.

•Vaadeldav nähtus (protsess) toimub mingis lõpmata suures kogumis, mida nimetatakse ***üldkogumiks***. Uurijat huvitab teha järeldusi üldkogumi kohta.

•Uuritava tunnuse jaotust üldkogumis (nn ***teoreetilist jaotust***) iseloomustab teadaolev jaotusseadus (väga sageli normaaljaotus), kuid selle jaotuse parameetrid ei ole teada.

•Uurija käsutuses olev statistiline andmestik on ***valim*** sellest üldkogumist.

•***Valimi maht*** on lõplik, selle tähiseks on $n$.

•Valimisse kuuluvad punktid/ objektid on üldkogumist juhuslikult valitud nii, et kõigil üldkogumi punktidel on võrdne tõenäosus valimisse sattuda ehk punkti valimisse sattumise tõenäosus on võrdeline üldkogumi tihedusfunktsiooniga selles punktis.

•Valimisse valitud punktid on omavahel sõltumatud.

### Teoreetiline ja konkreetne valim

Põhimõtteliselt käsitletakse valimit kahes mõttes:

•Konkreetne valim koosneb mõõtmistulemustest, st arvudest, kusjuures need arvud on antud teoreetilise jaotusega juhusliku suuruse väärtused.

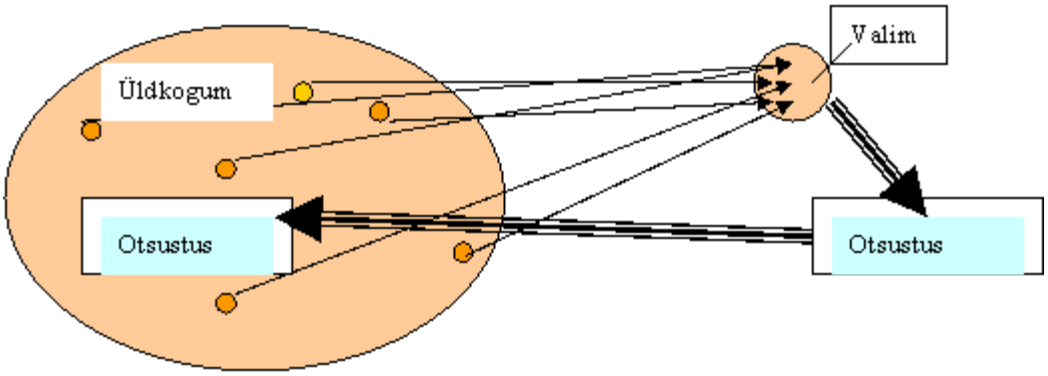
•Teoreetiline valim on juhuslik sõltumatute komponentidega $n$-mõõtmeline vektor, mille kõigi komponentide jaotuseks on üldkogumi jaotus.

Konkreetset jaotust kasutatakse ***statistiliste otsustuste tegemiseks üldkogumi***

***kohta***, teoreetilist jaotust – nende ***otustuseeskirjade omaduste selgitamiseks***.

**Matemaatilise statistika põhiülesande täpsustus**

Täpsemalt on matemaatilise statistika ülesandeks valimi põhjal otsustuste tegemine üldkogumi kohta.




**Valimi jaotus**

Konkreetse valimi jaotuse eeskirjaks on: $P(x_j) = 1/n$, st et ***kõigil valimi punktidel on võrdne tõenäosus***. Valimi jaotus on empiiriline jaotus. Selle jaotuse põhjal saab arvutada kõik jaotuse parameetrid, neid nimetatakse valim- ehk empiirilisteks parameetriteks.

**Järelemõtlemiseks**

1.Olgu üldkogum Bernoulli jaotusega. Missuguse jaotusega on siis valim?

2.Kas valim saab olla normaaljaotusega?

3.Oletame, et samast üldkogumist on võetud kaks sama mahuga valimit? Kas need on võrdsed või erinevad?

4.Kuidas muutub valimi jaotus, kui valimisse üks vaatlus lisatakse?

VASTUSED

**Üldkogum ja valim. Hindamine**

**PUNKTIHINNANG**

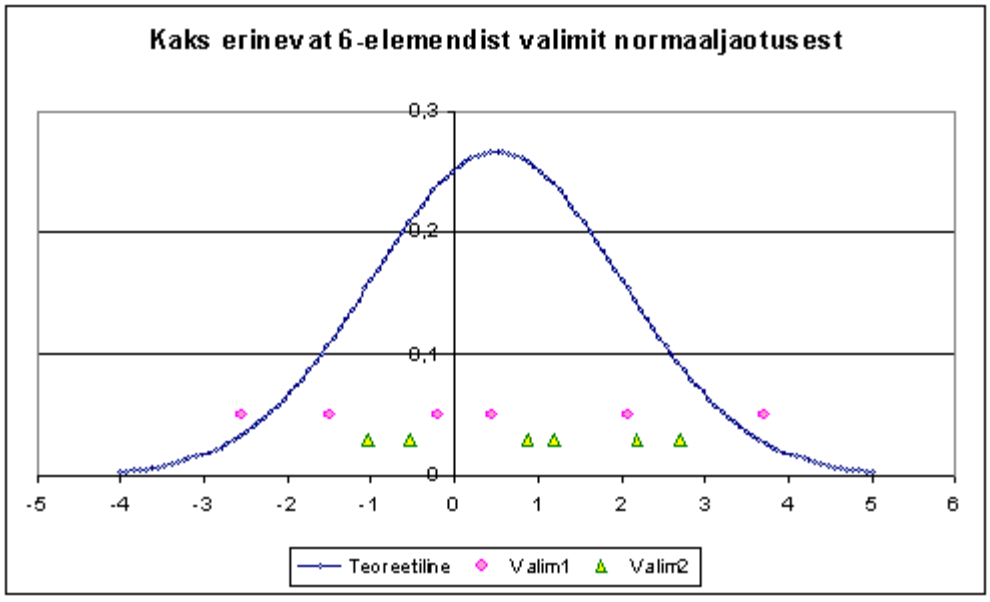
**Jaotusparameetrite hindamise ülesanne**

Tihti on tarvis valimi põhjal leida ***jaotusparameetrite hinnanguid***. Kõige sagedamini on tarvis hinnata ***keskväärtust***. Näiteks küsitakse vastsündinu keskmist kaalu vaadeldaval aastal, töötajate keskmist palka teataval kuul, ooteaja keskmist kestus pangas enne teenindamist, keskmist sademete arvu mingil kuul jne. Niisuguse ülesande lahendamiseks on kaks võimalust **– *kõikne*** ja ***valimipõhine***.

• ***Kõikse andmestiku põhjal on võimalik keskväärtus täpselt välja arvutada***. Näiteks kui aasta jooksul on kõik vastsündinud kaalutud, siis on täpselt (mõõtmisvigade täpsuseni) teada nende kaalu jaotus ning on võimalik arvutada välja ka täpne keskmine kaal, mis erineb selle "õigest" väärtusest vaid võimaliku mõõtmisvea poolest. Ekslik on kasutada kõikse valimi põhjal tehtud arvutuste korral metoodikat, mis eeldab, et andmed on valimipõhised.

• Valimi põhjal leitakse soovitavale ***keskväärtusele hinnang***, mis põhimõtteliselt peale mõõtmisvea sisaldab veel ka valimi juhuslikkusest tulenevat ***juhuslikku viga***. Selle põhjuseks on see, et ***kaks samast üldkogumist võetud samamahulist valimit ei ole ühti peaaegu mitte kunagi***, vt alljärgnev joonis.

Matemaatiline statistika tegeleb ***hinnangutega***, seega edaspidi ei vaatle me kõikse statistika abil lahendatavaid olukordi.




**Punkt- ja vahemikhinnang**

Matemaatilises statistikas lähtutakse eeldusest, et üldkogumi jaotusel on olemas ***õiged*** parameetri väärtused, mida uurija ei tea ja püüab hinnata. Selleks kasutatakse põhimõtteliselt kaht tüüpi hinnanguid. ***Punkthinnang on valimi põhjal arvutatud (üks) arv***, mis peaks eelduste kohaselt olema võimalikult lähedane hinnatava parameetri õigele väärtusele. ***Vahemikhinnang on otspunktide abil määratud vahemik***, milles (küllalt suure tõenäosusega) sisaldub hinnatava parameetri õige väärtus. Punkthinnang sisaldab vähem teavet, sest pole teada, kui palju parameetri õige erineb hinnangust.

**Punkthinnangu arvutamine valimi põhjal**

Sageli kasutatakse üldkogumi parameetri punkthinnangu arvutamiseks vastavat valimparameetrit. Sellisel hinnangul on rida häid omadusi:

• Ta on lihtsalt arvutatav, sest valimjaotus on diskreetne lõpliku hulga väärtustega ning seetõttu on enamus jaotusparameetritest vahetult arvutatavad;

• Enamasti on valimparameetrite jaoks küllalt lihtne leida täpsuse hinnanguid ja kontrollida nende omadusi.

***Näide 1.*** Joonisel on kujutatud kaks kuuepunktilist valimit üldkogumist normaaljaotusega keskväärtusega 0,5 ja dispersiooniga 2,25. Valimid on moodustatud põhimõtteliselt sarnaselt, kuid juba esmapilgul on näha, et nad on erinevad. Leiame nende valimite keskmised, kasutades valemit:

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

| | 1.valim | 2.valim |
|---|---|---|
| 1 | -1,478 | 2,701 |
| 2 | 2,08 | -0,534 |
| 3 | 0,444 | 0,878 |
| 4 | 3,704 | 2,198 |
| 5 | -2,561 | -1,039 |
| 6 | -0,203 | 1,187 |
| Summa | 1,985 | 5,39 |
| $\bar{x}$ | **0,331** | **0,898** |

Selgub, et ühe valimi põhjal arvutatud hinnang on õigest keskmisest (mis käesoleval juhul on 0,5) väiksem, teine aga suurem. Kahjuks pole punkthinnangut üksi kasutades selge, kui suure täpsusega ta on, selletõttu ongi tarvis täiendavalt kasutada hinnangu täpsust iseloomustavaid näitajaid.

**Hinnangu keskväärtus. Nihe**

Et valim on juhuslik, siis on ka valimi põhjal arvutatud suurused, sh ka hinnang, juhuslikud suurused. Kasutades teoreetilise valimi mõistet, on võimalik ***leida hinnangu keskväärtus***.

Kuna teoreetilise valimi puhul on iga valimi punkt üldkogumi jaotusega, siis saame keskväärtuse omadusi kasutades leida valimkeskmise keskväärtuse:

$$E\bar{x} = \frac{1}{n}\sum_{i=1}^{n} Ex_i = \frac{1}{n}\, nEX = EX.$$

Näeme, et keskmiselt annab valimkeskmine õigele keskmisele õige hinnangu, st ta ei üle- ega alahinda õiget keskmist süstemaatiliselt. Siinjuures on oluline see, et see omadus kehtib iga üldkogumi jaotuse korral. Edaspidi tähistame hinnangut tähega *t*, *t* on valimi funktsioon ja seega juhuslik suurus. Hinnatava parameetri õige väärtuse tähiseks on statistikakirjanduses kasutusel kreeka täht $\vartheta$(teeta).

**Kui hinnangu *t* keskväärtus E*t* ühtib hinnatava parameetri õige väärtusega $\vartheta$, siis öeldakse, et hinnang on *nihketa*.**

• Hinnangu keskväärtuse *Et* ja hinnatava parameetri õige väärtuse vahet nimetatakse hinnangu nihkeks. Nihke tähiseks on $b(t)$.
• Positiivse nihke puhul hinnang ülehindab, negatiivse nihke puhul – alahindab hinnatavat parameetrit $\vartheta$ süstemaatiliselt. Sellest ei järeldu siiski, et igal konkreetsel juhul oleks hinnangu viga sama märgiga.

**Dispersiooni hinnang. Nihke kõrvaldamine**

Teine oluline parameeter, mida on tarvis arvutada, on üldkogumi dispersioon *DX*. Üsna loomulik oleks selle hinnanguna kasutada valimdispersiooni, mille avaldis on

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

Arvutamisel selgub aga, et see hinnang on nihkega,

$$E\bar{\sigma}^2 = \frac{n-1}{n} DX,$$

seega alahindab õiget dispersiooni. Selletõttu kasutataksegi enamasti dispersiooni hindamiseks nihketa hinnangut

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

**Asümptootiline nihketus**

Kui hinnangu nihe läheneb valimi mahu kasvades nullile, siis on hinnang asümptootiliselt nihketa. Suurte valimite puhul pole suurt vahet, kas kasutada nihketa või asümptootiliselt nihketa hinnanguid.
Dispersiooni hinnang $\bar{\sigma}^2$on asümptootiliselt nihketa, sest tema nihe läheneb nullile:

$$b(\bar{\sigma}^2) = -\frac{1}{n} DX \to 0, \text{ kui } n \to \infty.$$

***Näide 2. Arvutame dispersiooni ja standardhälbe hinnangud eelmises näites esitatud andmete põhjal.***

| Hälvete ruutude summa | | |
|---|---|---|
| | 1.valim | 2.valim |
| 1 | 3,272 | 3,248 |

| | | |
|---|---|---|
| 2 | 3,058 | 2,052 |
| 3 | 0,013 | 0 |
| 4 | 11,37 | 1,689 |
| 5 | 8,364 | 3,753 |
| 6 | 0,285 | 0,083 |
| Summa | 26,37 | 10,83 |
| $s^2$ | 5,273 | 2,165 |
| Standardhälve | 2,296 | 1,471 |
| $\bar{\sigma}^2$ | 4,394 | 1,804 |
| Standardhälve | 2,096 | 1,343 |

Selgub, et erinevate valimite põhjal arvutatud dispersioonihinnangud on üsna erinevad, ning erinevad ka dispersiooni õigest väärtusest, mis käesoleval juhul on 2,25. Esimese valimi põhjal arvutatud nihketa hinnang on üle kahe korra suurem õigest dispersioonist, teise valimi põhjal arvutatud hinnang aga on sellest veidi väiksem. Samasugune on vahekord standardhälbe hinnangute korral.
Tabelis on esitatud ka nihkega hinnangud. Näeme, et käesoleval juhul esimese valimi põhjal arvutatud nihkega hinnang on õigele dispersioonile lähemal, kuid see on juhus, mida pole reaalselt võimalik kasutada. Teise valimi põhjal arvutatud nihkega hinnang on aga dispersiooni õigest väärtusest märksa väiksem.

**Hinnangu hajuvus ja dispersioon**

Hinnangu ebatäpsust põhjustab enamasti hinnangu hajuvus. Seda mõõdab hinnangu dispersioon. On selge, et mida väiksem on hinnangu dispersioon, seda täpsem on üldiselt hinnang. Hinnangu dispersiooni arvutamisel kasutatakse jälle teoreetilise valimi omadusi. Arvutame näitena valimkeskmise dispersiooni:

$$D(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} DX_i = \frac{1}{n}DX.$$

Jõudsime väga olulise tulemuseni – ***valimkeskmise dispersioon on pöördvõrdeline*** valimi mahuga, st et valimi mahtu vajaliku määrani suurendades on võimalik teha keskväärtuse hinnangu dispersioon kuitahes väikseks.

**Standardviga ja suhteline viga**

Keskväärtuse hinnangu standardhälvet nimetatakse ka **standardveaks**, selle suuruse tähiseks on m.. Standardviga on praktikas üks kõige levinumaid hinnangu täpsuse mõõte. Eelöeldust järeldub, et hinnangu ***standardviga on pöördvõrdeline ruutjuurega*** valimi mahust.
Teine hinnangu täpsust iseloomustav suurus on ka ***suhteline viga***, mis võrdub standardvea ja hinnatava suuruse keskväärtuse suhtega $\frac{m}{\bar{x}}$ ja mis avaldatakse tihti protsentides.

**Hinnangu efektiivsus**

***Nihketa hinnang on efektiivne, kui tema dispersioon on minimaalne kõigi sama mahuga nihketa hinnangute seas.***

Hinnangu efektiivsus e($\vartheta$)võrdub sama mahuga efektiivse hinnangu ja vaadeldava hinnangu dispersioonide sutega

$$e(\vartheta)=\frac{D\vartheta}{D\vartheta^*},$$

kus $\vartheta^*$ tähistab efektiivset hinnangut. Mida väiksem on hinnangu efektiivsus, seda suurem peab olema valimi maht soovitava hinnangutäpsuse saavutamiseks.

**Hinnangu mõjusus**

Hinnang on ***mõjus*** siis, kui ta koondub tõenäosuse järgi hinnatava parameetri õigeks väärtuseks,

$$\vartheta_n \xrightarrow{p} \vartheta.$$

Mõjusad on kõik nihketa või asümptootiliselt nihketa hinnangud, mille dispersioon valimi mahu suurenedes läheneb nullile.

***Näide 3.***

Leiame esitatud näiteandmestiku põhjal ka keskmise standardvea ja suhtelise vea.

Esimese valimi põhjal saame: $m^2 = 5{,}273/6 = 0{,}879$, $m = 0{,}937$, suhteline viga on aga 2,833 ehk 283%.

Teise valimi põhjal saame: $m^2 = 2{,}165/6 = 0{,}361$, $m = 0{,}601$; suhteline viga on 0,669 ehk 67%.

**Tõenäosuse hinnang ja hinnangu viga**

Sündmuse tõenäosuse hindamiseks valimi põhjal kasutatakse selle sündmuse suhtelist sagedust – see on mõttekäik, mis on tuttav juba statistilise tõenäosuse definitsioonist. Seega määrab ***statistiline tõenäosus tõenäosuse hinnangu***. Et suhtelist sagedust võib käsitleda Bernoulli jaotusega juhusliku suuruse keskväärtuse hinnanguna, siis saab selle dispersiooni määramisel kasutada keskväärtuse hinnangu dispersiooni arvutamise valemit:

$$D(k/n)=\frac{1}{n-1}\left(\hat{p}(1-\hat{p})\right)=\frac{k(n-k)}{(n-1)n^2}.$$

***Näide 4***

Olgu 10 katse jooksul tulemus *A* esinenud 7 korda. Siis on sündmuse *A* suhteline sagedus 0,7 ja selle sageduse hinnangu dispersioon on 21/900 = 0,023. Hinnangu standardviga on 0,153 ja suhteline viga 0,218 ehk 21,8%.

**Hindamismeetodid**

Hinnangute leidmiseks kasutatakse mitmeid erinevaid meetodeid, millest tuntuimad on vähimruutude meetod, suurima tõepära meetod ja momentide meetod.

***Momentide meetod***

• Momentide meetodi idee seisneb selles, et jaotusparameetrid avaldatavad juhusliku suuruse momentide kaudu:

$$\mu_k=E(X^k)$$

kaudu avaldatavad.

• Alati on leitavad ka valimimomendid

$$m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

• Asendades jaotusparameetrite avaldistesse teoreetiliste momentide asemele valimimomendid saamegi parameetritele hinnangud.

Momentide meetodil saadud hinnangud ei ole üldjuhul nihketa. Näiteks momentide meetodi kohta on hinnangud $\overline{x}$ ja $\overline{\sigma}^2$.

***Vähimuutude meetod***

• Vähimuutude meetodi idee seisneb selles, et otsitakse niisugust parameetri väärtust, mille korral valimi ja teoreetilise väärtuse vaheliste hälvete ruutude summa omandaks minimaalse väärtuse.

• Enamasti tuleb parameetrite hindamiseks vähimruutude meetodil lahendada ekstreemumülesanne.

***Suurima tõepära meetod***

• Suurima tõepära meetodi korral otsitakse niisuguseid parameetri väärtusi, mille korral olemasoleva valimi saamise tõenäosus oleks maksimaalne.

• Suurima tõepära meetodi puhul on oluline see, missugune eeldatakse olevat uuritava tunnuse jaotus.

• Normaaljaotuse korral ühtib suurima tõepära ja vähimruutude meetod, ülejäänud jaotuste korral võib suurima tõepära hinnang olla täpsem, kuid selle leidmine on tavaliselt töömahukam.

***<u>Näide 5 (vähimruutude meetod)</u>***

Selleks, et leida valimi $x_1$, $x_2$, ..., $x_n$ põhjal keskväärtuse hinnangut, otsime arvu $\mu$ nii, et minimeerida summa

$$\sum_{i=1}^{n} (x_i-\mu)^2.$$

Selleks arvutame saadud avaldisest tuletise $\mu$ järgi ja lahendame võrrandi:

$$\frac{d}{d\mu}\sum_{i=1}^{n} (x_i-\mu)^2 = 2\sum_{i=1}^{n} (x_i-\mu) = 0 \Rightarrow = \frac{d}{d\mu}\sum_{i=1}^{n} x_i.$$

Selgus, et niihästi momentide meetodi kui ka vähimruutude meetodi puhul on keskväärtuse hinnanguks valimkeskmine.

**Järelemõtlemiseks**

1.Kas on alust arvata, et mõni valim on parem ja mõni halvem?

2.Missugused suurused alljärgnevast loetelust on juhuslikud – hinnang, hinnangu nihe, valimkeskmine, valimdispersioon, efektiivsus?

3.Mis on suurem, kas juhusliku suuruse standardhälve või standardviga?

4.Missugune näitaja iseloomustab hinnangu süstemaatilist viga?

5.Osutus, et $n$ vaatluse põhjal leitud hinnangu suhteline viga oli kaks korda suurem kui soovitav. Kui palju tuleks teha täiendavaid katseid, et saada soovitava täpsusega hinnang?

6.Kas mõjusa/ nihketa/ nihkega/ mittemõjusa/ asümptootiliselt nihketa/ efektiivse

hinnangu täpsust saab suurendada katsete arvu suurendades?

**Üldkogum ja valim. Hindamine**

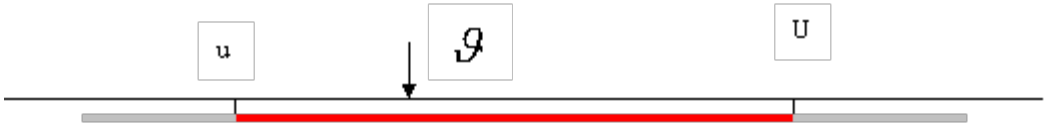
**VAHEMIKHINNANG**

**Usalduspiirid ja usaldustõenäosus**

Igati loomulik on mõte, et kui hinnangu juhuslikkuse tõttu ei saa täpselt määrata punkthinnangut, oleks mõistlik määrata vahemik, mis kindlasti otsitavat hinnangut kataks. Probleem on aga selles, et suurel osal juhtudest ei õnnestu ka sellist vahemikku ega piirkonda määrata, mis täiesti kindlasti otsitavat parameetrit sisaldaks. Kui aga niisugune hindamine peakski õnnestuma, ei paku see enamasti huvi, sest saadud piirkond on liiga lai ja tulemus tihti triviaalne. Praktiliselt pakub aga huvi ***usalduspiiride*** määramine, mis toimub järgmise skeemi kohaselt.

• Määratakse usaldusnivoo, st otsustatakse, kui suure tõenäosusega peaks vaadeldav piirkond õiget parameetrit sisaldama. Standardseks usaldusnivoo väärtuseks on 0,95, ning seda tähistatakse tavaliselt sümboliga $1-\alpha$.

• Valimi põhjal leitakse hinnatava parameetri väärtuste hulgas niisugune vahemik/ piirkond, mis kataks õiget parameetri väärtust tõenäosusega $1-\alpha$.




$$P(u < \vartheta < U) \geq 1-\alpha$$

Joonisel tähistab punane lõik parameetri $\vartheta$ $(1-\alpha)$ -usalduspiirkonda., selle otspunktid *u* ja *U* on usalduspiirid. Hallid lõigud tähistavad parameetri võimalike väärtuste hulka väljaspool usalduspiirkonda.

• Usalduspiirkonna otspunkte nimetatakse usalduspiirideks, neist väiksem on ***alumine***, suurem ***ülemine usalduspiir***.

• Põhimõtteliselt leidub palju niisuguseid lõike, mis rahuldavad usalduspiirkonna tingimusi. Kõige sobivam on valida nende hulgast välja ***minimaalse pikkusega usalduspiirkond***.

• Mida väiksema ulatusega (kitsam/ lühem) on usalduspiirkond, seda täpsem on hinnang.

• Kõige sagedamini otsitakse ***sümmeetrilist usalduspiirkonda***, mille puhul parameeter on võrdse tõenäosusega ülemisest usalduspiirist suurem ja alumisest usalduspiirist väiksem. See tõenäosus on tavaliselt $\alpha/2$.

**Usalduspiiride konstrueerimine**

Usalduspiirid *u* ja *U* on juhuslikud, valimist sõltuvad suurused. Nende määramise

juures on sageli abiks punkthinnang. Kui on teada:

• parameetri $\vartheta$ punkthinnang $t$ ja

• selle punkthinnangu jaotus, siis on võimalik määrata usalduspiirid nii, et kehtivad võrratused:

$$P(\vartheta < u) \le \alpha/2, \quad P(\vartheta > U) \le \alpha/2.$$

Paneme tähele, et siin on vastavalt $u$ ja $U$ juhuslikud suurused, $\vartheta$ aga konstant (mille väärtust me küll ei tea).

Kõige sagedamini arutletakse nii: Kui $t$ jaotuse kuju on teada, siis on sageli võimalik määrata ka selle jaotuse $\alpha/2$-kvantiil ning $(1-\alpha/2)$-kvantiil, mille korral vastavalt kehtivad võrratused

$$P(t < u) \le \alpha/2, \quad P(t > U) \le \alpha/2.$$

Kvantiilide hinnanguid $u$ ja $U$ sobib kasutada parameetri $\vartheta$ alumise ja ülemise $(1-\alpha)$-usalduspiirina.

**Normaaljaotuse keskväärtuse usalduspiiride määramine normaaljaotuse abil**

Vaatleme näitena üht kõige sagedamini esinevat ülesannet – normaaljaotusega juhusliku suuruse keskväärtuse $EX$ usalduspiiride arvutamist. Leiame kõigepealt punkthinnangu

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

Et see on normaaljaotusega juhuslike suuruste lineaarkombinatsiooniga, siis on ka $\bar{x}$ normaaljaotusega, ning sõltumatute liidetavate dispersiooni omaduse tõttu on

$$D(\bar{x}) = \frac{1}{n} DX.$$

Kui $DX$ oleks teada, oleks normaaljaotuse kvantiilide järgi väga lihtne leida ka vajalikke usalduspiire.

• 95% usalduspiiride jaoks leitakse normaaljaotuse tabelist 0,025- ja 0,975-kvantiilid, need on vastavalt –1,96 ja 1,96, ning usalduspiirid arvutatakse:

$$u = \bar{x} - 1{,}96 \frac{\sigma}{\sqrt{n}}, \quad U = \bar{x} + 1{,}96 \frac{\sigma}{\sqrt{n}}.$$

• 99% usalduspiiride jaoks leitakse 0,005- ja 0,995-kvantiilid ning usalduspiirid on:

$$u = \bar{x} - 2{,}58 \frac{\sigma}{\sqrt{n}}, \quad U = \bar{x} + 2{,}58 \frac{\sigma}{\sqrt{n}}.$$

Niisugust metoodikat saab kasutada siis, kui normaaljaotuse dispersioon on teada (mida juhtub väga harva), aga ka siis, kui ta on hinnatud väga suure valimi põhjal.

**Põhistatistikute jaotuste defineerimine**

Selleks, et leida üldisemal juhul normaaljaotuse keskväärtuse usalduspiire, kuid ühtlasi saada vahend rea statistikaülesannete lahendamiseks, on tarvis defineerida kaks jaotust: t-jaotus ja $\chi^2$-jaotus, mida nimetatakse ka põhistatistikute jaotuseks.

***$\chi^2$-jaotus***

Juhuslik suurus $Y_k$ on -jaotusega vabadusastmete arvuga $k$ ehk $\chi^2(k)$-jaotusega, kui ta on defineeritud alljärgnevalt:

$$Y_k = \sum_{i=1}^{k} X_i^2, \quad X_i \sim N(0,1), \quad i = 1,\ldots,k; \quad sõltumatud.$$

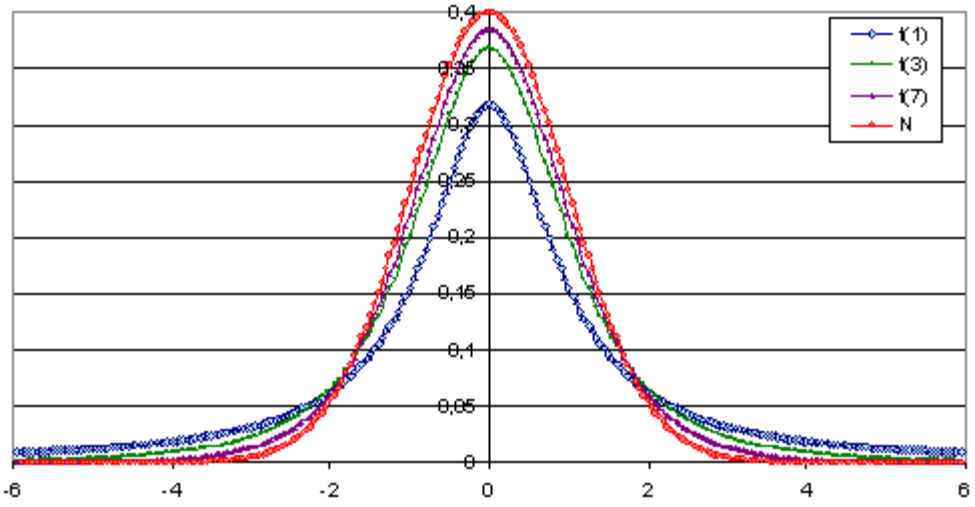
Lisatud jooniselt on näha, et vabadusastmete arvu $k$ suurenedes nihkub $\chi^2(k)$-jaotuse tihedusfunktsiooni graafik paremale (asendikarakteristik suureneb) ja muutub lamedamaks (hajuvuse parameeter suureneb samuti). Definitsioonist järeldub, et selle jaotuse keskväärtus on $k$ ja dispersioon $2k$.

***t-jaotus***

Juhuslik suurus $Z_k$ on $t$-jaotusega vabadusastmete arvuga $k$, ehk $t(k)$-jaotusega, kui ta on defineeritud alljärgnevalt:

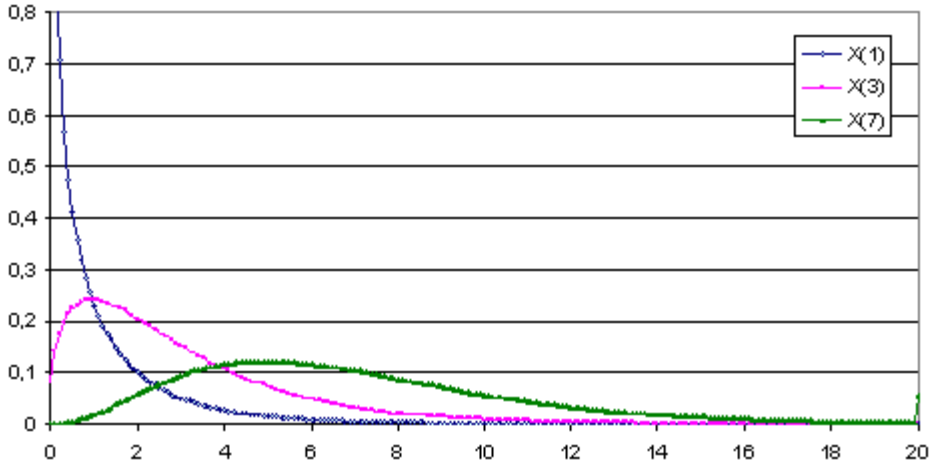
$$Z_k = \frac{X}{\sqrt{\frac{Y_k}{k}}}, \quad X \sim N(0,1); \quad Y_k \sim \chi^2(k); \quad sõltumatud.$$

**Standardiseeritud normaaljaotuse ja t-jaotuse tihedusfunktsioon**




Lisatud jooniselt on näha, et $t$-jaotuse tihedusfunktsiooni graafik muutub vabadusastmete arvu suurenedes suhteliselt vähe – vabadusastmete arvu suurenedes muutuvad jaotuse sabad kergemaks ja tihedusfunktsioon kontsentreerub keskväärtuse lähedusse. Samas ei muutu jaotuse asetus ja ta on kõigi vabadusastmete arvude korral sümmeetriline.

**X-ruut-jaotuse tihedusfunktsioon**




***t*-statistiku jaotus**

Kui $X \sim N(\mu, \sigma)$, siis ilmselt kehtib seos

$$\sum_{i=1}^{k} \frac{(x_i - \mu)^2}{\sigma^2} \sim \chi^2(k),$$

ning samuti on võimalik tõestada, et

$$\sum_{i=1}^{k} \frac{(x_i - \bar{x})^2}{\sigma^2} \sim \chi^2(k-1).$$

Võrreldes saadud avaldist dispersiooni nihketa hinnangu avaldisega $s^2$ on lihtne näha, et

$$(n-1)\frac{s^2}{\sigma^2} \sim \chi^2(n-1).$$

Defineerime nüüd *t*-statistiku, kus $\mu$, $\sigma$ on teoreetilise jaotuse parameetrid ja $\bar{x}$, $s$ – nende hinnangud.

$$t = \frac{\bar{x} - \mu}{s}\sqrt{n}.$$

Selgub, et *t*-statistik on *t*-jaotusega vabadusastmete arvuga $n-1$. Selle tõestamiseks tuleb vaid statistiku avaldist pisut teisendada:

$$t = \frac{\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{s^2(n-1)}{\sigma^2(n-1)}}},$$

ning me oleme saanudki kahe juhusliku suuruse jagatise, kus lugejas on standardiseeritud normaaljaotusega juhuslik suurus, nimetajas aga ruutjuur $\chi^2(n-1)$-jaotusega juhuslikust suurusest, mis on jagatud vabadusastmete arvuga. Seda, et lugeja ja nimetaja on sõltumatud, on samuti võimalik tõestada.

**Normaaljaotuse keskväärtuse usalduspiiride määramine *t*-statistiku abil**

Tehtud teisenduste mõte on selles, et saime nüüd statistiku, mis sisaldab teoreetilise dispersiooni asemel tema hinnangut. Seda kasutades saame normaaljaotuse usalduspiiride jaoks avaldise:

$$\begin{cases} u = x - \frac{s}{\sqrt{n}}\, t_{n-1}(\alpha/2); \\ u = x - \frac{s}{\sqrt{n}}\, t_{n-1}(\alpha/2). \end{cases}$$

Siin $t_{n-1}(\alpha/2)$ tähistab $t$-jaotuse $(1-\alpha/2)$-kvantiili ehk $\alpha/2$-täiendkvantiili, mille korral

$$P(X > t(\alpha/2)) = \alpha/2,$$

kusjuures $t$-jaotuse vabadusastmete arv on $n-1$. Paneme tähele, et mida rohkem on tehtud vaatlusi, seda kitsamaks muutuvad usalduspiirid. Kui vaatluste arv on üle saja, siis on $t$-jaotus juba nii lähedane normaaljaotusele, et $t$-jaotuse asemel võib usalduspiiride arvutamisel kasutada normaaljaotust.

***Näide 6.***

Leiame usalduspiirid näidetes 1 ja 2 vaadeldud valimite põhjal, vt alljärgnevat tabelit.

| Arvutussamm | 1.valim | 2.valim |
|---|---|---|
| keskmise hinnang | 0,331 | 0,898 |
| standardhälbe hinnang | 2,296 | 1,471 |
| valimi maht | 6,000 | 6,000 |
| ruutjuur | 2,449 | 2,449 |
| standardviga | 0,937 | 0,601 |
| vabadusastmeid | 5,000 | 5,000 |
| usaldusnivoo | 0,950 | 0,950 |
| $\alpha/2$ | 0,025 | 0,025 |
| t($\alpha/2$) | 2,571 | 2,571 |
| usalduspiiri kaugus | 2,410 | 1,544 |
| u | -2,079 | -0,646 |
| U | 2,741 | 2,442 |

Selgub, et mõlemad usalduspiirkonnad sisaldavad õiget keskväärtust, mis käesoleval juhul on 0,5. Usalduspiirkonnad on väga laiad sellepärast, et vaatlusi on väga vähe. Ühtlasi on näha ka see, et kumbki usalduspiirkond sisaldab teise valimi põhjal saadud punkthinnangut.

**Järelemõtlemiseks**

1.Kuidas muutuvad usalduspiirid, kui usaldustõenäosust suurendada 0,95-lt 0,99-le?

2.Kuidas sõltuvad usalduspiirid lähtetunnuse hajuvusest?

3.Kuidas sõltuvad usalduspiirid lähtetunnuse hajuvuse hinnangust?

4.Mitu korda tuleks valimi mahtu suurendada, et usalduspiirid muutuksid kaks korda kitsamaks?

5.Kuidas muutuvad usalduspiirid, kui punkthinnang suureneb?

# STATISTILISTE HÜPOTEESIDE KONTROLLIMINE

## STATISTILISE HÜPOTEESIDE KONTROLLIMISE TEOORIA PÕHIMÕISTED

### Millal ja milleks on tarvis kontrollida statistilisi hüpoteese?

Matemaatiline statistika on metoodiliseks aluseks enamusele empiirilistest uuringutest – toimugu need mistahes valdkonnas – eluteadustes, ühiskonnateadustes, tehnikas jm. Empiirilistele teadusuuringutele on omane olukorda, kus mõõtmistulemuste põhjal tehakse sisulisi järeldusi, mis teatavas mõttes kontrollivad seniste teoreetiliste vaadete paikapidavust ja mõnikord viivad seisukohtade muutumiseni.

- Millal saab empiiriliste andmete põhjal öelda, et midagi on oluliselt muutunud?
- Kuidas eristada juhuslikke hälbeid ja sisulisi muutusi?
- Kui palju tuleb teha katseid, et midagi veenvalt kinnitada?

Nendele ja paljudele teistele küsimustele võimaldab anda vastuse statistiliste hüpoteeside kontrollimise teooria.

| |
|---|
| Statistiliste hüpoteeside teooria tulemuste rakendamiseks sõnastatakse sisulistele, erialaspetsiifilistele hüpoteesidele vastavad statistilised hüpoteesid, rakendatakse nende kontrollimiseks välja töötatud aparatuuri ja tõlgendatakse tulemusi erialaselt. |

### Statistiliste hüpoteeside paar

1.Eeldame, et on olemas üldkogum, mille jaotusseadus on teada, kuid pole teada seda jaotust identifitseerivad parameetrid.

2.Sellest üldkogumist on olemas esindav valim mahuga $n$, $x_1$, $x_2$, …, $x_n$.

3.Uurijat huvitab tõestada, et üldkogumi parameeter $\vartheta$ rahuldab mingit loogilist tingimust, näiteks kuulub mingisse väärtuste hulka, on mingist arvust suurem jne.

4.Uurija määrab otsustuse juures kasutatava olulisuse nivoo $\alpha$, mis näitab suurimat lubatavat vea tõenäosust soovitava hüpoteesi tõestamisel.

5.Uurija sõnastab teineteist välistavate hüpoteeside paari, mille puhul sisukas hüpotees ($H_1$) väljendab (tavaliselt) uurija poolt soovitud tulemust, selle eitus – nullhüpotees ($H_0$) – aga vastupidist olukorda. Näiteks:

$$H_1\colon\ \vartheta > c,$$
$$H_0\colon\ \vartheta \le c.$$

Näites esitatud sisukas hüpotees on väljendatud ühe võrratusega, sellist hüpoteesi

nimetatakse ühepoolseks hüpoteesiks. Teine väga levinud hüpoteesipaar koosneb ***lihtsast nullhüpoteesist*** ja kahepoolsest sisukast hüpoteesist:

$$H_1: \vartheta \neq c;$$
$$H_0: \vartheta = c.$$

Niisugust hüpoteesipaari kasutatakse siis, kui soovitakse tõestada mingi muutumise või arengu toimumist, kusjuures pole ette teada, mis suunas see muutumine on toimunud.

Matemaatilise statistika meetoditega ***saab üldiselt tõestada ainult sisukat hüpoteesi $H_1$***. See, kui arutluskäigu tulemusena võetakse vastu nullhüpotees, ei ole selle hüpoteesi tõestus.

**Vead statistiliste hüpoteeside kontrollimisel**

Kuna valimi põhjal tehtavad järeldused on juhuslikud, tuleb arvestada juhuslike vigade võimalust nende järelduste tegemisel. Et hüpoteesid ei ole sõnastatud sümmeetrilistena, vaid üks on nö soovitav, siis on ka vead erinevad. Tehtavaid vigu iseloomustab alljärgnev tabel:

| Valimi põhjal tehtav järeldus | | |
|---|---|---|
| | Võetakse vastu $H_1$ | Võetakse vastu $H_0$ |
| Vastab $H_1$-le | Õige otsus | II liiki viga |
| Vastab $H_0$-le | I liiki viga | Õige otsus |

Vigadest peetakse eriti ebasoovitavaks I liiki viga, mille sisuks on soovitava tulemuse ekslik tõestamine. Selle vea esinemise tõenäosus ei tohi olla liiga suur, ning selle piiramiseks on võetud kasutusele olulisuse nivoo mõiste.

**Olulisuse nivoo ja võimsus**

***Olulisuse nivoo*** on maksimaalne lubatav eksimise tõenäosus soovitava hüpoteesi $H_1$ tõestamisel, st suurim lubatav I liiki vea tõenäosus. Olulisuse nivoo tähiseks on $\alpha$ ja tema väärtuseks on kõige sagedamini 0,05 ja 0,01. Olulisuse nivoo määrab uurija, selle suuruses lepitakse tavaliselt kokku enne uuringu algust. Üldiselt on teada, et mida väiksem on olulisuse nivoo, seda veenvam on hüpoteesi $H_1$ tõestus. Tuleb aga arvestada ka seda, et väiksema olulisuse nivooga tõestamiseks on tavaliselt tarvis suurema mahuga valimit. Mõistlik olulisuse nivoo ei ületa kunagi väärtust 0,5.

Hüpoteeside kontrollimise juures ei ole võimalik korraga piirata I ja II liiki vea tõenäosusi, seetõttu võib II liiki vea tõenäosus olla konkreetsete ülesannete korral olla küllalt suur (teist liiki vea tõenäosuse maksimaalne võimalik väärtus on $1-\alpha$). Siiski on hüpoteeside kontrollimise eeskiri e. kriteerium üldiselt seda parem, mida väiksem on II liiki vea tõenäosus. Tõenäosust, et II liiki viga ei esine, nimetatakse ***võimsuseks***.

**Hüpoteesi kontrollimise eeskirja (kriteeriumi) konstrueerimine ja otsuse vastuvõtmine**

6.Tuleb leida mingi statistik (valimi funktsioon), mille jaotus nullhüpoteesile vastava olukorra puhul on teada.

7.Selle statistiku väärtus arvutatakse valimi põhjal välja.

8.Kasutades statistiku jaotust eeldusel, et kehtib nullhüpotees, leitakse tõenäosus valimi põhjal arvutatud statistiku väärtuse (sellest vastavalt hüpoteesi sisule suurema/ väiksema) väärtuse saamiseks. Nii leitud tõenäosust nimetatakse olulisuse tõenäosuseks.

9.Otsustuse vastuvõtmiseks võrreldakse olulisuse tõenäosust olulisuse nivooga.

•Kui olulisuse tõenäosus on väiksem kui olulisuse nivoo, siis on valim selline, et selle saamine nullhüpoteesile vastavast üldkogumist on vähe tõenäone ja selletõttu kummutatakse nullhüpotees ning võetakse vastu sisukas hüpotees.

•Kui olulisuse tõenäosus on suurem kui olulisuse tõenäosus, siis on valimi pärinemine nullhüpoteesile vastavast üldkogumist küllaltki tõenäone ja sellega nullhüpoteesi ei kummutata.

Statistikute olulisuse tõenäosused arvutatakse tavaliselt programmiliselt. Klassikaline hüpoteeside kontrollimise viis on järgmine (erinevus alates 8. sammust):

8) Kasutatakse statistiku jaotust nullhüpoteesi eeldusel, mis on tabuleeritud. Leitakse olulisuse nivoole vastav täiendkvantiil e. protsentpunkt.

9) Võrreldakse arvutatud statistiku väärtust selle protsentpunktiga. Kui olulisuse tõenäosus on suurem kui olulisuse nivoo, siis loetakse sisukas hüpotees tõestatuks, vastasel korral võetakse vastu nullhüpotees.

**Seos hüpoteeside kontrollimise ja vahemikhindamise vahel**

Kui on teada selle parameetri vahemikhinnang, mille kohta soovitakse hüpoteese kontrollida, siis on viimane ülesanne lihtsalt lahendatav.

• Kui hinnatava parameetri $\vartheta$ jaoks on leitud $(1-\alpha)$-usalduspiirid $[u, U]$,

• siis saame kontrollimaks hüpoteesipaari

$$H_1: \ \vartheta \neq c;$$
$$H_0: \ \vartheta = c.$$

• olulisuse nivool $\alpha$ järgmise kriteeriumi:

$$u<c<U \Rightarrow H_0; \quad c \leq u \text{ või } c \geq U \Rightarrow H_1.$$

Kõige olulisemaid statistilisi hüpoteese saabki kontrollida eelmisel loengul tuletatud statistika põhijaotuste – $t$- ja $\chi^2$-jaotuse abil.

**Järelemõtlemiseks**

1.Olgu antud kriteerium ühepoolse hüpoteesi kontrollimiseks. Kas seda saab kasutada ka siis, kui vahetada sisukas ja nullhüpotees?

2.Missuguse ülesande lahendamiseks on kasutusele võetud $t$-jaotus?

3.Kas $t$-jaotus sõltub kasutatava valimi mahust? Kuidas?

4.Missuguses vahemikus saab muutuda usaldusnivoo? Olulisuse nivoo?

5.Kas usaldusnivoo/ olulisuse nivoo on tavaliselt suur (lähedane 1-le) või väike (lähedane 0-le)?

6.Missugune on olulisuse nivoo ja olulisuse tõenäosuse vahekord?

VASTUSED

# STATISTILISTE HÜPOTEESIDE KONTROLLIMINE

## HÜPOTEESIDE KONTROLLIMINE ÜLDKOGUMI KESKVÄÄRTUSE KOHTA

**Kõige sagedamini kontrollitavad hüpoteesid ühe keskväärtuse kohta**

Ühe juhusliku suuruse keskväärtuse kohta kontrollitakse enamasti üht järgmistest hüpoteesipaaridest:

1.Soovitakse tõestada, et keskväärtus on mingist konstandist suurem (näiteks: tõestatakse mingi näitaja kasvu, arengut, suurenemist võrreldes teatava standardiga). Siis sõnastatakse alljärgnev hüpoteesipaar:

$$H_1\colon\ EX > c$$
$$H_0\colon\ EX \le c.$$

2.Soovitakse tõestada, et keskväärtus on mingist konstandist (standardist) väiksem. Niisuguseid hüpoteese on vahel tarvis tõestada töökindluse teoorias jm. Sõnastatakse järgmised hüpoteesid:

$$H_1\colon\ EX > c$$
$$H_0\colon\ EX \ge c.$$

3.Soovitakse tõestada, et on toimunud muutus, kuigi pole teada, mis suunas see toimus. Sõnastatakse järgmised hüpoteesid:

$$H_1\colon\ EX \ne c$$
$$H_0\colon\ EX = c.$$

| |
|---|
| Viimases hüpoteesipaaris ei saa sisukat ja nullhüpoteesi vahetada. Statistikameetodite abil pole võimalik tõestada, et keskväärtus võrdub teatava konstandiga! |

**Ühepoolse hüpoteesi *EX>c* kontrollimine normaaljaotuse keskväärtuse kohta**

Vaatame, kuidas tõestada esimest võrratust eeldusel, et *X* on normaaljaotusega.

1. Kõigepealt lepitakse kokku, missugust olulisuse nivood kasutada. Oletame, et sobib $\alpha$=0,05.

2. Järgmine ülesanne on leida sobiv statistik. Selleks sobiks valimkeskmise ja tema standardhälbe jagatis, mis on normaaljaotusega, kuid kahjuks pole tavaliselt standardhälve teada, ning selle asemel tuleb kasutada tema hinnangut. Valimkeskmise ja tema standardhälbe suhe on *t*-jaotusega ja **põhiliselt kasutataksegi keskväärtuse kohta käivate hüpoteeside kontrollimiseks *t*-jaotust.**

3. Järgmine ülesanne on selgitada, missuguste statistiku väärtuste korral võetakse

sisukas hüpotees vastu. *Seda statistiku väärtuste hulka, mille korral sisukas hüpotees vastu võetakse, nimetatakse* ***kriitiliseks piirkonnaks*** *ja selle määrab nn* ***kriitiline väärtus.*** On loogiliselt mõistetav, et selleks, et võtta vastu sisukas hüpotees $EX>c$, peab olema valimkeskmine $\bar{x}$ suurem kui $c$, kuid kui palju suurem? On tarvis leida kriitiline väärtus $C^*$ nii, et siis, kui nullhüpotees on õige, ületaks statistik kriitilist väärtust tõenäosusega, mis ei ületaks olulisuse nivood, st et kehtiks alljärgnev võrratus tingliku tõenäosuse jaoks:

$$P(x > C^* \mid EX \le c) \le \alpha.$$

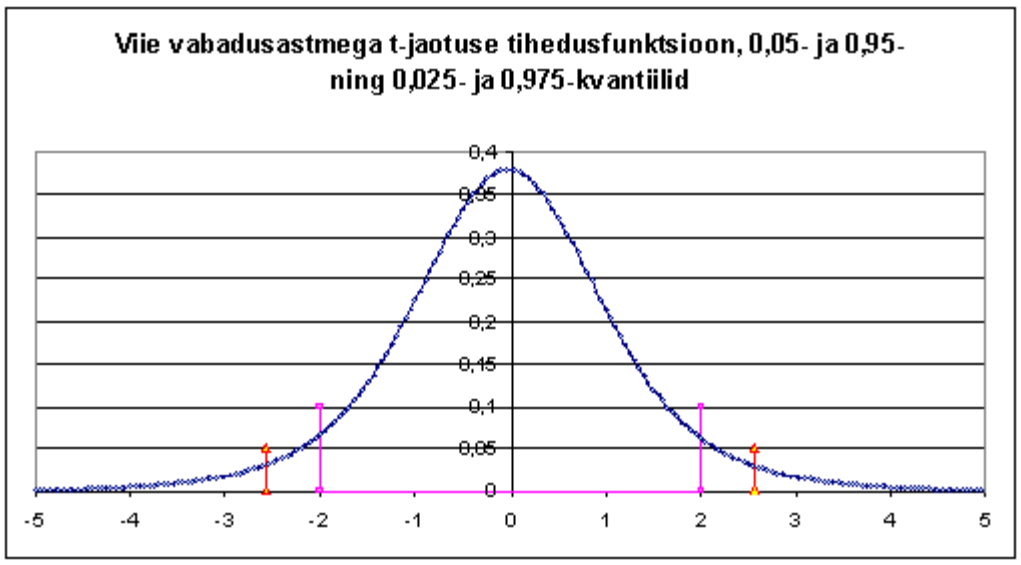
4. Kasutame eelmisel loengul kindlaks tehtud tõsiasja, et valimkeskmist ja standardhälbe hinnangut sisaldav $t$-statistik on $t$-jaotusega vabadusastmete arvuga $n-1$.

$$\frac{\bar{x}-EX}{s}\sqrt{n} \sim t_{n-1}.$$

5. See tähendab, et tema jaotus on teada ja vastavalt antud tõenäosusele saab leida kõik vajalikud kvantiilid. Seega kehtivad võrratused

$$P(t < -t_{n-1}(\alpha)) = \alpha \text{ ja } P(t > t_{n-1}(\alpha)) = \alpha,$$

kus $t$ on $t$-jaotusega statistik ja $-t_{n-1}(\alpha)$ ning $t_{n-1}(\alpha)$ on vastavalt $\alpha$-*kvantiil* ja $\alpha$-täiendkvantiil (vt roosad lõigud lisatud joonisel. Statistikatabelites esitatakse sageli nimelt ***täiendkvantiile***, so väärtusi $t(\alpha)$, mis rahuldavad tingimust $P(t > t(\alpha)) = \alpha$.




6. Võttes $t$-statistiku avaldises *EX väärtuseks hüpoteesis esineva väärtuse c, mis on äärmiseks väärtusek, mille korral nullhüpotees kehtib*, saame kirjutada võrratuse:

$$P\left(\frac{\bar{x}-c}{s}\sqrt{n} > t_{n-1}(\alpha)\right) = \alpha \Rightarrow P\left(\bar{x} - c + \frac{s}{\sqrt{n}} t_{n-1}(\alpha)\right) = \alpha.$$

Siit järeldub, et otsitav ***kriitiline väärtus*** on

$$C^* = c + \frac{s}{\sqrt{n}} t_{n-1}(\alpha).$$

Seega saame kriteeriumi:

•Kui $\bar{x} < C^*$, siis kehtib (on tõestatud olulisuse nivool $\alpha$) sisukas hüpotees $H_1$;

•kui aga $\bar{x} \geq C^*$, siis võetakse vastu nullhüpotees, mis tähendab, et sisukat hüpoteesi ei õnnestu tõestada.

***Sellega ei ole nullhüpotees tõestatud***!

**Ühepoolse hüpoteesi *EX<c* kontrollimine normaaljaotuse keskväärtuse kohta**

Esitatud mõttekäiku korrates näeme, et sisuka hüpoteesi $EX < c$ jaoks saame samuti tuletada kriitilise väärtuse $C^*$, mille tulemusena saame kriteeriumi:

$$Kui\ \bar{x} < c - \frac{s}{\sqrt{n}} t_{n-1}(\alpha),\quad siis\ kehtib\ H_1 (EX<c);$$

$$Kui\ \bar{x} < c - \frac{s}{\sqrt{n}} t_{n-1}(\alpha),\quad siis\ H_1\ ei\ kehti,\ võetakse\ vastu\ H_0.$$

Võrreldes esitatud kriteeriume on lihtne veenduda selles, et sisukas ja nullhüpotees ei ole sümmeetrilised.

**Kahepoolse hüpoteesi *EX≠c* kontrollimine normaaljaotuse keskväärtuse kohta**

Kahepoolse hüpoteesi kontrollimiseks saame samuti kasutada mõttekäiku, mida rakendasime ka ühepoolse hüpoteesi kontrollimiseks, kuid paneme tähele, et sisuka hüpoteesi vastuvõtmise piirkond koosneb kahest osast – niihästi suurtest kui ka väikestest statistiku väärtustest. Seetõttu saab I liiki viga tekkida kahel viisil ning on loomulik eeldada, et kummagi viisi tõenäosus on ühesugune, suurusega $\alpha/2$.

| |
|---|
| Seega saame kahepoolse hüpoteesi vastuvõtmise piirkonna konstrueerida kahe ühepoolse hüpoteesi vastuvõtmise piirkonna summana, kusjuures need on leitud, kasutades olulisuse nivood $\alpha/2$. |

**Näide 1**.

Vaatleme kolme ülesannet, mis tuleb lahendada olulisuse nivool 0,05.

A.On tarvis tõestada, et juhusliku suuruse keskväärtus on positiivne, st

$$H_1: EX > 0;$$

B.On tarvis tõestada, et juhusliku suuruse keskväärtus ei ole suurem kui 5, st

$$H_1: EX = 3;$$

C. On tarvis tõestada, et juhusliku suuruse keskväärtus erineb väärtusest 1,5, st

$$H_1: EX \neq 1{,}5.$$

Nende hüpoteeside kontrollimiseks saame kasutada kaht 6-elemendilist valimit, mida on kirjeldatud eelmises peatükis näidetes 1 ja 2. Lisaks sellele ühendame need valimid ja moodustame uue, 12-elemendilise valimi. Vajalikud arvutustulemused on esitatud alljärgnevas tabelis.

| | Valim 1 | Valim 2 | Ühendvalim 3 |
|---|---|---|---|
| | | | |

| | | | |
|---|---|---|---|
| Keskmine | 0,331 | 0,898 | 0,615 |
| Dispersioon | 5,273 | 2,165 | 3,469 |
| Standardhälve | 2,296 | 1,471 | 1,862 |
| Ruutjuur mahust | 2,449 | 2,449 | 3,464 |
| Standardviga | 0,937 | 0,601 | 0,538 |

| | Vabadusastmete arv | Vabadusastmete arv |
|---|---|---|
| Tõe-näosus | 5 | 11 |
| 0,05 | 2,015 | 1,796 |
| 0,025 | 2,571 | 2,201 |

Selleks, et konstrueerida hüpoteesi vastuvõtmise piirkondi, on tarvis leida t-jaotuse täiendkvantiilid. Jaotuse sümmeetrilisuse tõttu kehtib seos

$$P(t < -t(\alpha)) = P(t > t(\alpha)).$$

1. Üleasande A lahendamiseks saame 1., 2. ja 3. valimi puhul kriitiliseks piirkonnaks vastavalt:

$C_1^* = 0 + 0{,}937 \times 2{,}015 = 1{,}889$; samal viisil saame $C_2^* = 1{,}21$ ja $C_3^* = 0{,}966$.

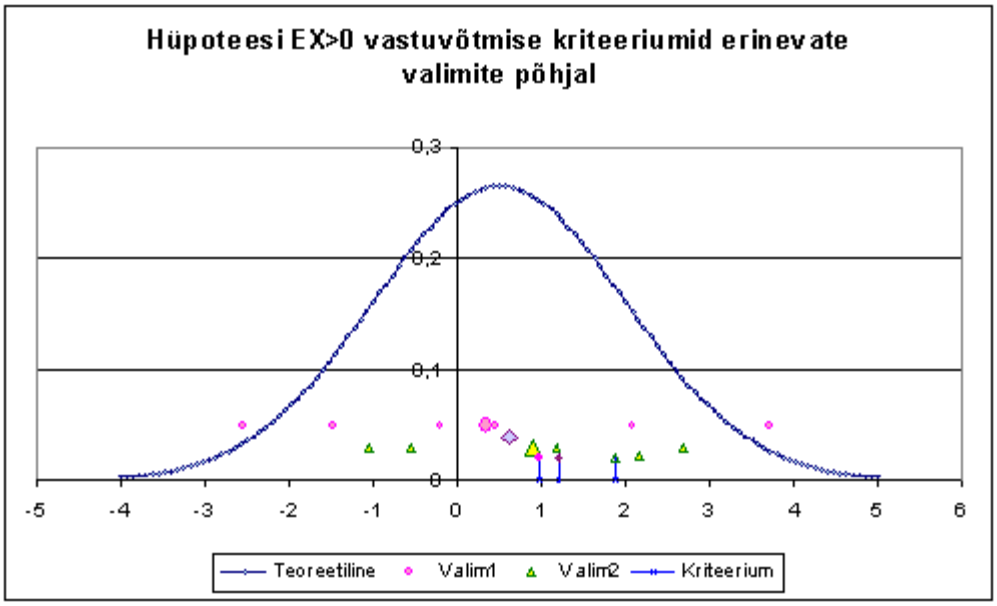
Kõigi kolme kriteeriumi puhul on tulemus sama, sest iga valimkeskmine on kriitilisest väärtusest väiksem. ***Me ei saa võtta vastu sisukat hüpoteesi, et uuritava juhusliku suuruse keskväärtus on positiivne.***

2.Ülesande B lahendamisel saame kriitilise piirkonna arvutamisel kasutada juba tehtud arvutusi. Kriitiline piir hüpoteesi $EX < 3$ tõestamiseks on vastavalt esimese valimi põhjal 3–1,889 = 1,111, teise valimi põhjal – 1,79 ja summaarse valimi põhjal 2,034. Selgub, et iga valimi keskmine on kriitilisest piirist väiksem, seega saame kõigil juhtudel sisuka hüpoteesi vastu võtta.

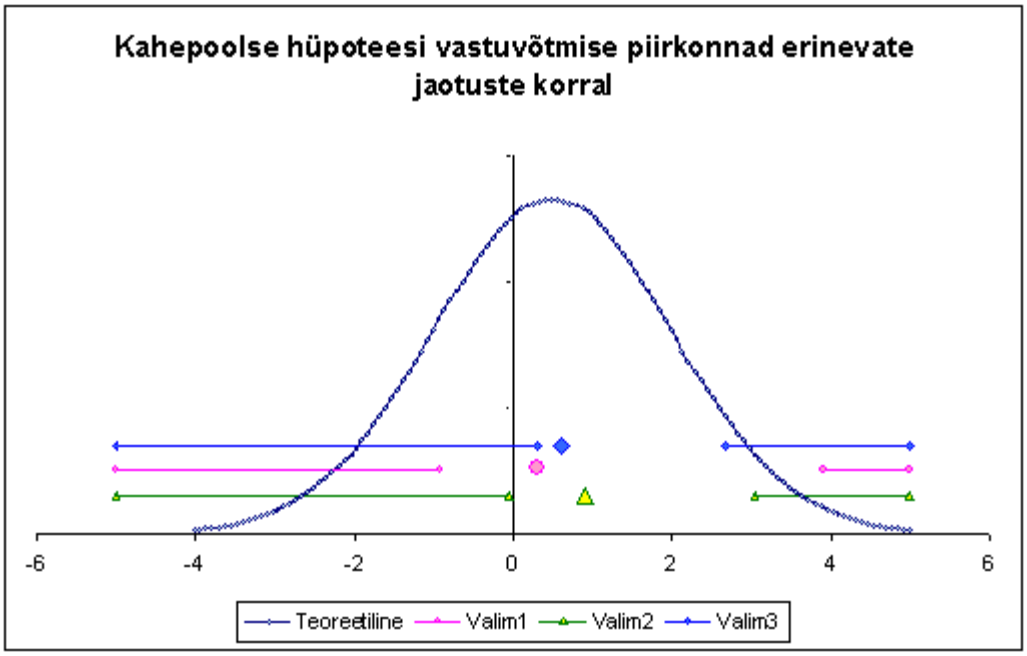
3.Ülesande C lahendamisel koosneb sisuka hüpoteesi vastuvõtmise piirkond kahest osast,

| | 1.valim | 2.valim | 3.valim |
|---|---|---|---|
| parandusliige | 2,40986 | 1,54413 | 1,183 |
| $C^*_1$ | -0,9099 | -0,04413 | 0,317 |
| $C^*_2$ | 3,90986 | 3,04413 | 2,683 |

sisukas hüpotees võetakse vastu siis, kui $\overline{x} < C^*_1$, samuti ka siis, kui $\overline{x} > C^*_2$. Antud kolme valimi põhjal arvutatud kriitilised väärtused on lisatud tabelis.




Lisatud joonisel on erinevate värvidega kujutatud kõigi valimite põhjal arvutatud hüpoteesi $H_1$: $EX$ = 1,5 vastuvõtmise piirkonnad Et keskväärtuse hinnang ühelgi juhul sellesse piirkonda ei satu, ei õnnestu seda hüpoteesi tõestada.




Kokkuvõttes saime tõestada ainult seda, et juhusliku suuruse keskväärtus on väiksem kui 3, kuid ei saanud kinnitust ei väitele, et keskväärtus on positiivne ega õnnestunud tõestada ka seda, et ta erineb arvust 0,5. Saadud tulemused ei ole vastuolulised, sest neist kaks on vaid vastuvõetud nullhüpoteesid, mitte tõestused.

**Tulemuste üldistamine juhule, kui lähtejaotus erineb normaaljaotusest**

Kogu esitatud mõttekäik kehtib täpselt vaid sel juhul, kui lähtejaotus on normaaljaotusega. Osutub siiski, et saadud tulemus on suhteliselt ***robustne*** ehk stabiilne jaotuse suhtes. Ka sel juhul, kui lähtejaotus erineb normaaljaotusest, kuid pole väga ebasümmeetriline või eriti raskete sabadega, läheneb *t*-statistiku jaotus *t*-jaotusele, nii et esitatud metoodika on kasutatav.

**Kahe normaaljaotuse keskmiste võrdlemine (sõltuvad vaatlused)**

Teine väga sageli esinev ülesanne puudutab kahe normaaljaotusega juhusliku suuruse keskmiste võrdlemist valimite põhjal, st ühe hüpoteesipaari kontrollimist järgmiste hüpoteesipaaride hulgast:

$$H_1: EX_1 > EX_2,$$
$$H_0: EX_1 \leq EX_2;$$

$$H_1: EX_1 < EX_2,$$
$$H_0: EX_1 \geq EX_2;$$

$$H_1: EX_1 \neq EX_2,$$
$$H_0: EX_1 = EX_2;$$

Siin eristatakse põhiliselt kaht juhtu – sõltuvate ja sõltumatute vaatluste juht.

| Sõltuvate vaatluste juhuga on tegemist siis, kui valim koosneb samadest objektidest, keda on mõõdetud kaks korda. |
|---|

Kui esimese mõõtmise tulemused moodustavad juhusliku suuruse $X_1$ ja teise mõõtmise tulemused juhusliku suuruse $X_2$, siis on kõige sobivam leida vahe $Y = X_2 - X_1$ ning rakendada $EY$ jaoks eelmistes punktides arendatud metoodikat.

**Kahe normaaljaotuse keskmiste võrdlemine**
**(sõltumatud vaatlused)**

Olgu tarvis kahe üldkogumi $X_1 \sim N(\mu_1, \sigma)$ ja $X_2 \sim N(\mu_2, \sigma)$ keskmiste kohta kontrollida üks kolmest ülalmärgitud hüpoteesipaarist. Eeldame, et kummastki juhuslikust suurusest on olemas valim vastavalt mahuga $n_1$ ja $n_2$. Niisuguseid vaatlusi loetakse sõltumatuteks. Lisaeelduseks on see, et mõlemad juhuslikud suurused on sama hajuvusega.

Siis saab kõiki ülaltoodud hüpoteese kontrollida $t$-testi abil. Selleks arvutatakse mõlema valimi põhjal keskmise ja standardhälbe hinnangud $\bar{x}_1$, $\bar{x}_2$, $s_1$ ja $s_2$.. Nende abil arvutatakse $t$-statistik:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}, \quad s = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}; \quad t \sim t_{n_1+n_2-2}.$$

• Kui soovitakse kontrollida ühepoolset hüpoteesi $H_1: EX_1 > EX_2$, siis võrreldakse $t$-tatistiku väärtust $t$-jaotuse a-täiendkvantiiliga: kui $t > t(\alpha)$, siis kehtib $H_1$.

• Kui soovitakse kontrollida ühepoolset hüpoteesi $H_1: EX_1 < EX_2$, siis võrreldakse $t$-statistiku väärtust $t$-jaotuse $\alpha$-kvantiiliga (see on negatiivne): kui $t < t(\alpha)$, siis kehtib $H_1$.

• Kui soovitakse kontrollida kahepoolset hüpoteesi $H_1: EX_1 \neq EX_2$, siis võrreldakse $t$-statistiku absoluutväärtust $t$-jaotuse $\alpha/2$-täiendkvantiiliga: $|t| < t(\alpha/2)$, siis kehtib $H_1$.

**Ühepoolse ja kahepoolse hüpoteesi vahekord**

Tõestatav sisukas hüpotees tuleb sõnastada enne uuringu algust, st varasema info põhjal.

Kui pole eelteavet selle kohta, kumb keskmistest on suurem, tuleb valida tõestamiseks kahepoolne hüpotees. Kui aga on tarvis tõestada, et mingi näitaja on suurenenud, et ühes (konkreetses) üldkogumis on keskmine suurem kui teises jne, on otstarbekas valida tõestamiseks ühepoolne hüpotees. Ühepoolse hüpoteesi tõestamine kasutab vaatlusi ökonoomsemalt.

**Järelemõtlemiseks**

1.Uurija soovib tõestada, et juhusliku suuruse keskväärtus on > 3. Selgub, et valimkeskmine on 2,8. Mida uurija peaks edasi tegema? Kas tulemus sõltub olulisuse nivoost (eeldades, et see on < 0,5).

2.Uurija soovib tõestada etteantud olulisuse nivool, et juhusliku suuruse keskväärtus on > 3. Valimkeskmine on küll 3,5, kuid sellest ei piisa sisuka hüpoteesi vastuvõtmiseks. Mida saab uurija nüüd teha?

3.Uurija soovib tõestada, et juhusliku suuruse keskväärtus on > 3. Valimkeskmine on küll 3,5, kuid sellest ei piisa sisuka hüpoteesi vastuvõtmiseks. Puudub ka võimalus täiendavaid katseid teha. Mida saab uurija teha?

4.Uurijal on võimalik teha kahe keskväärtuse võrdlemiseks t-testi abil kokku $n$ katset. Kuidas oleks tal kõige otstarbekam need katsed jaotada esimese ja teise juhusliku suuruse valimiteks? Miks?

5.Missugust jaotust saab $t$-jaotuse asemel kasutada keskväärtuste võrdlemiseks suurte valimite ($n > 100$) korral?

6.Missugust jaotust saab $t$-jaotuse asemel kasutada keskväärtuste võrdlemiseks siis, kui võrreldavate juhuslike suuruste dispersioonid on täpselt teada?

VASTUSED

## STATISTILISTE HÜPOTEESIDE KONTROLLIMINE

## MÕNINGATE MUUDE STATISTILISTE HÜPOTEESIDE KONTROLLIMINE

### Hüpoteeside kontrollimine korrelatiivse seose kohta

Vaatleme juhuslikku vektorit $(X,Y)$ ja eeldame, et sellest on tehtud $n$ vaatlust ning vaatluste põhjal arvutatud välja korrelatsioonikordaja hinnang

$$r = \frac{\sum_{i=1}^{n} (x_1 - \bar{x}_2)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x}_2)^2 \cdot \sum_{j=1}^{n} (y_j - \bar{y})^2}}.$$

Kontrollimist vajavad tavaliselt hüpoteesid selle kohta, kas korrelatiivne sõltuvus on statistiliselt oluline, st et kas üldkogumis on korrelatsioonikordaja nullist erinev. Vastavalt sellele on tarvis kontrollida hüpoteesipaare üldkogumi korrelatsioonikordaja $r(X, Y)$ kohta:

$$H_1: r(X, Y) > 0,$$
$$H_0: r(X, Y) \leq 0;$$

$$H_1: r(X, Y) < 0,$$
$$H_0: r(X, Y) \geq 0;$$

$$H_1: r(X, Y) \neq 0,$$
$$H_0: r(X, Y) = 0.$$

Osutub, et valimi korrelatsioonikordaja abil saame asümptootiliselt $t$-jaotusega statistiku, mille abil õnnestub küllalt edukalt loetletud hüpoteese kontrollida.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}.$$

***Näide 2.***

Olgu antud 10 üliõpilase tulemused kontrolltööl ja õppetööl osalemise sagedused. Soovitakse kontrollida (olulisuse nivool 0,05) hüpoteesi nende näitajate korrelatiivse seose kohta. Et eelduse kohaselt õppetööl osalemine mõjub tulemustele positiivselt, valime kontrollimiseks ühepoolse sisuka hüpoteesi $r(X, Y) > 0$.

Andmed ja arvutustulemused on esitatud lisatud tabelis:

| Jrk | osalus | tulemus | näitaja | osalus | tulemus |
|---|---|---|---|---|---|

| nr | | | | | |
|---|---|---|---|---|---|
| 1 | 5 | 65 | keskmine | 5,9 | 72,5 |
| 2 | 7 | 90 | dispersioon | 1,449 | 20,58 |
| 3 | 5 | 70 | | Korrelatsioonikordaja r | 0,531 |
| 4 | 6 | 100 | | $r^2$ | 0,282 |
| 5 | 3 | 40 | | $1-r^2$ | 0,718 |
| 6 | 6 | 70 | | $\sqrt{n-2}$ | 2,828 |
| 7 | 7 | 40 | | $t$ | 2,091 |
| 8 | 8 | 95 | | 0,05-täiendkvantiil | 1,86 |
| 9 | 7 | 80 | | 0,025-täiendkvantiil | 2,306 |
| 10 | 5 | 75 | | | |

Korrelatsioonikordaja on 0,531. Et *t*-statistiku väärtus on 2,091 ja 2,091 > 1,86, siis on võimalik sisukat hüpoteesi $H_1$ tõestada – õppetöös osalemise ja kontrolltöö tulemuste vahel on positiivne korrelatsioon.

On näha, et ilma eelinformatsioonita korrelatiivse seose märgi kohta ei olnuks käesoleval juhul võimalik seose olulisust tõestada.

**Hüpoteeside kontrollimine jaotuste erinevuse kohta (jaotuste võrdlemine)**

Olgu teoreetiline jaotus *P* tõenäosusfunktsiooniga $P(X = a_i) = p_i$; i = 1, 2, ..., *m*, ning olgu tehtud *n* vaatlust, mille tulemusena on saadud väärtuste $a_i$ sageduseks $k_i$, $i=1, ..., m$, kusjuures $k_1 + k_2 + ... + k_m = n$.

Nende andmete põhjal saab kontrollida järgmist hüpoteesipaari:

$H_1$: Üldkogum (millest pärinevad vaatlused) ei ole jaotusega *P*,

$H_0$: Üldkogum on jaotusega *P.*

Tähelepanu tuleb pöörata tõsiasjale, et jaotuste kooskõla väidab nullhüpotees, seega pole võimalik tõestada valimi pärinemist mingist teoreetilise jaotusega üldkogumist, küll aga on võimalik tõestada, et vaadeldav valim ei saa teatava teoreetilise jaotusega üldkogumist pärineda või see on väga väikese tõenäosusega sündmus.

**Empiirilise ja teoreetilise jaotuse võrdlemine $\chi^2$-statistiku abil**

Üks võimalusi jaotuste kooskõla kontrollimiseks tugineb $\chi^2$-statistiku kasutamisele. Selle meetodi puhul tuleb aga arvestada, et tegemist on asümptootilise meetodiga, mida ei saa rakendada väikeste valimite korral. Meetodi aluseks on eeldus, et pika katseseeria korral sagedus läheneb tõenäosusele, kusjuures sageduse jaotus läheneb normaaljaotusele, nii et ligikaudselt kehtib iga *i* korral järgmine seos:

$$\frac{k_i}{n} \sim N(p_i ; p_i(1 - p_i)) \text{ ehk } \left( \frac{k_i}{n} - p_i \right) / \sqrt{p_i(1 - p_i)} \sim N(0,1).$$

Kui tõenäosusfunktsioon määrab ära m suhtelist sagedust, siis on katseseeriaga määratud m ligikaudu normaaljaotusega $N(0,1)$ juhuslikku suurust, neist igaüks vastab ühe väärtuse $a_i$ esinemissageduselae. Võttes kõik need avaldised ruutu ja liites, saame juhusliku suuruse $Y$, mis võrdub $m$ standardiseeritud normaaljaotusega juhusliku suuruse ruudu summaga. Kui need liidetavad oleksid sõltumatud, oleks $Y$ jaotuseks $\chi^2$-jaotus vabadusastmete arvuga $m$. Tegelikult nad pole sõltumatud, sest nende summa on $n$. Osutub aga, et nn hii-ruut-statistik $H$,

$$H = \sum_{i=1}^{n} \frac{(k_i - n^2 p_i)^2}{np_i},$$

mis on saadud nimetatud summast teisendamise teel, on $\chi^2$-jaotusega vabadusastmete arvuga $m - 1$.

Selletõttu toimub püstitatud hüpoteesi kontrollimine järgmiselt:

•Määratakse kasutatav olulisuse nivoo $\alpha$;

•Arvutatakse statistiku $H$ väärtus;

•Leitakse $\chi^2$-jaotuse $a$-täiendkvantiil $h_{m-1}(\alpha)$;

•Kui $H > h_{m-1}(\alpha)$, siis võetakse vastu sisukas hüpotees $H_1$, vastasel korral jäädakse nullhüpoteesi juurde – valim võib pärineda jaotusega $P$ üldkogumist.

$H$-statistiku sisu on väga hästi arusaadav – kui empiirilise ja teoreetilise jaotuse vahel on väga hea kooskõla, st et

$$p_i \approx \frac{k_i}{n}$$

iga $i$ korral, siis on statistik $H$ väike ning jaotused hästi kooskõlas. Siis võetakse vastu nullhüpotees. Kui aga mõned suhtelised sagedused tõenäosustest väga tugevasti erinevad, siis on $H$-statistik suur ja nullhüpotees kummutatakse. Siinjuures on näha, et eriti tugevasti mõjustavad $H$-statistiku väärtust väikese tõenäosusega juhusliku suuruse väärtused $a_i$.

Meetodi asümptootikast tingitud võimaliku vea vähendamiseks soovitatakse järgmist:

•Kui juhusliku suuruse $X$ mõne väärtuse $a_i$ tõenäosus on nii väike, et kehtib võrratus $np_i < 5$, siis soovitatakse mitte selle väärtuse esinemist eraldi sündmusena vaadelda, vaid ühendada mitu naaberväärtust ühiseks sündmuseks.

Olukord, kus teoreetiline jaotus on täpselt teada, esineb praktikas harva. Sagedamini on teada jaotuste pere, milles konkreetne jaotus identifitseeritakse parameetrite abil. Sel juhul hinnatakse parameetreid sama valimi põhjal, kusjuures sellega tekitatakse täiendavaid seoseid liidetavate $X_i$ vahel.

•Kui teoreetilise jaotuse parameetrid on valimi põhjal hinnatud, siis on hüpoteesi kontrollimisel kasutatava $\chi^2$-jaotuse vabadusastmete arv $m - r - 1$, kus $r$ on valimi põhjal hinnatud parameetrite arv.

•Enamus statistikaprogramme väljastab koos $H$-statistikuga tema olulisuse tõenäosuse $p$, mis näitab esimest liiki vea tegemise tõenäosust siis, kui $H_1$ vastu võetaks. Kui $p < \alpha$, siis tuleb võtta vastu $H_1$.

•Jaotuste kooskõla hindamise ülesannete puhul pakuvad huvi ka ***suure olulisuse***

***tõenäosusega*** juhud, näiteks kui $p > 0{,}5$. Sellisel korral öeldakse, et empiiriline jaotus on teoreetilise jaotusega ***hästi kooskõlas*** – kuigi pole võimalik tõestada, et valim nimelt selle teoreetilise jaotusega üldkogumist pärineb.

**Kahe jaotuse võrdlemine $\chi^2$-statistiku abil**

Esitatud metoodika on suhteliselt lihtsalt üldistatav ka juhule, kus soovitakse võrrelda valimite põhjal kaht teoreetilist jaotust $P_1$ ja $P_2$, millel on samad või ühitatavad väärtused $a_1$, $a_2$, ..., $a_m$. Seega kontrollitakse antud olulisusenivool $\alpha$ hüpoteesipaari:

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

Hüpoteeside kontrollimise aluseks on empiirilised jaotused, selletõttu nimetatakse seda ülesannet sageli ka kahe empiirilise jaotuse võrdlemiseks.

•Olgu esimesest empiirilisest jaotusest valim mahuga $n_1$, kusjuures üksikväärtuste sagedused on $k_{11}$, $k_{12}$, .., $k_{1m}$.

•Teise empiirilise jaotuse vastavad sagedused on $k_{21}$, $k_{22}$, .., $k_{2m}$ ja valimi maht $n_2$.

•Sel juhul loetakse teoreetiliseks jaotuseks ühisest valimist hinnatud sagedusi

$$p_i = \frac{k_{i1} + k_{i2}}{m_1 + m_2}.$$

•Hii-ruut-statistik H sisaldab kummagi valimi põhjal leitud hälberuute,

$$H = \sum_{j=1}^{2} \sum_{i=1}^{mj} \frac{(k_{ji} - n_j^2 p_i)^2}{n_j p_i}.$$

•Hii-ruut-statistik on asümptootiliselt $\chi^2$ jaotusega, kusjuures vabadusastmete arv on $m$ –1. Selle põhjenduseks on tõsiasi, et teoreetilise jaotuse iga tõenäosus on arvutatud valimi põhjal, seega leitakse vabadusastmete arv seosest $2m – m – 1$.

•Edasine mõttekäik on standardne – leitakse statistiku $H$ väärtus ning tema olulisuse tõenäosus $p$.

•Kui $p < \alpha$, siis on tõestatud sisukas hüpotees – valimid pärinevad erineva jaotusega üldkogumitest.

•Kui $p \geq \alpha$ , siis võivad mõlemad valimid pärineda ka ühesuguse jaotusega üldkogumist (mõlema valimi teoreetiline jaotus olla sama). Mida suurem on olulisuse tõenäosus, seda paremini on jaotused $P_1$ ja $P_2$ kooskõlas.

T**õenäosuste võrdlemine $\chi^2$-statistiku abil**

Eelmise ülesande erijuhuks on olukord, kus võrreldavad jaotused on binaarsed – nendel on ainult kaks väärtust – sündmuse $A$ esinemisel väärtus 1 ja mitte-esinemisel väärtus 0. Sel juhul on tegemist tõenäosuste võrdlemise ülesandega – kontrollimist vajab hüpotees, et sündmuse $A$ esinemise tõenäosus $p_1$ on mõlemas võrreldavas üldkogumis sama:

$$H_1: p_{11} \neq p_{21},$$

$$H_0: p_{11} = p_{21}.$$

Loomulikult järeldub sündmuse *A* tõenäosuste võrdsusest ka vastandsündmuse tõenäosuste võrdumine. Käesoleval juhul on hii-ruut-statistiku

$$H = \sum_{j=1}^{2} \sum_{i=1}^{2} \frac{(k_{ji} - n_j p_i)^2}{n_j p_i}$$

vabadusastmete arvuks 4–2–1 =1. Paneme tähele, et hii-ruut statistiku abil puudub võimalus ühepoolse hüpoteesi kontrollimiseks, et seda teha, tuleks kasutada lähendamist normaal- või *t*-jaotusega.

**Järelemõtlemiseks**

1.Miks ei saa tõestada kahe jaotuse võrdumist?

2.Kas põhimõtteliselt on võimalik tõestada seda, et kahe normaaljaotusega juhusliku suuruse erinevus on väiksem kui mingi fikseeritud suurus *d*?

3.Oletagem, et mingi valimi puhul kontrollitakse rida hüpoteese: kas on tegemist valimiga normaaljaotusest, ühtlasest jaotusest, Poissoni jaotusest jne, ning igal juhul saadakse tulemuseks nullhüpotees. Kuidas saadud tulemust interpreteerida?

4.Vaadeldakse kolme katseseeriat sündmuse *A* tõenäosuse määramiseks ning kontrollitakse, kas need annavad sama tulemuse. Selgus, et

•I ja II katseseeria tulemuste võrdlemisel võeti vastu nullhüpotees – tõenäosus võib olla sama.

•II ja III katseseeria tulemuste võrdlemisel võeti vastu nullhüpotees – tõenäosus võib olla sama.

•I ja III katseseeria tulemuste võrdlemisel võeti vastu sisukas hüpotees – tõenäosused on erinevad. Kuidas saadud tulemust tõlgendada?

VASTUSED


TAGASI

**Statistiline sõltuvus ja statistiline mudel**

## STATISTILINE SÕLTUVUS

Statistiline sõltuvus kahe tunnuse vahel
Ühe tunnuse tinglikud jaotused teise tunnuse suhtes
Statistilise sõltuvuse/ sõltumatuse tähtsus
Statistilise sõltuvuse tugevus ja olulisus
Statistilise seose tugevust iseloomustavad seosekordajad
Täielik statistiline sõltuvus
Statistilise sõltuvuse olulisus
Statistilise sõltuvuse tugevus. Crameri seosekordaja
Üldisemad statistilise seose kordajad
Järelemõtlemiseks

## LINEAARNE MUDEL

Hajuvusdiagramm ehk korrelatsiooniväli
Lineaarne mudel
Mudeli parameetrite määramine vähimruutude printsiibil
Prognoosid ja prognoosivead
Järelemõtlemiseks

## MUDELI STATISTILINE OLULISUS

Lineaarne mudel ja konstantne mudel
Mudeli olulisuse mõiste
F-jaotus
F-jaotuse kasutamine mudeli olulisuse kontrollimisel
Dispersioonanalüüsi tabel
Mudeli parameetrite usaldusvahemikud
Regressioonisirge usalduspiirid
Järelemõtlemiseks

## ÜLDISEMAD MUDELID

Mudelite üldistamise teed
Mitmene lineaarne regressioon
Argumentide funktsioonide kasutamine mudelis
Libatunnustega mudelid
Funktsioontunnuse teisendamine
Aegrea tüüpi mudelid
Järelemõtlemiseks

**Statistiline sõltuvus ja statistiline mudel**

**STATISTILINE SÕLTUVUS**

1.See ei muutu.

2.See ei muutu.

3.Mitteoluline seos võib muutuda oluliseks.

4.Lühema skaalaga tunnus on üldjuhul paremini määratud (teisest tugevamalt sõltuv) kui pikema skaalaga tunnus.

**Sündmus. Klassikaline ja geomeetriline tõenäosus**

Selle vastus on:

Täringuviske abil määratud 6 elementaarsündmuse abil on defineeritud:

- 6 ühest elementaarsündmusest koosnevat sündmust;
- 15 kahest elementaarsündmusest koosnevat sündmust;
- 20 kolmest elementaarsündmusest koosnevat sündmust,
- 15 neljast elementaarsündmusest koosnevat sündmust;
- 6 viiest elementaarsündmusest koosnevat sündmust;
- 1 kindel sündmus, mis sisaldab kõiki kuut elementaarsündmust;
- 1 võimatu sündmus, mis ei sisalda ühtki elementaarsündmust.

Seega kokku 64 =26 erinevat sündmust.

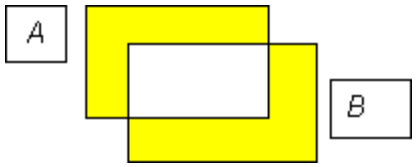
## Sündmus. Klassikaline ja geomeetriline tõenäosus

### SÜNDMUSE MÕISTE

1. Sellepärast, et kindel sündmus sisaldab kõiki võimalikke katsetulemusi.
2. Sündmuse *A* vastandsündmus on kindla sündmuse ja sündmuse *A* vahe.
3. Kui sündmused *A* ja *B* on võrdsed (sisaldavad samu elementaarsündmusi), siis kehtivad võrdused $A \cup B = A \cap B = A = B$
4. Liidetavad kuuluvad summasse, korrutis kuulub teguritesse. $A \subset A \cup B$ ja $B \subset A \cup B$ ; $A \cap B \subset A$ ja $A \cap B \subset B$;

Vahe $A \setminus B$ kuulub esimesesse sündmusesse *A* (vähendatavasse), kuid ei kuulu teise *B* (lahutatavasse).

5. Sümmeetriline vahe: $(A \setminus B) \cup (B \setminus A)$




joonisel kollasega märgitud osa, kus A ja B on lõikuvad ristkülikud.

6. Need on $A \setminus B$, $B \setminus A$ ja $A \cap B$

## Sündmus. Klassikaline ja geomeetriline tõenäosus

### KLASSIKALINE TÕENÄOSUSE MÕISTE

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

2. $A \subset B \Rightarrow P(A) \leq P(B)$.

3. $P(A) \leq P(A \cup B)$, $P(A) \geq P(A / B)$.

4. $P(A \cup B \cup C)= P(A \cup B) + P(C)-P((A \cup B) \cap C)=$
   $P(A)+ P(B)+ P(C)- P(A \cap B)- P(A \cap C)- P(B \cap C)+ P(A \cap B \cap C)$.

5. Niisugustel väidetel ei ole sisulist mõtet.

**Sündmus. Klassikaline ja geomeetriline tõenäosus**

**GEOMEETRILINE TÕENÄOSUS**

1. $0 \le P(A \setminus B) \le P(A) = 1/5$
   $0 \le P(B \setminus A) \le P(B) = 2/5$

2. Võimatu sündmus on see, kui katse tulemusena ei valita (tabata) ühtki vaadeldava tõenäosusruumi (märklaua) punkti.

3. Sündmuse tõenäosus on nii väike, et praktikas seda ei toimu kunagi.

4. Ei saa.

5. $0{,}8 \le P(A \cup B) \le 1$
   $0 \le P(A \setminus B) \le 0{,}2$
   $0{,}2 \le P(B \setminus A) \le 0{,}4$

## Statistiline tõenäosus ja suurte arvude seadus. Sündmuste sõltuvus

## STATISTILINE TÕENÄOSUS

1.Katseseeria pikendamisel ühe katse võrra muutuvad kõigi sündmuste tõenäosused, välja arvatud need, mille väärtus oli enne ja pärast uut katset kas 0 või 1. Tõepoolest, kui lisanduval katsel sündmus *A* ei toimu, kehtib võrratus:

$$\frac{k}{n+1} < \frac{k}{n},$$

seega tõenäosus väheneb, kui *k* erineb nullist, ning kui sündmus *A* toimub, siis

$$\frac{k+1}{n+1} > \frac{k}{n},$$

st et tõenäosus suureneb, kui *k* ja *n* ei ole võrdsed.

2.Kõige suurem võimalik erinevus endise ja uue tõenäosuse vahel on *k*-nda katse korral $1/k$.

3.See on võimalik erandliku katseseeria korral, kus sündmus *A* ei toimunud esimese katse korral, kuid toimub kõigi järgnevate katsete korral. Siis on statistiline tõenäosus *k*-ndal katsel $k/(k+1)$, ning suureneb iga katsega.

4.Kui eelmises ülesandes kirjeldatud katse korral on sündmuse *A* tõenäosuse väärtus 1 (näiteks määratud geomeetrilise tõenäosusena), siis on *k*-ndal katsel tõenäosuse ja suhtelise sageduse erinevus $1/k$ ja see jada väheneb järjest. Tegemist on aga erandliku olukorraga, üldiselt ei vähene tõenäosuse ja suhtelise sageduse erinevus monotoonselt.

5.Sel juhul tõenäosuse liitmise lause ei kehti.

**Statistiline tõenäosus ja suurte arvude seadus. Sündmuste sõltuvus**

**TINGLIK TÕENÄOSUS JA SÜNDMUSTE SÕLTUVUS**

1.Üksteist välistavad sündmused ei ole sõltumatud. Vaatleme nullist erineva tõenäosusega sündmusi *A* ja *B*. Kui nad on välistavad, siis on nende korrutis võimatu, st

$$P(A \cap B) = 0.$$

Siis aga saab võrdus (5) kehtida ainult siis, kui vähemalt üks sündmustest oleks null-tõenäosusega.

2.Vaatleme olukorda, kus

$$A \subset B.$$

Siis kehtib ilmne võrdus

$$A \cap B = A$$

Kasutame seda tinglike tõenäosuste arvutamiseks:

- $P(A \mid B) = P(A \cap B) / P(B) = P(A) / P(B) > P(A).$
- $P(B \mid A) = P(B \cap A) / P(A) = P(A) / P(A) = 1.$

Viimane seos on hästi mõistetav, sest sisaldusseos tähendabki seda, et ühe sündmuse toimumisest järeldub (kindlasti) teise sündmuse toimumine, seega ongi vastav tinglik tõenäosus võrdne ühega.

3.Kui üks sündmus järeldub teisest, on sündmused sõltuvad. Tõepoolest, siis ei kehti sõltumatuse tingimus, sest korrutise tõenäosus on võrdne ühe sündmuse tõenäosusega; järelikultv seos (5) saab kehtida üksnes triviaalsel erijuhul, kui teise sündmuse tõenäosus on 1.

4.Et katsetulemused (elementaarsündmused) on üksteist välistavad, siis kehtib ülesande 1. vastus: nad ei ole sõltumatud. Nende sõltuvus seisneb selles, et kui üks esineb, ei saa ükski teine esineda.

5.Üksteist välistavate sündmuste korrutis on võimatu sündmus.

6.Et sündmus ja tema täiend- ehk vastandsündmus on üksteist välistavad, siis kehtib ülesande 1. vastus: nad ei ole sõltumatud. Nende sõltuvus seisneb selles, et kui üks esineb, ei saa teine esineda.

## Statistiline tõenäosus ja suurte arvude seadus. Sündmuste sõltuvus

## BAYESI TEOREEM

1.Järeltõenäosuste summa on võrdne ühega, sest ka peale sündmuse $A$ toimumist moodustavad sündmused $H_i$ täissüsteemi.

2.Jah, võib. See juhtub siis, kui sündmuse $A$ tinglik tõenäosus vastava hüpoteesi $H_i$ toimumisel on 0.

3.Sel juhul on sündmus $A$ sõltumatu kõigist sündmustest $H_i$.

**Juhuslik suurus ja vektor. Jaotus ja tema esitused**

**JUHUSLIK SUURUS JA VEKTOR**

1.Olgu juhusliku suuruse väärtused $a$ ja $b$. Siis kolme liidetava summal on järgmised võimalikud väärtused: $3a$, $2a+b$, $a+2b$, $3b$. Nende tõenäosused ei ole võrdsed.

2.Olgu esimese, teise ja kolmanda juhusliku suuruse väärtused vastavalt $a_1$ ja $b_1$, $a_2$ ja $b_2$, ja $a_3$ ja $b_3$. Siis on juhuslike suuruste summa võimalikud väärtused:

- $a1+ a2+ a3$;
- $a1+ a2+ b3$;
- $a1+ b2+ a3$;
- $b1+ a2+ a3$;
- $a1+ b2+ b3$;
- $b1+ a2+ b3$;
- $b1+ b2+ a3$;
- $b1+ b2+ b3$.

Seega on summa võimalike erinevate väärtuste arv 8, kuid võib juhtuda, et mõned väärtused langevad kokku. Näiteks, kui kehtib võrdus $a_2+ b_3 = b_2+ a_3$, siis on teine ja kolmas summa väärtus omavahel võrdsed.

3.Moodustame mündiviske tulemuste tabeli

| Kombinatsiooni nr | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1-sendine | vapp | kiri | vapp | vapp | kiri | kiri | vapp | kiri |
| 5-sendine | vapp | vapp | kiri | vapp | kiri | vapp | kiri | kiri |
| 10-sendine | vapp | vapp | vapp | kiri | vapp | kiri | kiri | kiri |
| X väärtus | 0 | 1 | 5 | 10 | 6 | 11 | 15 | 16 |
| Tõenäosus | 1/8 | 1/8 | 1/8 | 1/8 | 1/8 | 1/8 | 1/8 | 1/8 |

**Juhuslik suurus ja vektor. Jaotus ja tema esitused**

**DISKREETSED JAOTUSSEADUSED**

1.Kui Bernoulli jaotuse parameetriks on $p$, siis on kolme sõltumatu liidetava summa binoomjaotusega $B(3, p)$.

2. $\sum_{k=1}^{\infty} pq^{k-1} = q\sum_{k=0}^{\infty} p^{k} = \frac{q}{1-p} = 1.$

3.Ei saa, nad on alati sõltuvad, sest neid seob ühine lineaarne tingimus

$$\sum_{i=1}^{m} X_i = n.$$

4.Et alati $p > pq^k$, kui $q < 0$ ja $k > 0$, siis on alati esimene väärtus $k = 1$ suurim.

**Juhuslik suurus ja vektor. Jaotus ja tema esitused**

**PIDEVAD JAOTUSSEADUSED**

1. $F(x)=\int_{-\infty}^{x} f(t)dt=\int_{0}^{x}\lambda e^{-\lambda t}dt=-e^{-\lambda t}\Big|_{0}^{x}=1-e^{-\lambda x}$.

2.Lõigu [*a*, *b*] pikenemisel väheneb tihedusfunktsiooni väärtus pöördvõrdeliselt (piirkonnas, kus ta erineb nullist). Nimetatud piirkonda vastavalt pikeneb.

3.Tihedusfunktsioon võrdub nulliga argumendi negatiivsete väärtuste puhul ja läheneb nullile argumendi piiramatul kasvamisel.

4.Diskreetse juhusliku suuruse tõenäosusfunktsioonil ja pideva juhusliku suuruse tihedusfunktsioonil on järgmised sarnased omadused:

•Mõlemad funktsioonid on määravad jaotusfunktsiooni/ on jaotusfunktsiooni poolt üheselt määratud.

•Mõlemad funktsioonid võrduvad nulliga nende argumentide korral, mis ei ole vastava juhusiku suuruse väärtusteks.

•Mõlemad funktsioonid erinevad nullist nende argumentide korral, mis kuuluvad juhusliku suuruse väärtuste hulka.

•Mõlema funktsiooni kaudu saab määrata juhusliku suuruse abil defineeritud sündmuste tõenäosusi; selleks kasutatakse tõenäosusfunktsiooni korral summeerimist, tihedusfunktsiooni korral integreerimist.

•Tõenäosusfunktsiooni summa üle kõikvõimalike väärtuste ja tihedusfunktsiooni integraal (miinus lõpmatusest lõpmatuseni) võrdub ühega.

•Mõlema funktsiooni väärtused üldised kahanevad, kui argument läheneb lõpmatusele või miinus lõpmatusele.

•Tõenäosusfunktsiooni väärtus on suurem niisuguste juhusliku suuruse väärtuste korral, mille esinemistõenäosus on suurem; tihedusfunktsiooni väärtus on suurem piirkonnas, millesse juhusliku suuruse väärtus satub suurema tõenäosusega.

**Juhusliku suuruse jaotusparameetrid**

**JUHUSLIKU SUURUSE ASENDIKARAKTERISTIKUD**

1. Ei ole. Näide: täringuviske tulemused on täisarvud, nende keskväärtus on 3,5.

2. Jah, sest mood on alati juhusliku suuruse väärtus.

3. Kui juhuslik suurus on mittenegatiivne, siis on ka tema minimaalne väärtus mittenegatiivne,

$$\min(x_i) \geq 0.$$

Väide järeldub nüüd vahetult keskväärtuse (mediaani, moodi) monotoonsuse omadusest.

4. Sel juhul on keskväärtus ühtlasi sümmeetriakeskpunktiks ja mediaaniks.

5. Siis kehtib ka võrratus $EX < EY$.

6. Jah, sest lõplikul arvul liidetavatel on alati summa olemas.

7. Jah, sest vastav integraal eksisteerib alati.

8. Ülesande 4 vastusest järeldub, et siis on $EX = 0$.

9.

• Bernoulli jaotuse keskväärtus on $p$ (sündmuse tõenäosus).

• Binoomjaotuse keskväärtus:

$$EX = \sum_{k=1}^{n} k \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} = np \sum_{k=1}^{n-1} \frac{(n-1)!}{(k-1)!(n-k)!} p^{(k-1)} (1-p)^{n-k} = np,$$

sest summamärgi all on tõenäosusfunktsiooni summa (summeeritakse $k' = k-1$ järgi, kusjuures kasutatakse seda, et esimeses summas on esimene liidetav 0), mis tõenäosusfunktsiooni põhiomaduse tõttu võrdub ühega. Sama tulemuseni jõuaksime

ka arvestades, et vastavalt definitsioonile on binoomjaotus $n$ sõltumatu Bernoulli jaotusega juhusliku suuruse summa, seega keskväärtuse aditiivsuse omadusest järeldub, et $EX = np$.

• Poissoni jaotuse keskväärtuse arvutamisel kasutame sarnast võtete binoomjaotuse keskväärtuse arvutamisega, teisendades summat nii, et see annaks tõenäosusfunktsiooni summa:

$$EX = \sum_{k=1}^{\infty} k \frac{e^{-\lambda}\lambda^{k}}{k!} = \lambda \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^{k-1}}{(k-1)!} = \lambda.$$

• Ühtlase jaotuse keskväärtuse leiame integreerimise teel:

Sama tulemuseni võinuksime jõuda ka tõdemusest, et tegemist on sümmeetrilise jaotusega, mille sümmeetria keskpunkt paikneb otspunktide vahelise lõigu keskel.

$$EX = \frac{1}{b-a}\int_{a}^{b} x dx = \frac{x^2}{2(b-a)}\Big|_{a}^{b} = \frac{a+b}{2}.$$

• Eksponentjaotuse keskväärtuse leiame, kasutades ositi integreerimist ja tõsiasja, et

$$xe^{-\lambda x} = 0$$

nii $x$ lähenemisel nullile kui ka lõpmatusele:

$$EX = \int_{0}^{\infty} x\lambda e^{-\lambda x} dx = -xe^{-\lambda x}\Big|_{0}^{\infty} - \int_{0}^{\infty} e^{-\lambda x} dx = \frac{1}{\lambda}.$$

10.

• Ühtlase jaotuse mediaan võrdub keskväärtusega.

• Eksponentjaotuse mediaani leiame tingimusest:

$$F(med) = 0{,}5 \Rightarrow 1 - e^{-\lambda med} = 0{,}5 \Rightarrow e^{\lambda med} = 2 \Rightarrow med = \frac{\ln 2}{\lambda}.$$

11. Bernoulli jaotuse mood on 1, kui $p>0{,}5$ ja 0, kui $p<0{,}5$.

• Binoomjaotusel on enamasti üheselt määratud mood. Kaks võrdset kõrvuti paiknevat moodi on juhul, kui $p=0{,}5$ ja $n$ on paarisarv.

• Poissoni jaotusel võib samuti olla kaks kõrvutipaiknevat moodi, nii on olukord siis, kui parameeter $\lambda$ on täisarvuline; siis on moodiks $\lambda-1$ ja $\lambda$.

• Ühtlasel jaotusel mood puudub.

• Eksponentjaotus on L-jaotus, sellel ei eksisteeri moodi, kuid tihedusfunktsioon kasvab piiramatult nullile lähenemisel.

**Juhusliku suuruse jaotusparameetrid**

**JUHUSLIKU SUURUSE HAJUVUSKARAKTERISTIKUD**

Tõestada see valem!

$$DX = E(X - EX)^2 = E(X^2 - 2EX \times X + (EX)^2) = EX^2 - 2EX \times EX + (EX)^2 = EX^2 - (EX)^2.$$

Tõestamise juures kasutati keskväärtuse omadusi (lineaarsust, aditiivsust ja konstandi keskväärtust).

**Juhusliku suuruse jaotusparameetrid**

**JUHUSLIKU SUURUSE HAJUVUSKARAKTERISTIKUD**

1. Et $D(X + Y) = D(X - Y) = DX + DY$, siis on mõlemal juhul standardhälbeks

$$\sqrt{DX + DY}.$$

2. Leiame jaotuste hajuvuskarakteristikud

• Bernoulli jaotus:

$$DX = E(X-EX)^2 = p(1-EX)^2+(1-p)(0-EX)^2 = p(1-p)^2+(1-p)p^2$$
$$= p(1-p)(1-p+p) = p(1-p);$$
$$\sigma(X) = \sqrt{p(1-p)}.$$
$$V(X) = \sqrt{p(1-p)} / p = \sqrt{\frac{1-p}{p}}.$$

• Binoomjaotus.

Kasutame seda, et binoomjaotus on sõltumatute Bernoulli jaotusega juhuslike suuruste summa, seega

$$\sigma(X) = \sqrt{np(1-p)},$$
$$V(X) = \sqrt{\frac{n(1-p)}{p}},$$
$$DX = np(1-p).$$

• Poissoni jaotus

Kasutame dispersiooni avaldist momentide kaudu, $DX = E(X^2) - (EX)^2$. Esimest järku moment on $\lambda$. Leiame ka teist järku momendi:

$$E(X^2) = \sum_{k=1}^{\infty} k^2 \frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k=1}^{\infty} k(k-1)\frac{e^{-\lambda}\lambda^k}{k!} + \sum_{k=1}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} = \lambda^2 \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^{(k-2)}}{(k-2)!} + \lambda = \lambda^2 + \lambda.$$

Teine liidetav on keskväärtus, esimese puhul saame summamärgi all tõenäosusfunktsiooni summa, kusjuures tuleb taas arvestada, et esimesed liidetavad võrduvad nulliga. Nüüd leiame ka soovitavad hajuvuskarakteristikud:

$$DX = \lambda^2 + \lambda - \lambda^2 = \lambda,$$
$$\sigma(X) = \sqrt{\lambda},$$
$$V(X) = 1/\sqrt{\lambda}.$$
$$EX^2 = \frac{1}{b-a}\int_a^b x^2 dx = \frac{b^3-a^3}{3(b-a)} = \frac{b^2+ab+a^2}{3};$$
$$DX = \frac{b^2+ab+a^2}{3} - \frac{a^2+2ab+b^2}{4} = \frac{(a-b)^2}{12};$$
$$\sigma(X) = \frac{b-a}{2\sqrt{3}};$$
$$V(X) = \frac{b-a}{4\sqrt{3}(a+b)}.$$

• Ühtlane jaotus

3. Lõplik haare on olemas:

• Bernoulli jaotusel, selle suurus on 1;

• Binoomjaotusel, selle suurus on $n$;

• Ühtlasel jaotusel, selle suurus on $b - a$.

• Poissoni jaotuse väärtused pole tõkestatud, sellel puudub lõplik maksimum ja seega ka haare.

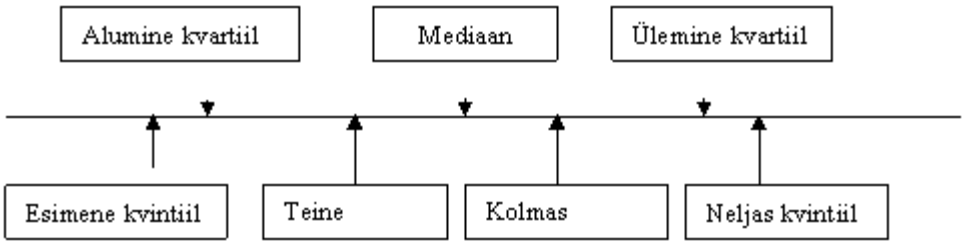
• Samal põhjusel puudub haare ka eksponentjaotusel.

**Juhusliku suuruse jaotusparameetrid**

**JAOTUSE KUJU ISELOOMUSTAVAD KARAKTERISTIKUD**

1.Siis on juhuslik suurus kõdunud konstandiks, kvantiile pole võimalik leida; võib ka öelda, et kõik kvantiilid langevad ühte selle juhusliku suuruse ainsa väärtusega.

2. Kvartiilide ja kvintiilide paiknemine




3.Mittenegatiivsete väärtustega juhusliku suuruse keskväärtus sisaldab nii positiivseid kui ka negatiivseid liikmeid, hajuvuse kordajates kasutatakse ainult positiivseid hälbeid, selletõttu võib suhte nimetaja muutuda väikseks (ka võrdseks nulliga) ja näitaja sisu muutub ning ei ole hästi interpreteeritav.

4.Sümmeetrilise juhusliku suuruse sümmeetriakeskmeks on keskväärtus, mis langeb ühte mediaaniga. Pideva sümmeetrilise juhusliku suuruse puhul paikneb iga $q$-kvantiil sümmeetriliselt oma täiendkvantiili, $(1 - q)$-kvantiili suhtes.

5.Ekstsess ei saa kindlasti olla väiksem kui –3, sest ekstsessi avaldise esimene liige on mittenegatiivne. Detailsem analüüs näitab, et esimene liidetava minimaalne võimalik väärtus on 1, seega ei saa ekstsess olla väiksem kui –2.

6.Bernoulli jaotus on sümmeetriline siis, kui $p = 0{,}5$.

7.Ühtlase jaotuse kvartiilid on:

• alumine kvartiil: $(3a + b)/4$; mediaan: $(a + b)/2$, ülemine kvartiil: $(a + 3b)/4$.

• Ühtlase jaotuse asümmeetriakordaja on 0, sest jaotus on sümmeetriline.

• Arvutame lihtsuse mõttes nullpunkti suhtes sümmeetrilise ühtlase jaotuse U(-$a$, $a$) ektsessi; kuna jaotuse kuju lineaarteisenduse tagajärjel ei muutu, kehtib

esitatud arvutus ka üldjuhul.

$$DX = \frac{1}{2a}\int_{-a}^{a} x^2\,dx = \frac{1}{2a}\,\frac{2a^3}{3} = \frac{a^3}{3}; \quad (DX)^2 = \frac{a^4}{9};$$

$$E(X)4 = \frac{1}{2a}\int_{-a}^{a} x^4\,dx = \frac{a^4}{5};$$

$$e(X) = \frac{a^4 \cdot 9}{5 \cdot a^4} - 3 = -1{,}2.$$

**Juhusliku suuruse jaotusparameetrid**

**TŠEBÕŠEVI VÕRRATUS JA SUURTE ARVUDE SEADUSE TÕESTUS**

1. Jah, on küll, kuigi praktiliselt esineb äärmiselt harva (väga väikse tõenäosusega).

2. Põhimõtteliselt võib ka lühikese katseseeria korral esineda kuitahes suuri hälbeid. Kui aga kasutada tõenäosuste hindamiseks Tšebõševi võrratust, siis on näha, et tõenäosus saada hälve, mis on suuremad kui 10-kordne standardhälve, on alla 1/100, seega on üsna vähe tõenäoline nii suurt hälvet saada 10-katselise seeria korral.

## Normaaljaotus ja tsentraalne piirteoreem. Lineaarne korrelatsioonikordaja

### NORMAALJAOTUS

1.Standardiseeritud normaaljaotuse kvartiilid on määratud tingimustega $F(q_1) = 0{,}25$, $F(q_2) = 0{,}5$, $F(q_3) = 0{,}75$. Vastavad väärtused on: $q_1 = -0{,}6745$, $q_2 = 0$ ja $q_3 = 0{,}6745$.

2.Otsitavad kvantiilid on –1,960 ja 1.960, nende vahele jääb juhuslik suurus tõenäosusega 0,975–0,025 = 0,95.

3.$P(X > 1) = 0{,}159$, $P(X < -1) = 0{,}159$.

# Normaaljaotus ja tsentraalne piirteoreem. Lineaarne korrelatsioonikordaja

## TSENTRAALSED PIIRTEOREEMID

1.Jah on. Siin on $X_i$ Bernoulli jaotusega, $n$ sõltumatu liidetava summa on binoomjaotusega, sellest lahutatakse tema keskväärtus ja nimetajas on binoomjaotuse, so summa standardhälve.

2.Vigu võib käsitleda kui paljude ühesuguse jaotusega sõltumatute faktorite mõjude summeerumise resultaati.

3.Jah sõltub. Normaaljaotuseks koondumine toimub kõige kiiremini siis, kui $p$ = 0,5, sest siis on binoomjaotus sümmeetriline. Mida väiksem või suurem (vastavalt 0 või 1 lähedasem) on $p$, seda ebasümmeetrilisem on binoomjaotus ja koondumine sümmeetriliseks normaaljaotuseks toimub aeglasemalt.

4.Ei, selle puhul ei käsitleta tsentreeritud juhuslikke suurusi.

5. Normaaljaotus sobib lähendiks siis, kui $p$ on 0,5 lähedal, Poissoni jaotus aga siis, kui $p$ on 0 lähedal.

**Normaaljaotus ja tsentraalne piirteoreem. Lineaarne korrelatsioonikordaja**

**NORMAALJAOTUSEGA JUHUSLIK VEKTOR. KORRELATSIOONIKORDAJA**

1. Kui $r = 0$, siis on kahemõõtmelise normaaljaotuse tihedusfunktsiooni avaldis esitatav korrutisena:

$$f(x,y)=\frac{1}{2\pi}e^{-\frac{(x^2+y^2)}{2}}=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}},$$

millest järeldubki komponentide sõltumatus.

2. Korrelatsioonikordaja absoluutväärtus ei muutu, kuid märk muutub, kui teisendatud tunnuse kordaja on negatiivne.

3. $r(X, Y) = r(X, -X) = -1$.

4. $Y$ jaotus on samuti standardiseeritud normaaljaotus.

5. $Y$ jaotus on $N(-\mu, s)$.

6. $2X \sim N(2\mu, 2\sigma)$;
$X+Y \sim N(\mu_1+\mu_2, \sqrt{\sigma_1^2+\sigma_2^2})$;
$2X-3Y \sim N(2\mu_1-3\mu_2, \sqrt{4\sigma_1^2+9\sigma_2^2})$.

**Üldkogum ja valim. Hindamine**

**ÜLDKOGUM JA VALIM**

1.Üldiselt samuti Bernoulli jaotusega, kusjuures on ka võimalus, et valim on konstantse jaotusega (kõik objektid omavahel võrdsed).

2.Ei saa, valim on alati diskreetne, normaaljaotus on pidev. Küll aga saab olla valimi jaotus normaaljaotusele lähedane.

3.Üldiselt erinevad. Võimalik on ka see, et nad on võrdse, kuid see on väga väikse tõenäosusega sündmus.

4.Toimub üks kahest võimalusest:

•Lisandub üks uus punkt tõenäosusega $1/(n+1)$ ja kõigi teiste punktide tõenäosus väheneb $(n+1)/n$ korda.

•Ühe punkti tõenäosus suureneb: $k/n$ asemele tuleb $(k+1)/(n+1)$ ja kõigi teiste punktide tõenäosus väheneb $(n+1)/n$ korda.

**Üldkogum ja valim. Hindamine**

**PUNKTIHINNANG**

1. Võrdse mahu ja ühesuguse eeskirjaga moodustatud valimid on statistika seisukohast samaväärsed. Suurema mahuga valim annab põhimõtteliselt täpsema hinnangu.

2. Juhuslikud on hinnang, valimkeskmine ja valimi dispersioon (mis on samuti hinnangud). Hinnangu nihe ja efektiivsus ei ole juhuslikud, sest on defineeritud hinnangu keskväärtuse kaudu.

3. Juhusliku suuruse standardhälve on suurem, sest standardviga ehk keskmise standardhälve on määratud valemiga

$$m=\sigma/\sqrt{n},$$

kusjuures valimi maht $n$ on suurem kui 1.

4. Süstemaatilist viga iseloomustab hinnangu nihe.

5. Kui eeldada, et keskmise hinnang ja dispersiooni hinnang jäävad samaks, siis on tarvis suhtelise vea kahekordseks vähendamiseks valimi mahtu suurendada $2^2 = 4$ korda.

6. Kõigi hinnangute täpsust saab suurendada katsete arvu suurendamisel, sest kõik hinnangud sisaldavad juhuslikku viga, mis sõltub katsete arvust. Nihkega hinnang sisaldab aga lisaks nihet, mida katsete arvu suurendamisega ei ole võimalik kõrvaldada.

**Üldkogum ja valim. Hindamine**

**VAHEMIKHINNANG**

1. Kui muud näitajad jäävad samaks, siis usaldustõenäosuse $1-\alpha$ suurendamisel (ühele lähenemisel) usalduspiirkond laieneb.

2. Mida suurem on lähtetunnuse hajuvus, seda laiemad on üldiselt usalduspiirid.

3. Suurema hajuvuse korral on ka hajuvuse hinnangud suuremad; suuremate hajuvuse hinnangute korral on usalduspiirid laiemad.

4. Valimi mahtu tuleks suurendada $2^2 = 4$ korda, kui oletada, et hajuvuse hinnang jääb samaks.

5. Punkthinnangu suurenes suurenevad ka mõlemad usalduspiirid. Kui selle juures hajuvus ei suurene, jääb usaldusvahemiku ulatus (pikkus) samaks.

## STATISTILISTE HÜPOTEESIDE KONTROLLIMINE

## STATISTILISE HÜPOTEESIDE KONTROLLIMISE TEOORIA PÕHIMÕISTED

1.Ei saa, sest siis vahetuksid esimest ja teist liiki vead, mis aga pole sümmeetrilised.

2.*t*-jaotuse abil saab konstrueerida normaaljaotusega juhusliku suuruse keskväärtusele usalduspiire ja kontrollida hüpoteese normaaljaotuse keskväärtuste kohta. Vastav metoodika töötab ka siis, kui tunnuse teoreetiline jaotus mõnevõrra erineb normaaljaotusest.

3.*t*-jaotuse parameetriks on vabadusastmete arv *f*, mis avaldub valimi mahu kaudu: $f = n - 1$. Kui vabadusastmete arv läheneb lõpmatusele, läheneb *t*-jaotus standardiseeritud normaaljaotusele. Seetõttu, kui valimi maht on üle 100, *t*-jaotus peaaegu ei sõltu enam valimi mahust.

4.Usaldusnivoo on tõenäosus, seetõttu ta saab muutuda vahemikus [0, 1]. Praktilist huvi pakuvad väärtused, mis on 1 lähedal, näiteks 0,9, 0,95, 0,99 ja 0,999. Ka olulisuse nivoo on tõenäosus, seetõttu ta saab muutuda vahemikus [0, 1]. Praktilist huvi pakuvad väärtused, mis on 0 lähedal, näiteks 0,1, 0,05, 0,01 ja 0,001.

5.Olulisuse nivoo $\alpha$ (lubatud esimest liiki vea tõenäosus) on tavaliselt väike, usaldusnivoo aga määratakse kui $1-\alpha$ ja seetõttu on ta enamasti suhteliselt suur, st lähedane arvule 1.

6.Sama ülesande puhul valitakse tavaliselt olulisuse nivoo ja usaldusnivoo nii, et nende summa on 1, st kui olulisuse nivoo on a, siis usaldusnivoo on $1 - \alpha$.

## STATISTILISTE HÜPOTEESIDE KONTROLLIMINE

## HÜPOTEESIDE KONTROLLIMINE ÜLDKOGUMI KESKVÄÄRTUSE KOHTA

1.Tuleb võtta vastu nullhüpotees, pole mõtet teha arvutusi. Tulemus ei sõltu olulisuse nivoost (kui see on mõistlik, st väiksem kui 0,5).

2.Teha täiendavaid katseid, et suurendada valimi mahtu.

3.Suurendada olulisuse nivood, kui see ei aita, siis jääda nullhüpoteesi juurde.

4.Kui pole eelinfot juhuslike suuruste hajuvuse vahekorra kohta või on dispersioonid ligikaudu võrdsed, siis jagada võrdselt, sest sel juhul on tegur $(n_1 \ n_2)/(n_1+n_2)$ maksimaalne. Kui hajuvused erinevad palju, tuleb suurema hajuvusega üldkogumist võtta suurem valim.

5.Normaaljaotust.

6.Normaaljaotust.

## STATISTILISTE HÜPOTEESIDE KONTROLLIMINE

## MÕNINGATE MUUDE STATISTILISTE HÜPOTEESIDE KONTROLLIMINE

1.See on nullhüpotees, mida ei saa tõestada.

2.See on võimalik. Selleks on vaja:

•Määrata kahe juhusliku suuruse vahet $D$ mõõtev statistik $\delta$;

•Leida statistiku $\delta$ jaotus;

•Kontrollida hüpoteesi $H_1$: $D < d$ täidetust soovitaval olulisuse nivool.

3.Tegemist on valimiga, mis võib kuuluda niihästi normaal-, Possoni kui ka ühtlase jaotusega üldkogumisse. Niisugune olukord tekib tavaliselt väikest valimite või eriti väikese olulisuse nivoo korral.

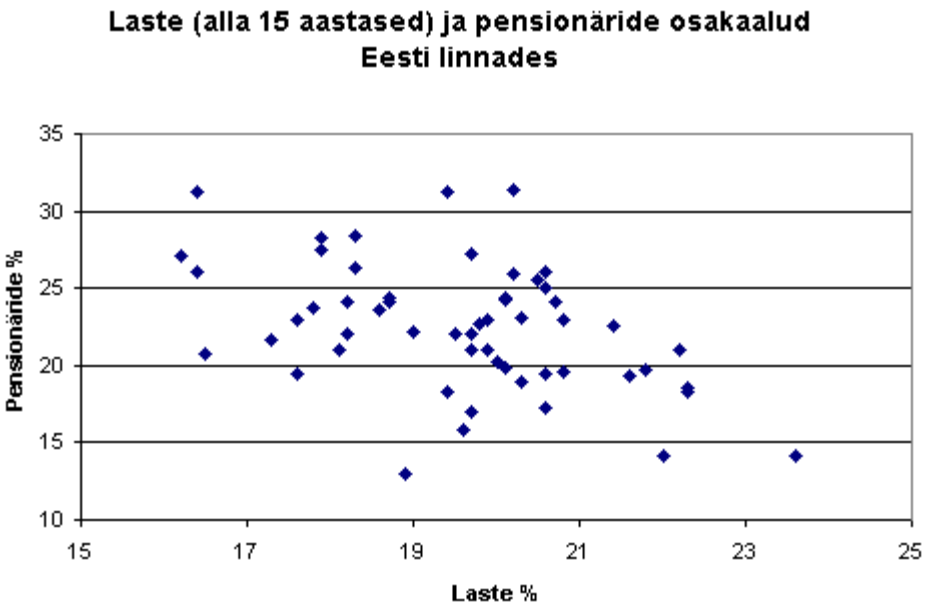
4.Nii tulebki tulemust tõlgendada: I ja III katseseeria erinevad omavahel, II ei erine oluliselt kummastki (on kas vahepealne, väga väike või tugevasti hajuv).

**Statistiline sõltuvus ja statistiline mudel**

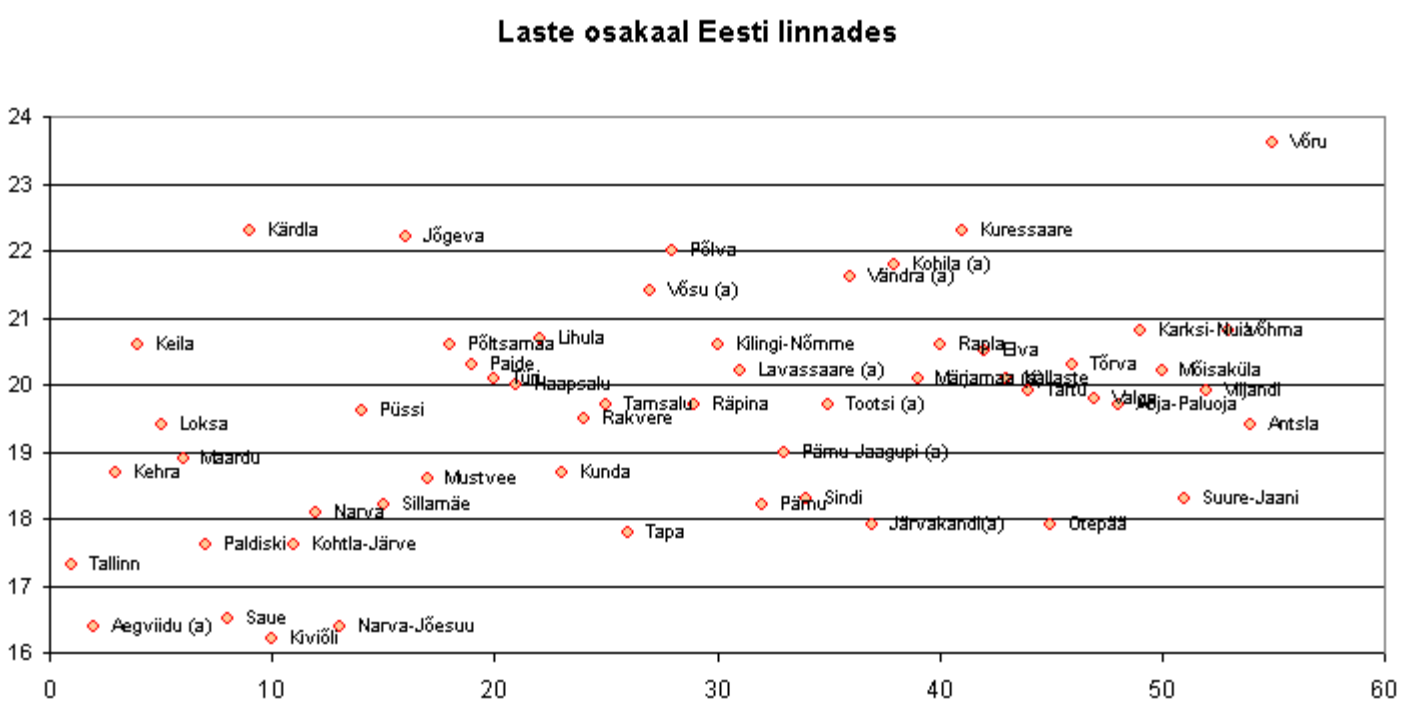
**LINEAARNE MUDEL**

**Hajuvusdiagramm ehk korrelatsiooniväli**

Kahe pideva arvtunnuse *X* ja *Y* empiirilist ühisjaotust ilmestab hästi graafik, mida nimetatakse hajuvusdiagrammiks ehk korrelatsiooniväljaks.




Sellele graafikule on kantud kõik valimi punktid oma *X*- ja *Y*-koordinaatidega, ning lähtetunnuste pidevuse tõttu on üliväike tõenäosus selleks, et mõned punktid kattuksid. Praktiliselt seda siiski juhtub piiratud mõõtmis- ja esitustäpsuse tõttu, vt lisatud joonised.




Lisatud joonisel on horisontaalteljel linna/ alevi järjekorranumber, vertikaalteljel aga alla 15-aastaste laste osakaal linnaelanikkonnas (protsentides). Sellel hajuvusdiagrammil on iga punkt identifitseeritud, tavaliselt pole seda võimalik teha punktide suure arvu tõttu, vt eelmist joonist.

**Lineaarne mudel**

Kahe arvtunnuse vahelist statistilist sõltuvust saab esitada mudeli abil. Mudeli puhul valitakse üks tunnustest ***prognoositavaks*** ehk ***funktsioontunnuseks*** ja teine argumenttunnuseks (vahel öeldakse ka seletav tunnus). Funktsioontunnust tähistatakse tähega *Y*, argumenttunnus on *X*.

Mudeli otsimist alustatakse tavaliselt lihtsaimast võimalikust mudelist, see on lineaarne mudel

$$Y = a + bX + \varepsilon,$$

mida tasandil kujutab sirge, vt alljärgnev joonis. Siin tähistab $\varepsilon$ *juhuslikku viga*, mis on samuti nagu *X* ja *Y* juhuslik suurus, mis iga vaatluse korral omandab üldiselt erineva väärtuse.

**Mudeli parameetrite määramine vähimruutude printsiibil**

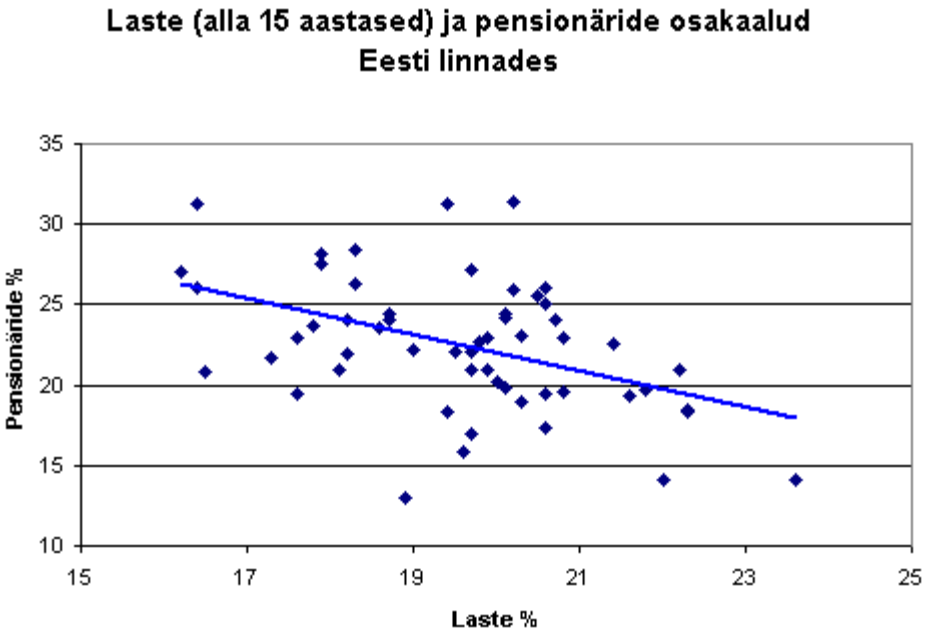
Selleks, et määrata tasandil sirge $Y = a + bX$, on tarvis valimi andmete põhjal arvutada – ***hinnata*** – sirge võrrandi kordajad *a* ja *b*, so ***mudeli parameetrid***. Põhimõtteliselt on mudeli parameetrite hindamiseks terve rida meetodeid, tutvume siin vähimruutude meetodiga, mille sisuks on määrata mudeli parameetrid nii, et juhusliku vea ruutude summa oleks antud valimi korral minimaalne:

$$\sum_{i=1}^{n} (y_i - (a + bx_i))^2 \Rightarrow min.$$

Selle saavutamiseks tuleb lahendada ekstreemumülesanne, mis määrab nn

normaalvõrrandite süsteemi:

$$\begin{cases} \frac{\partial}{\partial a}\sum_{i=1}^{n}(y_i-(a+bx_i))^2=0 \\ \frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i-(a+bx_i))^2=0. \end{cases}$$




Peale tuletiste arvutamist saame võrrandisüsteemi parameetrite *a* ja *b* määramiseks:

$$\begin{cases} an+b\sum_{i=1}^{n}x_i=\sum^{n}y_i \\ a\sum_{i=1}^{n}x_i+b\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}x_iy_i. \end{cases}$$

Selle võrrandisüsteemi lahendamisel saame regressioonikordajale *b* ja vabaliikmele $\alpha$ järgmised hinnangud, mida me tähistame "katusega", märkimaks, et tegemist on valimi põhjal arvutatud statistikutega.

$$\hat{b}=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x^2-(\sum_{i=1}^{n}x_i)^2}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x^2}$$

$$\hat{a}=\bar{y}-\hat{b}\bar{x}.$$

**Prognoosid ja prognoosivead**

Iga valimi punkti jaoks saab arvutada tema prognoosi leitud mudeli põhjal ja samuti prognoosivea:

$$\hat{y}_i=\hat{a}+\hat{b}x_i,\ \varepsilon_i=y_i-\hat{y}_i.$$

Prognoosiviga avaldub tunnuste keskmiste ja regressioonikordaja hinnangu kaudu kujul:

$$\varepsilon_i=(y_i-\bar{y})-\hat{b}(x_i-\bar{x}).$$

Siit järeldub, et prognoosivigade keskmine üle kõigi valimi punktide võrdne nulliga, st prognoos on keskmiselt õige, ei sisalda süstemaatilist viga. Vabaliikme avaldisest järeldub, et ka valimkeskmise prognoos ei sisalda prognoosiviga.

Prognoosiviga on juhuslik suurus, selle hajuvust mõõdab valimi põhjal leitud prognoosivigade ruutude summa:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - 2\hat{b} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) + \hat{b}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Arvutame kahe viimase liidetava avaldised:

$$-2\hat{b}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = -2\hat{b}n \cdot \hat{cov}(X,Y); \quad \hat{b}^2 \sum_{i=1}^{n} (x_i - x)^2 = \hat{b}^2 n s_x^2 = \hat{b}n \cdot \hat{cov}(X,Y).$$

Seega saame prognoosivigade hajuvuse hindamiseks suuruse

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = n s_y^2 - n \frac{(\hat{cov}(X,Y))^2}{s_x^2},$$

kus $\hat{cov}(X,Y)$ tähistab tunnuste *X* ja *Y* kovariatsiooni $E(X–EX)(Y–EY)$ hinnangut ja nende korrelatsioonikordaja hinnang avaldub kovariatsiooni kaudu:

$$\hat{r} = \frac{\hat{cov}(X,Y)}{s_x s_y}.$$

Seda arvestades saame prognoosivea dispersioonile hinnangu

$$s_\varepsilon^2 = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 = s_y^2(1 - \hat{r}^2).$$

Siit on näha lineaarse korrelatsioonikordaja tõlgendus: korrelatsioonikordaja ruut näitab, kui suure osa funktsioontunnuse dispersioonist kirjeldab mudel. Regressioonikordaja avaldub korrelatsioonikordaja kaudu, see võimaldab hinnata regressioonimudeli parameetreid ka jaotuskarakteristikute põhjal.

$$\hat{b} = \frac{s_x \hat{r}}{s_y}.$$

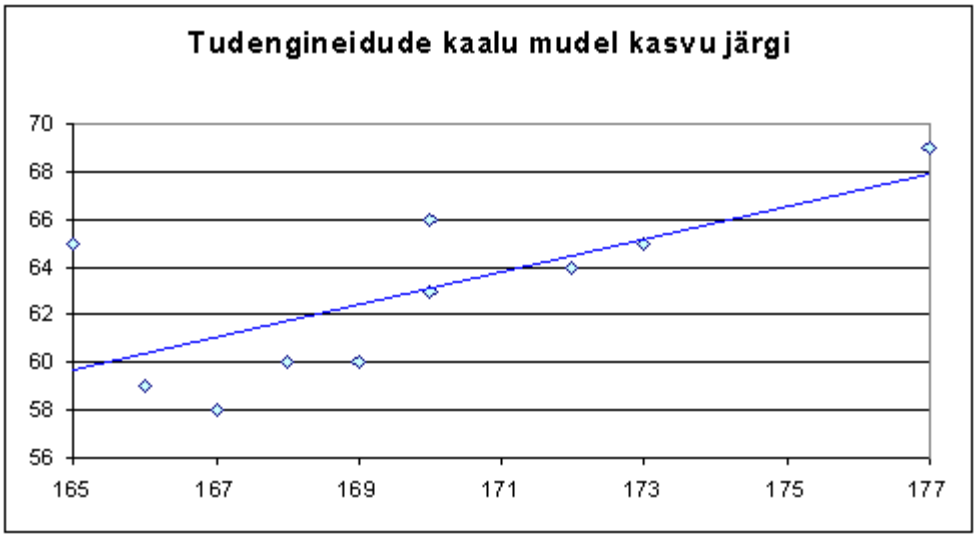
***Näide.*** Olgu antud kümne tudengineiu pikkused ja kaalud alljärgnevas tabelis, mida illustreerib ka lisatud graafik.

| Jrk nr | Pikkus | Kaal | Jrk nr | Pikkus | Kaal |
|---|---|---|---|---|---|
| 1 | 173 | 65 | 6 | 167 | 58 |
| 2 | 168 | 60 | 7 | 166 | 59 |
| 3 | 165 | 65 | 8 | 169 | 60 |
| 4 | 177 | 69 | 9 | 172 | 64 |
| 5 | 170 | 63 | 10 | 170 | 66 |

Leiame tunnuste põhikarakteristikud

| | Kasv | Kaal |
|---|---|---|
| keskmine | 169,7 | 62,9 |
| standardhälve | 3,592 | 3,542 |

Leiame kaalu lineaarse mudel sõltuvalt kasvust




| | |
|---|---|
| r | 0,696 |
| $r^2$ | 0,485 |
| $1-r^2$ | 0,515 |
| $\sqrt{(n-2)}$ | 2,828 |
| t | 3,82 |
| p (ühep) | 0,003 |

Leiame kõigepealt korrelatsioonikordaja ja kontrollime selle olulisust, kasutades selleks ühepoolset hüpoteesi (sisulisest kaalutlusest lähtudes on selge, et korrelatsioon saab olla ainult positiivne). Selgub, et seos on statistiliselt oluline, seega on mudeli parameetrite hindamine hoolimata väikesest valimist mõttekas. Mudeli parameetrite hindamisel saame kasutada juba leitud valimi arvkarakteristikuid.

$$b = \frac{\hat{r}s_y}{s_x} = (0{,}696 \cdot 3{,}542)/3{,}592 = 0{,}686;$$

$$a = 62{,}9 - 0{,}686 \cdot 169{,}7 = -53{,}6.$$

| Jrk nr | Pikkus | Kaal | Mudelkaal | Prognoosiviga |
|---|---|---|---|---|
| 1 | 173 | 65 | 65,165 | -0,165 |
| 2 | 168 | 60 | 61,733 | -1,733 |
| 3 | 165 | 65 | 59,674 | 5,326 |

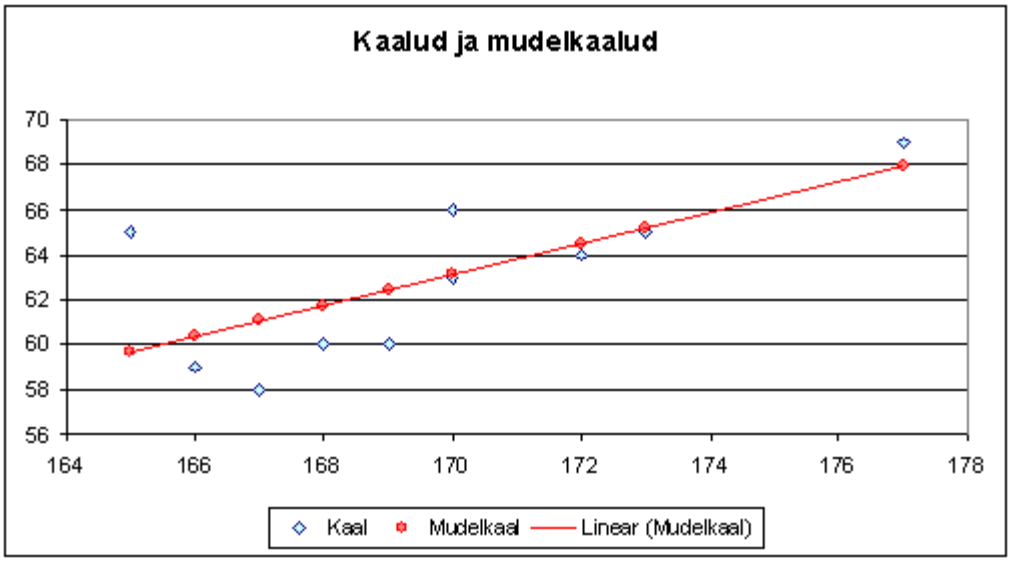
| | | | | |
|---|---|---|---|---|
| 4 | 177 | 69 | 67,911 | 1,089 |
| 5 | 170 | 63 | 63,106 | -0,106 |
| 6 | 167 | 58 | 61,047 | -3,047 |
| 7 | 166 | 59 | 60,360 | -1,360 |
| 8 | 169 | 60 | 62,419 | -2,419 |
| 9 | 172 | 64 | 64,479 | -0,479 |
| 10 | 170 | 66 | 63,106 | 2,894 |

otsitav mudel on:

**Kaal = 0,686•kasv –53,6.**

Leiame mudeli järgi ka vaadeldud neiude kaaluprognoosid (mudelkaalud antud kogumi suhtes) ja prognoosivead.

Viie kilogrammi võrra mudelkaalu ületava, samuti kahe-kolme kilogrammi võrra allpool mudelkaalu oleva neiu prognoosivead on varjutatud.




**Järelemõtlemiseks**

1.Missugused näites esitatud neidudest on ülejäänutega võrreldes suhteliselt raskemad ja missugused kergemad?

2.Kui ühe neiu kasv ja kaal võrdub täpselt valimi keskmiste näitajatega, kus asub sellele vastav punkt graafikul regressioonisirge suhtes.

3.Kuidas paikneks hajuvusdiagrammil konstantsele mudelile vastav regressioonisirge?

4.Kuidas näeks välja hajuvusdiagramm siis, kui *X* ja *Y* vahel on täielik lineaarne sõltuvus?

5.Kuidas muutuks mudeli graafik siis, kui regressioonikordaja korrutada –1-ga?

6.Kuidas muudab regressioonisirge graafikut vabaliikme muutmine?

7.Mis juhtub regressioonisirgega siis, kui vahetada argument- ja funktsioontunnus?

8.Kuidas muudaks regressioonisirge paiknemist üks teistest väga oluliselt erinev punkt (näiteks selline punkt, millel nii *X* kui *Y* väärtus on 10 korda suuremad kui

ülejäänutel?

**Statistiline sõltuvus ja statistiline mudel**

**MUDELI STATISTILINE OLULISUS**

**Lineaarne mudel ja konstantne mudel**

Iga mudeli puhul tuleb teha selgeks, kas mudel on üldistatav üldkogumile, ehk kas ta on statistiliselt oluline. Selleks võrreldakse mudelit konstantse mudeliga, mis ei sõltu üldse argumenttunnusest (ja mis selle tõttu ei ole tavalises mõttes statistiline mudel). Konstantne mudel vaadeldava tunnuse $Y$ jaoks on alati kujul

$$Y = EY,$$

sest keskväärtus on definitsiooni järgi tunnuse väärtuste parim lähend vähimruutude mõttes. Valimi põhjal tähendaks see, et konstantse mudeli hinnanguks on valimkeskmine,

$$Y = \bar{y}.$$

**Mudeli olulisuse mõiste**

Ühe argumendiga lineaarse regressioonimudeli olulisuse kindlakstegemiseks tuleb kontrollida statistiliste hüpoteeside paari:

$H_0$: Üldkogumis on parimaks mudeliks konstantne mudel, st $b = 0$

$H_1$: Üldkogumis leidub lineaarne mudel, mis on parem kui konstantne mudel, st $b \neq 0$.

Niipea, kui asume statistiliste hüpoteeside kontrollimisele, tuleb teha tunnuse $Y$ jaotuse kohta täiendavaid eeldusi.

•Eeldame, et $Y$ on normaaljaotusega;

•Eeldame, et prognoosivigade jaotus ei sõltu $X$ ja $Y$ väärtustest, seega ka prognoosivigade dispersioon on konstantne $\sigma^2$.

Igasuguse ülaltoodud tingimusi rahuldava mudeli olulisuse kontrollimiseks on olemas standardne metoodika, mille oluliseks osaks on ***dispersioonanalüüsi tabeli*** koostamine ja ***F-statistiku kasutamine***.

***F-jaotus***

Olgu juhuslikud suurused $X_1$ ja $X_2$ sõltumatud ning vastavalt $\chi^2$-jaotusega vabadusastmete arvudega $f_1$ ja $f_2$. Siis öeldakse et juhuslik suurus

$$F(f_1, f_2) = \frac{f_2 \cdot X_1}{f_1 \cdot X_2}$$

on $F$-jaotusega vabadusastmete arvudega $f_1$ ja $f_2$. F-jaotus on üks statistika

põhijaotusi, ning tema olulisemad täiendkvantiilid on tabuleeritud, samuti arvutavad statistikaprogrammid välja *F*-jaotusega statistikute olulisuse tõenäosused.

**F-jaotuse kasutamine mudeli olulisuse kontrollimisel**

Funktsioontunnuse *Y* hajuvust mõõdab ruutude summa $S^2$, mida kasutatakse ka dispersiooni hindamisel

$$S^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 \cdot s_y^2 = S^2/(n-1).$$

Ruutude summa $S^2$ on esitatav kahe liidetava summana:

$$S^2 = S_1^2 + S_0^2,\ S_1^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2,\ S_0^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Neist esimene iseloomustab prognoosi erinevust valimkeskmisest, teine üksikväärtuste erinevusi mudeliga ennustatud väärtustest. Kui *Y* on normaaljaotusega ja null-hüpotees on õige, st et parim mudel on konstantne, siis on leitud summad sõltumatud ja võrdelised hii-ruut jaotusega, kusjuures sulgudes on näidatud vabadusastmete arvud:

$$S^2/\sigma_y^2 \sim \chi^2(n-1),\ S_1^2/\sigma_y^2 \sim \chi^2(1),\ S^2/\sigma_y^2 \sim \chi^2(n-2).$$

Seega on suhe

$$F = \frac{S_1^2}{S_0^2} \cdot \frac{n-2}{1}.$$

*F*-jaotusega vabadusastmete arvudega 1 ja *n*–2 (esimesel kohal on lugejas asuva avaldise vabadusastmete arv). Vabadusastmete arv määratakse üldiselt kui sõltumatute vaatluste arv – kasutatud lineaarsete seoste arv. Siit on lihtne järeldada $S^2$ ja ${S_0}^2$ vabadusastmete arv, ${S^2}_1$ vabadusastmete arv võrdub eelmiste avaldiste vabadusastmete arvude vahega.

**Dispersioonanalüüsi tabel**

Mudeli olulisuse kontrollimise standardne metoodika tugineb nn dispersioonanalüüsi tabelil, mida esitab ka enamus rakendusstatistika programme. See tabel on ühesuguse kujuga nii keerukate kui suhteliselt lihtsate mudelite korral. Enamasti on tabelis 6 veergu, kusjuures teise ja kolmanda veeru puhul sisaldab viimane rida eelnevate ridade summat. Enamasti vastab tabelis igale mudeli liikmele rida, mis võimaldab kontrollida selle liikme olulisust mudelis.

| Varieeruvuse allikas | Ruutude summa | Vabadus-astmete arv | Keskruut | F-statistik | Olulisuse tõenäosus |
|---|---|---|---|---|---|
| Mudel | ${S^2}_1$ | 1 | ${S^2}_1$ | ${S^2}_1(n-1)/{S^2}_0$ | p |
| Viga | ${S^2}_0$ | n–2 | $s^2={S^2}_0/(n-$ | | |

| | | | 2) | | |
|---|---|---|---|---|---|
| Summaarne hajuvus | $S^2$ | n–1 | | | |

Neljas veerg tabelis sisaldab teises ja kolmandas veerus antud arvude jagatist. Otsuse vastuvõtmisel tuginetakse olulisuse tõenäosusele – kui see on väiksem kui etteantud olulisuse nivoo, siis kummutatakse nullhüpotees, mille sisuks on vastava mudeli liikme mitteolulisus. Nullhüpotees kummutatakse siis, kui *F*-statistiku väärtus on suur, ning see näitab, et vastava liikme poolt kirjeldatud funktsioontunnuse hajuvuse osa on mudeli veaga võrreldes küllalt suur. Lihtsaimates mudelites on üks ainus (vabaliikmest) erinev liige ja sel juhul kontrollitakse dispersioonanalüüsi tabeli abil terve mudeli olulisust.

***Näide.*** Koostame dispersioonanalüüsi tabeli kontrollimaks leitud kasvu-kaalu mudeli olulisust.

| Varieeruvuse allikas | Ruutude summa | Vabadus-astmete arv | Keskruut | *F*-statistik | Olulisuse tõenäosus |
|---|---|---|---|---|---|
| Mudel | 54,712 | 1 | 54,712 | 7,522 | 0,0253 |
| Viga | 58,188 | 8 | 7,273 | | |
| Summaarne hajuvus | 112,9 | 9 | | | |

Seega mudel on statistiliselt oluline olulisuse nivool 0,05.

**Mudeli parameetrite usaldusvahemikud**

Mudeli parameetrite hinnangud $\hat{a}$ ja $\hat{b}$ on statistikud, mille jaotuseks on tehtud eeldustel normaaljaotus. Selle normaaljaotuse keskväärtuseks on vastava parameetri tegelik väärtus üldkogumis, dispersiooni hinnangud on aga alljärgnevad (need tuletatakse vastavalt *a* ja *b* avaldistest):

$$s_a^2 = s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_x^2}\right),\ s_b^2 = \frac{s^2}{S_x^2},\ S_x^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

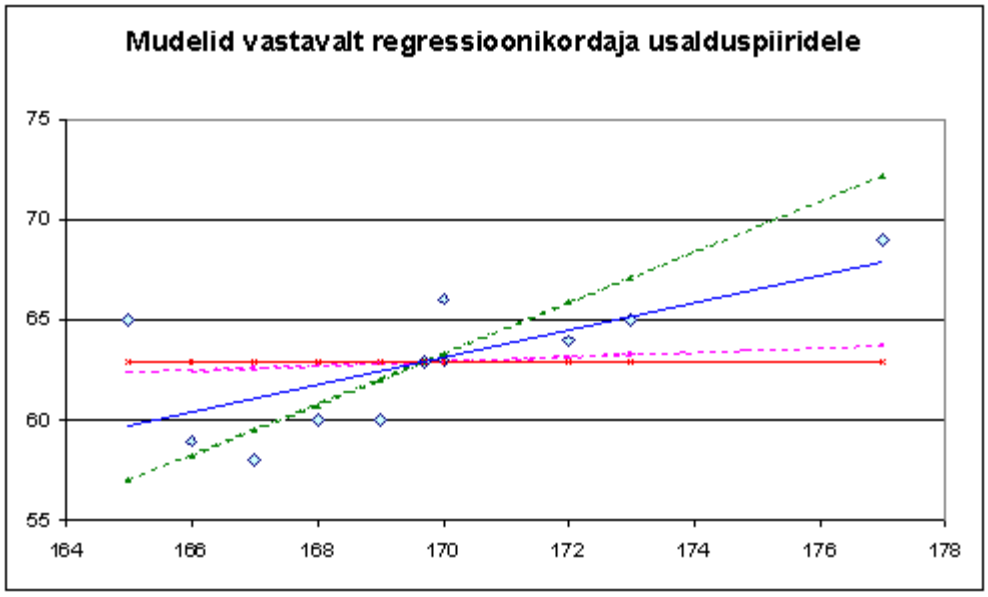
Nende hinnangute järgi on võimalik leida ka kummagi mudeli parameetri usalduspiirid:

$$\underline{a}, \bar{a} = \hat{a} \pm t_{\alpha/2}(n-2)\cdot s_a, \quad \underline{b}, \bar{b} = \hat{b} \pm t_{\alpha/2}(n-2)\cdot s_b.$$

***Näide.*** Leiame tuletatud mudeli parameetrite 95%- usalduspiirid:

| $S_a$ arvutamine | | $s_b$ arvutamine | | Usalduspiirid | |
|---|---|---|---|---|---|
| *s* | 2,697 | 1/*n* | 0,1 | $t_{0,025}(8)$ | 2,306 |
| $S_x^2$ | 116,1 | $\bar{x}^2$ | $169,7^2 = 28798,09$ | *a* usalduspiirid | *b* usalduspiirid |
| | | | | | |

| | | | | | |
|---|---|---|---|---|---|
| $S_x$ | 10,775 | $\bar{x}^2$ / $S_x^2$ | 248,046 | 0,109 | -151,563 |
| $s_a$ | 0,2503 | $s_a$ | 42,484 | 1,264 | 44,373 |




Arvutuse tulemus näitas, et regressioonikordaja on oluline, sest tema usalduspiirkonda ei kuulu nullpunkt. Seevastu vabaliige ei ole oluline, selle usalduspiirkond on väga lai ja sisaldab ka nullpunkti. Siiski ei mõjuta vabaliikme mitteolulisus mudeli olulisust. Ka ei ole mõttekas asendada saadud mudel homogeense, ilma vabaliikmeta mudeliga, sest see kirjeldab tunnust *Y* halvemini kui leitud mudel. Joonisel on näidatud, kuidas kulgeks mudel siis, kui regressioonikordaja väärtus vastaks tema alumisele ja ülemisele usalduspiirile, kusjuures on jälgitud tingimust, et regressioonisirge läbib valimi keskpunkti. Lisaks on joonisele kantud ka konstantne prognoos, millele vastab horisontaaljoon ($b$ = 0). Ka jooniselt on näha, et käesoleval juhul vastavad regressioonijoone mõlemad tõusvale (positiivse tõusunurgaga) regressioonisirgele.

**Regressioonisirge usalduspiirid**

Mudeli abil saadavad prognoosid on kõik juhuslikud suurused. Lisaks juba käsitletud prognoosiveale tuleb arvesse võtta ka seda, et mudeli kordajad ei ole täpselt teada, vaid on valimi põhjal hinnatud ja sisaldavad samuti, nagu me nägime, arvestatava suurusega juhuslikku viga.

Arvestades eeldust, et prognoosivead on normaaljaotusega ja sõltumatud, saame prognoosi jaoks standardhälbe hinnangu

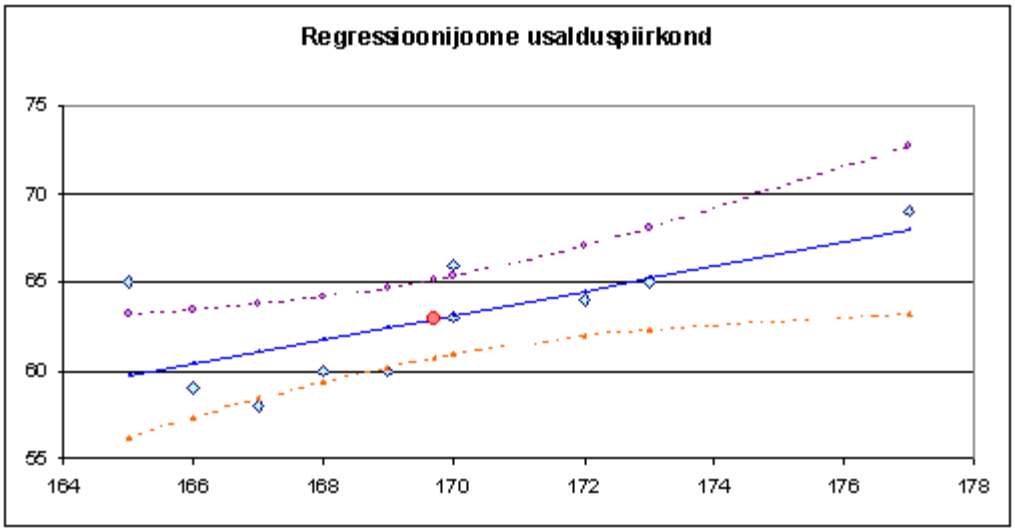
$$s_p = s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_x^2}}.$$

Sellest hinnangust on näha, et prognoosi standardhälve on valimi keskpunkti läheduses kõige väiksem, mis tuleneb regressioonijoone seotusest valimi keskpunktiga.

Esitatud seos kirjeldab regressioonisirge hajuvust oma teoreetilise väärtuse ümber,

sisuliselt on selles arvesse võetud niihästi regressioonikordaja kui ka vabaliikme hajuvus. Seda arvestades on võimalik leida regressioonisirge jaoks usalduspiirkond, arvutades igale $X$ väärtusele $x_0$ vastavalt prognoosi usalduspiirkond, kasutades standardset normaaljaotusega statistiku usalduspiiride arvutuseeskirja $t$-jaotuse abil:

$$\hat{y}_0 \pm t_{\alpha/2}(n-2)\cdot s_{\rho}.$$




Lisatud joonisel on kujutatud regressioonijoone usalduspiirkond, st piirkond, milles (valimi andmetel) paikneb mudelit esitav regressioonijoon tõenäosusega 0,95. Sellesse piirkonda satuvad sama tõenäosusega teoreetilised prognoosid.

**Järelemõtlemiseks**

1.Mida peaks tegema uurija, kes usub, et vaadeldavat nähtust saab mudeliga kirjeldada, kuid leitud mudel ei osutu statistiliselt oluliseks?

2.Kas sellest, et $X$ ja $Y$ vaheline korrelatsioonikordaja on/ei ole statistiliselt oluline, järeldub (samal olulisuse nivool) ka lineaarse mudeli olulisus/ mitteolulisus?

3.Millal on mudel statistiliselt oluline, kuid korrelatsioonikordaja on väike?

4.Kas see, kui regressioonikordaja on väike, näitab lineaarse seose nõrkust või midagi muud? Mida?

VASTUSED

**Statistiline sõltuvus ja statistiline mudel**

**ÜLDISEMAD MUDELID**

**Mudelite üldistamise teed**

Väga sageli ei piisa nähtuste kirjeldamiseks lihtsast lineaarsest mudelist, ning mõõtmisandmete põhjal on tarvis konstrueerida keerukamaid mudeleid. Kõige tavalisemad võimalused selleks on järgmised:

1.Lisada juurde argumente, saades sellega mitmesed regressioonimudelid;

2.Lisada juurde argumentide funktsioone, saades sel kujul näiteks polünomiaalsed mudelid;

3.Lisada juurde mittearvulisi tunnuseid esindavaid nn libatunnuseid (dummy variables);

4.Teisendada funktsioontunnust – mis üldjuhul viib välja lineaarsete mudelite klassist üldistatud lineaarsete funktsioonide klassi;

5.Eritüüpi mudeliteks on ajast sõltuvad mudelid, sh aegread, mis sisaldavad teatavaid eriomadustega liikmeid.

Üldiselt lisandub mudelite keerukamaks muutumisega ka täiendavaid probleeme seoses mudeli eelduste kontrollimisega. Siiski sobivad väga paljudel juhtudel esitatud vähimruutude metoodika põhjal saadud mudelid vähemalt esimesteks lähenditeks, mis saavad olla aluseks spetsiaalsetele süvauuringutele.

**Mitmene lineaarne regressioon**

Juhuslik suurus $Y$ võib sõltuda enam kui ühest arvtunnusest $X_1$, $X_2$, ..., $X_q$. Sel juhul on uuringu eesmärgiks otsida mudelit

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_q X_q + \varepsilon,$$

kus $Y$ ja $X_1$ ... $X_q$ on tunnused, $b_0$, ..., $b_q$ valimi põhjal hinnatavad mudeli parameetrid ja $\varepsilon$ juhuslik viga, mille kohta eeldatakse, et tema jaotus on normaalne ning üksikvaatluste vead on sõltumatud.

•Mudeli parameetrite hindamiseks saab kasutada vähimruutude meetodit, mis annab parameetrite hindamiseks samuti lineaarse võrrandisüsteemi nagu ka ühest argumendist sõltuva mudeli puhul.

•Mudeli headuse hindamiseks saab kasutada mitmest korrelatsioonikordajat, mis on tavalise korrelatsioonikordaja analoog (sisuliselt korrelatsioonikordaja parima prognoosi ja funktsioontunnuse vahel) ja mille ruut – determinatsioonikordaja iseloomustab mudeli poolt kirjeldatud osa funktsioontunnuse hajuvuses.

Mitmese lineaarse regressiooni olulisemad probleemid:

1.argumentide otstarbekaim valik siis, kui potentsiaalsete argumentide hulk on suur;

sobiv on valida mudelisse vaid neid argumente, mille lisamine mudelit oluliselt parandab.

2.multikollineaarsus, mis ilmneb selles, et omavahel tihedalt korreleeritud argumentide korral muutub lahend ebastabiilseks ja raskesti interpreteeritavaks, seetõttu tuleks vältida korreleeritud argumente mudelis.

***Näide***

Vaatleme taas neidude pikkusi ja kaalusid kirjeldavat andmestikku, kusjuures lisame juurde teise argumendina talje ümbermõõdu. Selgub, et antud juhul multikollineaarsust pole, sest talje ümbermõõdu ja pikkuse korrelatsioonikordaja on vaid 0,14. Saime uue mudeli kujuga:

"Kaal" = 0,614 "Pikkus" + 0,718 "Taljeümbermõõt"– 91,5.

Kõik kolm mudelis sisalduvat liiget on statistiliselt olulised olulisuse nivool 0,05, mudelit iseloomustav determinatsioonikordaja on 0,757 ($R^2$ = 0,87).

Seega paranes varem vaadeldud mudel varasemaga võrreldes tunduvalt, nagu näitavad ka prognoosivead lisatud tabelis:

| Jrk nr | Pikkus | Talje ümbermõõt | Kaal | Ennustatud kaal | Prognoosi-viga |
|---|---|---|---|---|---|
| 1 | 173 | 72 | 65 | 66,434 | -1,434 |
| 2 | 168 | 69 | 60 | 61,210 | -1,210 |
| 3 | 165 | 75 | 65 | 63,674 | 1,326 |
| 4 | 177 | 70 | 69 | 67,455 | 1,545 |
| 5 | 170 | 70 | 63 | 63,156 | -0,156 |
| 6 | 167 | 68 | 58 | 59,878 | -1,878 |
| 7 | 166 | 65 | 59 | 57,110 | 1,890 |
| 8 | 169 | 69 | 60 | 61,824 | -1,824 |
| 9 | 172 | 71 | 64 | 65,102 | -1,102 |
| 10 | 170 | 70 | 66 | 63,156 | 2,844 |

**Argumentide funktsioonide kasutamine mudelis**

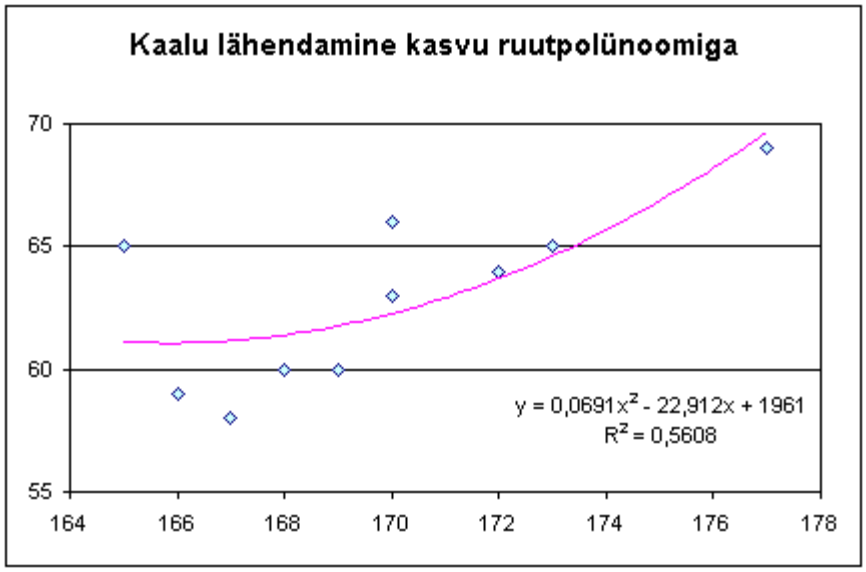
Tunnusega $X$ koos on üheselt määratud ka selle funktsioonid $X^2$, ln$X$, $e^X$ jne, mida samuti saab kasutada mudeli argumentidena (jälgides vaid vastavate funktsioonide määramispiirkondi). Kasutades mitmese regressioonimudeli põhimõtet võib konstrueerida mudeleid, mis sisaldavad argumendi funktsioone. Üks populaarsemaid võimalusi on polünomiaalne regressioon:

$$Y = b_0 + b_1 X + b_2 X^2 + \ldots + b_q X^q + \varepsilon$$

mille korral on oluline jälgida võimaliku multikollineaarsuse mõju, mistõttu maksimaalne aste $q$ üldiselt ei tohiks olla suur.

***Näide***

Püüame nüüd neidude kaalu prognoosida nende kasvu ruutfunktsioonina, vt lisatud joonis




Kuigi saadud lähend on mõnevõrra parem lineaarsest lähendist (determinatsioonikordaja suurenes 0,48-lt 0,56-ni), osutub saadud mudel statistiliselt mitteoluliseks – kõigi kolme liikme olulisuse tõenäosus on suurem kui 0,3. Selle põhjuseks on pikkuse ja pikkuse ruudu omavaheline äärmiselt tugev korrelatsioon ($r$ = 0,9999), mis põhjustab mudeli mittekollineaarsust. Ei saaa järeldada, et leitud mudel üldkogumis hästi töötaks!

**Libatunnustega mudelid**

Regressioonimudeleid saab konstrueerida põhimõtteliselt ainult arvuliste argumentide abil. Selgub aga, et mõnigi kord leidub mittearvulisi tunnuseid, millel on tugev mõju funktsioontunnusele. Niisuguste tunnuste mudelisse lülitamiseks kasutatakse neile vastavaid libatunnuseid (indikaatoreid).

Olgu tunnusel $X$ $m$ mittearvulist väärtust $a_1$, $a_2$, …, $a_m$. Sellele tunnusele $X$ vastab $m$ –1 libatunnust $Z_1$, $Z_2$, …, $Z_{m-1}$, mis on defineeritud nii:

$$X = a_1 \Leftrightarrow Z_1 = 1, Z_2, \ldots, Z_{m-1} = 0;$$

$$X = a_2 \Leftrightarrow Z_2 = 1, Z_j = 0, \text{ kui } j \neq 2,$$

$$X = a_m \Leftrightarrow Z_j = 0, j = 1, \ldots, m-1.$$

Need libatunnused on kõik arvulised ja neid võib lisada tavalisel viisil regressioonimudelisse. Mudeli tõlgendamisel tuleb aga arvestada, et juhul, kui $X = a_i$, tuleb mudeli vabaliikmele lisada kordaja $b_i$.

***Näide***

Oletagem, et vaadeldud neidude hulgast osa pärinevad linnast, osa maalt. Moodustame libatunnuse $Z$, mis omandab väärtuse 1 maalt pärinevate neiude puhul (järjekorras 3., 7. ja 10.), ning lisame selle tunnuse mudelisse. Selgus, et saime mudeli kujuga:

"*Kaal*" = 0,998 "*Pikkus*" + 4,47 "*Maalt*"– 107,9.

Kõik liikmed on olulisuse nivool 0,05 statistiliselt olulised, determinatsioonikordaja on 0,756, $R^2$ = 0,87. Käeolevat mudelit võime tõlgendada nii:

"Linnaneiu kaal" = 0,998 "Pikkus" – 107,9; Maaneiu kaal = 0,998 "Pikkus" – 103,4.

**Funktsioontunnuse teisendamine**

Mõnikord on tarvis teisendada ka funktsioontunnust. Kui kasutatakse varasemaga sarnast lähenemist, tuleb arvesse võtta, et mudel kaotab oma optimaalsuse algse tunnuse suhtes, sellepärast sobivad niisugused lihtsad mudelid ainult esialgsesk lähendiks.

***Tõenäosuse prognoosimine***

Sageli on tarvis leida mudeleid mingite sündmuste tõenäosuse prognoosimiseks. Siin on aga probleemiks see, et tõenäosus ei ole normaaljaotusega, tema väärtused on piiratud lõiguga [0, 1]. Et sellest probleemist üle saada, kasutataksegi mitmesuguseid teisendusi, milledest olulisim on ***šanss***.

• Sündmuse *A* šansiks nimetatakse suhet $O(A) = P(A)/(1 - P(A))$, kus $P(A)$ on sündmuse tõenäosus.

Leidnud mudeli šansi jaoks, on suhteliselt lihtne siit leida ka mudel tõenäosuse jaoks. Šansside jaoks optimaalsete mudelite leidmisega tegeleb üldistatud lineaarsete mudelite teooria.

***Multiplikatiivsed mudelid***

Argument- ja funktsioontunnuste logaritmimine võimaldab leida multiplikatiivseid mudeleid, kasutades lineaarsete mudelite jaoks välja töötatud tehnikat. Tõepoolest, leides mudeli $\ln(Y)$ kirjeldamiseks argumentide $\ln(X_1)$ ja $\ln(X_2)$ kaudu saame $Y$ avaldada argumentide (astmete) korrutisena.

$$ln(Y) = b_0 + b_1\, ln(X_1) + b_2\, ln(X_2) \Rightarrow ln(Y) =$$
$$ln(B_0 X_1^{b_1} X_2^{b_2}) \Rightarrow Y = B_0 X_1^{b_1} X_2^{b_2}, B_0 = e^{b_0}.$$

**Aegrea tüüpi mudelid**

Mudeli argumendiks võib olla ka aeg. Kui mingit tunnust *X* mõõdetakse võrdsete ajavahemike tagant, siis moodustavad mõõtmistulemused andmestiku, mida nimetatakse aegreaks. Aegridade puhul on sageli võimalik kasutada teatavat lisainformatsiooni, mille allikaks on teiste tegurite ajast sõltuv mõju.

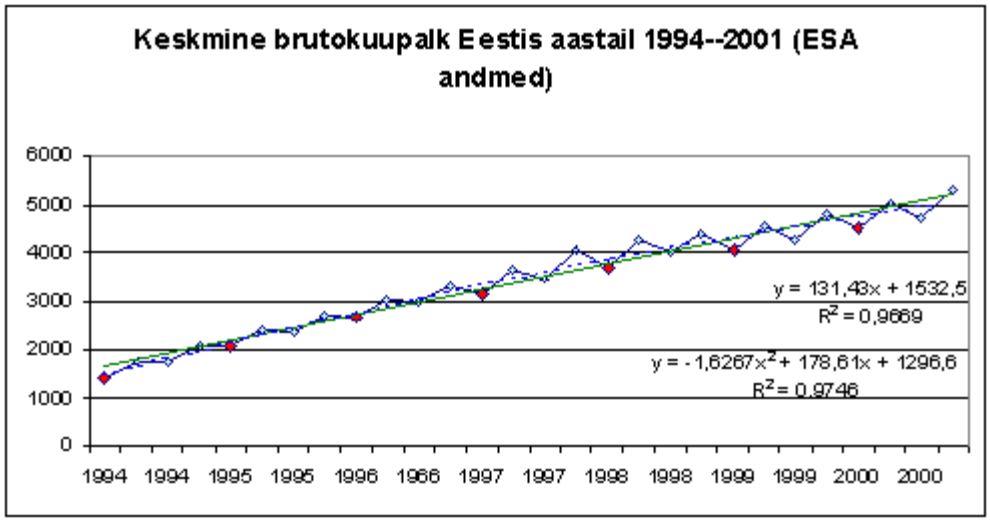
Sellest lähtuvalt esitatakse aegrida sageli järgmiselt:

$$X_t = f(t) + s(t) + \varepsilon_t,$$

kus $f(t)$ tähistab trendi, st aja põhimõtteliselt mittejuhuslikku funktsiooni, mida kirjeldava mudeli parameetrid hinnatakse vaatluste põhjal. Teine liige, $s(t)$ on sesoonnsus, see muutub perioodiliselt vastavalt aja tsüklilisele iseloomule – sesoonsuse perioodiks võib olla, aasta, nädal, ööpäev jm. Lisandub juhuslik viga. Selleks, et leida trendi kirjeldav mudel on tavaliselt tarvis enne aegreast eemaldada sesoonsus, seejärel on võimalik kasutada varem kirjeldatud mudelite leidmise meetodeid; võimalik on aga ka tegutsemine teises järjekorras.

***Näide***

Eesti brutopalga muutus 7 aasta jooksul. Joonisel on I kvaratali palgad märgitud punaste suuremate sümbolitega, lisatud on ka lineaarne ja ruut-trend.




**Järelemõtlemiseks**

1.Olgu mudeli jaoks võimalikke potentsiaalseid argumente palju. Miks tuleb teha nende seas valik?

2.Kuidas muutub mitmene korrelatsioonikordaja argumendi lisamisel mudelisse? Miks?

3.Kas determinatsioonikordaja on mitmesest korrelatsioonikordajast suurem või väiksem?

4.Kas mitmene regressioonikordaja on suurem või väiksem võrreldes funktsioontunnuse ja argumentide vaheliste korrelatsioonikordajatega?

VASTUSED

**Statistiline sõltuvus ja statistiline mudel**

**LINEAARNE MUDEL**

1. Suhteliselt raskeim on neiu nr 3, temale järgnevad neiud nr 10 ja 4. Kergeim on neiu nr 6, järgnevad 8, 2 ja 7.

2. See punkt asub regressioonisirgel.

3. Konstantsele mudelile vastab horisontaalne sirge, mille paiknemise määrab *EY* väärtus.

4. Kõik hajuvusdiagrammi punktid paikneksid regressioonisirgel.

5. Graafiku siht muutuks, senise tõusva sirge asemele tuleks langev sirge (või vastupidi).

6. Vabaliikme muutumisega kaasneb graafiku paralleelnihe. Vabaliikme suurenedes nihkub graafik ülespoole, vähenedes aga allapoole.

7. Graafik muutub, tuleb arvutada uued regressiooniparameetrid.

8. See punkt “kallutaks” graafikut enese suunas, kusjuures üksiku punkti mõju on seda suurem, mida vähem on valimis punkte ja mida kaugemal see punkt ülejäänutest paikneb.

**Statistiline sõltuvus ja statistiline mudel**

**MUDELI STATISTILINE OLULISUS**

1. Suurendama valimi mahtu, see viib alati sihile, kui mudel tegelikult kehtib, kuid mõnikord võib vajalik punktide arv olla väga suur. Teine võimaldus on suurendada olulisuse nivood, kuid sellega langeb mudeli usaldusväärsus (näiteks kui mudel on oluline olulisuse nivoo 0,25 puhul, ei ole see kuigi usaldusväärne).

2. Jah.

3. See võib juhtuda siis, kui valimi maht on suur.

4. See näitab regressioonisirge paiknemist graafikul (tõus on väike), ning see ei ole seotud seos tugevusega.

**Statistiline sõltuvus ja statistiline mudel**

**ÜLDISEMAD MUDELID**

1.Üldiselt on tõlgendamise seisukohast lihtsam mudel parem. On ka selge, et suure argumentide arvuga mudelisse täiendavate argumentide lisamine saab mudelit ainult vähe parandada (sest maksimaalne kirjeldatuse tase on piiratud 100%-ga). Suurema hulga argumentide puhul on reaalsem ka multikollineaarsuse oht.

2.Kas suureneb või jääb samaks, sest mudel optimeerib kirjelduse taseme; kui lisanduv argument seda ei paranda, on vastav regressioonikordaja 0 ja mitmene korrelatsioonikordaja jääb samaks.

3.Väiksem.

4.Suurem või võrdne. Ühe argumendi puhul on mitmene korrelatsioonikordaja võrdne tavalise korrelatsioonikordaja absoluutväärtusega. Mitme argumendi puhul on üldiselt suurem kummagi argumendi ja funktsioontunnuse korrelatsioonikordaja absoluutväärtusest.