Sisukord

Peatükk 1

Meeldetuletus

Mõistel *juhuslik suurus* ja *juhusliku suuruse jaotus* on matemaatilises statistikas äärmiselt tähtis koht. Jaotus iseloomustab juhusliku suuruse väärtuste paiknemist, määrates võimalike väärtuste hulga ja esinemistõenäosused. Järgnevalt tuletame meelde diskreetse ja pideva juhusliku suuruse mõisteid ning põhilisi jaotusi, mida oleme õppinud kursusest 'Tõenäosusteooria ja statistika I' . Järgnev tekst põhineb õpikutel Traat (2006) ja Pärna (2013).

1.1 Diskreetne juhuslik suurus

Diskreetseks juhuslikuks suuruseks nim. funktsiooni $X : \Omega \to \mathcal{R}$, mis võtab kas lõpliku või loenduva arvu väärtuseid $x_1, x_2, ..., x_{(n)}$.

Diskreetsete juhuslike suuruste korral on kogu tõenäosusarvutuste jaoks vajalik info kirjas jaotuses ehk paarides (x_i, p_i), kus $x_i,\ i \in I$ on võimalikud väärtused ja $p_i = P(\{X = x_i\})$ on nende väärtuste tõenäosused. Samuti teame, et kõikide juhuslike suuruste korral on kogu tõenäosusarvutuste jaoks vajalik info olemas jaotusfunktsioonis $F_X(x) = P(\{X \leq x\})$.

Keskväärtus ja dispersioon arvutatakse valemitega

$$EX = \sum_i x_i \cdot p_i,$$

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

Näide 1 *Ühe ettevõtte osakond (kokku 6 inimest) on leidnud, et töötaja haigestumise tõenäosus on 0,1. Olgu juhuslik suurus X haigestunud töötajate arv hommikul.*

Siis on selle juhusliku suuruse jaotust võimalik ette anda näiteks tabelina (eeldades, et töötajad istuvad igaüks oma kabinetis, ehk haigestuvad üksteisest sõltumatult):

$X = x_i$	0	1	2	3	4	5	6	Σ
p_i	0,531	0,354	0,098	0,015	0,002	0	0	1

Seda jaotust on võimalik ette anda ka graafiliselt vertikaalsete sirglõigete abil, mis algavad x-telje väärtustest 0, 1,..,6 ja mille pikkus on võrdne väärtusega p_i, $i = 0, 1, ..., 6$. Võimalusi on veelgi: tõenäosusfunktsiooni abil (ei pruugi alati leiduda, siin sobiks binoomjaotuse

valem); jaotusfunktsiooni abil (harjutus lugejale) ja ka jaotusfunktsioonile vastava graafiku abil (samuti harjutus lugejale).

1.1.1 Tuntumad diskreetsed jaotused

Bernoulli jaotus – kahe võimaliku väärtusega {0,1} jaotus:

$$X \sim B(1,p),$$

kus $P(X=1)=p, \quad P(X=0)=1-p$. Tõenäosusfunktsiooniks on

$$p(x)=p^x(1-p)^{1-x}, \quad x \in \{0,1\}.$$

Keskväärtus ja dispersioon on

$$EX = 0 \cdot p(0) + 1 \cdot p(1) = p,$$

$$DX = (0-p)^2 \cdot p(0) + (1-p)^2 \cdot p(1) = p(1-p).$$

Bernoulli jaotuse väärtused 1 ja 0 võivad olla koodid mingi omaduse A esinemisele või mitteesinemisele objektil. Jaotuse parameeter p näitab omaduse A tõenäosust katses, lõpliku üldkogumi korral ka A osakaalu üldkogumis. Bernoulli jaotusega on 'jah'/'ei' tunnused, kus 'jah' võib tähendada mingi arvamuse, haiguse jm. olemasolu.

Binoomjaotus – väärtustega $x \in \{0,1,\ldots,n\}$ jaotus, mida tähistatakse

$$X \sim B(n,p),$$

kus tavalise interpretatsiooni kohaselt on n katseseeria pikkus, milles vaadeldakse sündmuse A esinemist, ning p on sündmuse A esinemise tõenäosus ühes katses. Tõenäosusfunktsioon on

$$p(x)=C_n^x p^x (1-p)^{n-x} \quad , x \in \{0,1,\ldots,n\}, \quad C_n^x = \frac{n!}{x!(n-x)!}.$$

Keskväärtus ja dispersioon on vastavalt

$$EX = np,$$

$$DX = np(1-p).$$

Bernoulli jaotus on binoomjaotuse erijuht $n=1$ korral. Statistikas esineb binoomjaotus sageli mingi omaduse/sündmuse esinemiste arvu jaotusena valimis. Oletagem, et Eestis on 7% töötuid. Ülaltoodud valemid annavad vastuse küsimustele: mis jaotusega on töötute arv 100ses juhuslikus valimis; kui palju on selles valimis oodatavalt töötuid?

Geomeetriline jaotus väärtustega $x=1,2,...$ on jaotus, mida tähistatakse $X \sim Geom(p)$, mis tekib siis kui vaadeldakse katsete arvu huvipakkuva sündmuse A esimese toimumiseni (katsed on sõltumatud). Tõenäosusfunktsiooniks on

$$p(x)=(1-p)^{x-1} \cdot p,$$

kus $p=P(A)$. Keskväärtus ja dispersioon on

$$EX = \frac{1}{p}, \quad DX = \frac{1-p}{p^2}.$$

Poissoni jaotus – loenduva arvu väärtustega, $x \in \{0, 1, \ldots\}$, jaotus, mida tähistatakse

$$X \sim Po(\lambda).$$

Väärtuse x esinemise tõenäosus arvutatakse tõenäosusfunktsiooniga

$$p(x) = e^{-\lambda}\frac{\lambda^x}{x!}.$$

Parameeter λ on antud juhul nii keskväärtus kui ka dispersioon:

$$EX = \lambda, \quad DX = \lambda.$$

Poissoni jaotusega on sageli mingi 'sündmuse esinemiste arv ajavahemikul', näiteks 'õnnetusjuhtumite arv Tallinn-Tartu maanteel septembri esimesel nädalal'. Täpsemini öeldes, on Poissoni jaotus niisuguste juhuslike suuruste jaoks sageli sobivaimaks mudeliks.

1.2 Pidevad juhuslikud suurused

Juhuslikku suurust X nimetatakse pidevaks, kui tema jaotusfunktsioon on esitatav kujul

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

mingi funktsiooni f korral. Funktsiooni f nimetatakse juhusliku suuruse X tihedusfunktsiooniks. Pideva juhusliku suuruse väärtuspiirkonnaks on reaaltelg või selle mingi osa. Tema keskväärtus ja dispersioon avalduvad tihedusfunktsiooni abil seostega:

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

$$DX = \int_{-\infty}^{\infty} (x - EX)^2 \cdot f(x)dx.$$

1.2.1 Tuntumad pidevad jaotused

Ühtlane jaotus, $X \sim U(a, b)$, on määratud lõplikul lõigul $[a, b]$ ja tema tihedusfunktsioon avaldub kujul

$$f(x) = \frac{1}{b-a}, \ kui \ x \in [a, b].$$

Jaotuse keskväärtus on lõigu keskpunkt $EX = (a+b)/2$ ja dispersioon $DX = (b-a)^2/12$. Ühtlase jaotusega $U(0, b)$ on näiteks bussi ootamise aeg, kui minna peatusse juhuslikult ja bussid läbivad seda intervalliga b.

Eksponentjaotus $X \sim Exp(\theta)$, omab tihedusfunktsiooni

$$f(x) = \theta e^{-x\theta}, \quad x \geq 0.$$

Jaotuse keskväärtust on $EX = \frac{1}{\theta}$ ja dispersioon on $DX = \frac{1}{\theta^2}$. Eksponentjaotus lihtsaimaks mudeliks tunnuse 'eluiga' (ka seadmete oma) jaotuse kirjeldamisel.

Normaaljaotusega juhuslikku suurust X keskväärtusega μ ja dispersiooniga σ^2 tähistatakse

$$X \sim N(\mu, \sigma).$$

Tema tihedusfunktsioon esitub valemiga

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

Tähtis omadus on lineaarteisendus:

$$X = \sigma Y + \mu \sim N(\mu, \sigma),$$

kus $Y \sim N(0, 1)$. Normaaljaotusega $N(1, \sigma)$ on näiteks 'piima hulk 1 liitrises piimapakis', kus σ iseloomustab pakkimisliini täpsust. Mõõtes rea piimapakkide täituvust, saame valimi, mille väärtused varieeruvad 1 liitri ümber.

Peatükk 2

Juhuslikud suurused ja vektorid

Jätkame teadmiste omandamist sellest, kuidas kirjeldada juhuslike suuruseid ja juhuslikke vektoreid, mis on nendega seotud põhimõisted ja võimalused tõenäosusarvutuste teostamiseks. Järgnev peatükk põhineb R. Kangro konspektil aastal 2015 ja õpikul Meyer (1970).

2.1 Momente genereeriv funktsioon

Teame, et diskreetsete juhuslike suuruste korral on kogu tõenäosusarvutuste jaoks vajalik info kirjas jaotuses ehk paarides (x_i, p_i), kus $x_i,\ i \in I$ on võimalikud väärtused ja $p_i = P(\{X = x_i\})$ on nende väärtuste tõenäosused.

Samuti teame, et kõikide juhuslike suuruste korral on kogu tõenäosusarvutuste jaoks vajalik info olemas jaotusfunktsioonis $F_X(x) = P(\{X \leq x\})$. Pidevate juhuslike suuruste korral on aga paljude arvutuste jaoks mugavam kirjeldus tihedusfunktsiooni f_X kaudu.

Osutub, et võimalikke kirjeldusi, mis on eriti mugavad mitmete tõenäosusarvutuste tegemiseks, on veelgi. Üks sellistest kirjeldustest on momente genereeriv funktsioon, mis leidub ainult osadel juhuslikel suurustel, kuid mis on väga mugav mõningate teoreetiliste tulemuste näitamisel.

Defineerime vajalikud mõisted.

Definitsioon 1 *Juhusliku suuruse X k-ndat järku momendiks nimetatakse arve*

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

Seega keskväärtus on esimest järku moment m_1 ning dispersioon avaldub kujul $DX = m_2 - m_1^2$.

Osutub, juhusliku suuruse momentide arvutamine on seotud järgneva funktsiooniga.

Definitsioon 2 *Juhusliku suuruse X momente genereerivaks funktsiooniks M_X nimetatakse funktsiooni*

$$M_X(t) = E(e^{tX}), t \in \mathrm{R}.$$

On selge, et iga juhusliku suuruse momente genereeriv funktsioon on defineeritud $t = 0$ korral ning $M_X(0) = 1$. Samas leidub juhuslikke suuruseid, mille momente genereeriv

funktsioon ei ole defineeritud ühegi teise t väärtuse korral (näiteks Cauchy jaotus tihedusfunktsiooniga $f_X(x) = \frac{1}{\pi(1+x^2)}$). Sellistel puhkudel ei anna momente genereeriv funktsioon meile mingit kasulikku infot jaotuse kohta ning seetõttu sageli öeldakse, et juhuslikul suurusel (või jaotusel) on olemas momente genereeriv funktsioon ainult siis, kui selle väärtused on defineeritud mingis 0-punkti sisaldavas vahemikus.

Näide 2 *Olgu $X \sim Exp(2)$. Leiame suuruse X momente genereeriva funktsiooni:*

$$\begin{aligned} M_X(t) = E(e^{tX}) &= \int_0^\infty e^{tx} \cdot 2e^{-2x}\,dx \\ &= \int_0^\infty 2e^{(t-2)x}\,dx \underset{t\neq 2}{=} \left.\frac{2e^{(t-2)x}}{t-2}\right|_{x=0}^{\infty} = \frac{2}{t-2}(\lim_{x\to\infty} e^{(t-2)x} - e^0) \\ &\underset{t<2}{=} \frac{2}{t-2}(0-1) = \frac{2}{2-t},\ t<2. \end{aligned}$$

Arvutustest on näha, et nii $t = 2$ kui $t > 2$ korral on integraal lõpmatu.

Momente genereeriva funktsiooni nime õigustab järgmine tulemus.

Lemma 1 *Kui juhusliku suuruse X momente genereeriv funktsioon eksisteerib $t = 0$ mingis ümbruses (st vahemikus $|t| < \delta$ mingi $\delta > 0$ korral on M_X väärtused lõplikud), siis k-s moment avaldub selle funktsiooni k-ndat järku tuletise kaudu kujul*

$$m_k = M_X^{(k)}(0).$$

Tõestus. Siin kursuses tõestame selle tulemuse lõpliku arvu väärtustega diskreetsete juhusliku suuruse korral, täielik tõestus antakse magistritaseme kursuses „Tõenäosusteooria II“.

Olgu X diskreetne juhuslik suurus väärtustega $x_1, x_2, \ldots, x_n$, siis keskväärtuse lineaarsuse omaduse põhjal

$$M_X(t) = \sum_{i=1}^{n} e^{tx_i} p_i,$$

kus $p_i = P(\{X = x_i\})$. Tuletise lineaarsuse tõttu võime lõplikus summa tuletise leida liikmete tuletiste summana:

$$M_X^{(k)}(t) = \sum_{i=1}^{n} p_i \frac{d}{dt^k}(e^{tx_i}) = \sum_{i=1}^{n} x_i^k p_i e^{tx_i}.$$

Seega

$$M_X^{(k)}(0) = \sum_{i=1}^{n} x_i^k p_i,$$

mis ongi võrdne k-ndat järku momendiga $E(X^k)$.

Kui me teaksime, et $M_X(t)$ on tehtud eeldustel alati lõpmatult palju kordi diferentseeruv ja et tuletise võtmise ja keskväärtuse leidmise järjekorda saab praegusel juhul muuta, oleks ka üldjuhul tõestus lihthe:

$$M_X^{(k)}(t) = \frac{d}{dt^k} E(e^{tX}) = E(\frac{d}{dt^k} e^{tX}) = E(X^k e^{tX}),$$

kust $t = 0$ korral saaksime võrduse $M_X^{(k)}(0) = E(X^k)$. Samas nii lõpmatute summade kui integraalide puhul ei ole selline järjekorra vahetamine (tuletisega keskväärtuse alla minemine) alati õigustatud ja nõuab põhjalikku põhjendamist. Nagu mainitud, tõestatakse see tulemus üldkujul hilisemas kursuses. □.

Näide 3 *Eelmise näite põhjal teame, et $X \sim Exp(2)$ korral $M_X(t) = \frac{2}{2-t}$. Kuna*

$$M'_X(t) = \frac{2}{(2-t)^2},\ M''_X(t) = \frac{4}{(2-t)^3},$$

siis eelneva lemma põhjal $EX = m_1 = M'_X(0) = \frac{1}{2}$, $E(X^2) = m_2 = M''_X(0) = \frac{1}{2}$ ning seega $DX = m_2 - m_1^2 = \frac{1}{4}$. Need tulemused on muidugi juba varasemast teada, kuid sageli on juhusliku suuruse momente lihtsam leida momente genereerivat funktsiooni diferentseerides kui vastavat keskväärtust otse arvutades.

Sageli tuleb kasuks ka teadmine, kuidas on juhusliku suuruse lineaarse funktsiooni abil defineeritud juhusliku suuruse momente genereeriv funktsioon seotud esialgse juhusliku suuruse momente genereeriva funktsiooniga. Selleks tõestame järgmise tulemuse.

Lemma 2 *Kui juhusliku suuruse X momente genereeriv funktsioon on M_X, siis juhusliku suuruse $Y = aX + b$ momente genereeriv funktsioon on avaldatav kujul*

$$M_Y(t) = e^{bt} M_X(at), t \in \mathbf{R}.$$

Tõestus. Definitsiooni kohaselt

$$M_Y(t) = E(e^{tY}) = E(e^{t(aX+b)}) = E(e^{(at)X} e^{bt}) \underset{E(cX)=cEX}{=} e^{bt} M_X(at). \square$$

Näide 4 *Kasutame eelnevat tulemust, et leida normaaljaotusega $N(\mu, \sigma)$ juhusliku suuruse momente genereeriva funktsiooni avaldis. Kõigepealt leiame standardse normaaljaotusega juhusliku suuruse X momente genereeriva funktsiooni avaldise:*

$$\begin{aligned} M_X(t) = E(e^{tX}) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2} + tx - \frac{t^2}{2} + \frac{t^2}{2}}\, dx \\ &= e^{\frac{t^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-t)^2}{2}}\, dx \\ &= e^{\frac{t^2}{2}},\ t \in \mathbf{R}, \end{aligned}$$

kus eelviimasel real olev integraal võrdub ühega seetõttu, et integraali all on jaotuse $N(t, 1)$ tihedusfunktsioon. Normaaljaotuse lineaarse teisenduse omaduse põhjal on juhusliku suuruse $X \sim N(0, 1)$ korral juhuslik suurus $Y = \sigma X + \mu$ jaotusega $N(\mu, \sigma)$. Seega, jaotuse $N(\mu, \sigma)$ juhusliku suuruse momente genereerivaks funktsiooniks on

$$M_Y(t) = e^{\mu t} M_X(\sigma t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}.$$

Oluline on ka teadmine, et kui momente genereeriv funktsioon omab lõplikke väärtuseid mingis nullpunkti ümbruses, siis on selle funktsiooni põhjal võimalik kindlaks teha vaadeldava juhusliku suuruse jaotus. Nimelt kehtib järgmine ühesuse tulemus.

Lemma 3 *Kui juhuslike suuruste X ja Y momente genereerivad funktsioonid M_X ja M_Y omavad mõlemad lõplikke ning võrdseid väärtuseid mingis nullpunkti sisaldavas vahemikus, siis on juhuslikud suurused X ja Y sama jaotusega.*

See lemma tõestatakse samuti kursuses „Tõenäosusteooria II“. Nagu me hiljem näeme, võimaldab momente genereerivate funktsioonide ühesuse omadus sageli kindlaks teha sõltumatute juhuslike suuruste summa jaotust: leiama summa momente genereeriva funktsiooni ning kui vastab mõne tuntud jaotuse momente genereerivale funktsioonile, siis ongi summa jaotus kindlaks tehtud.

2.2 Tinglik keskväärtus ja tinglik jaotus. Diskreetsed juhuslikud vektorid. Juhusliku vektori jaotusfunkctioon

Selles peatükis tutvume tingliku keskväärtuse mõiste ja lihtsamate omadustega. Samuti tuletame meelde diskreetsete juhuslike vektoritega seotud mõisteid ning omadusi ning juhusliku vektori jaotusfunktsiooni omadusi.

2.2.1 Tinglik jaotus ja tinglik keskväärtus.

Aines „Tõenäosusteooria ja statistika I“ tutvusime tingliku tõenäosuse mõistega. Osutub, et praktikas pakub sageli huvi nn tinglik keskväärtus.

Näide 5 *Veeretatakse kahte täringut. Olgu X_1 esimesel täringul saadud silmade arv ja X_2 vastavalt teisel. Olgu $Y = X_1 + X_2$ ehk kahel täringul saadud silmade summa. Oskame leida juhusliku suuruse Y keskväärtuse (mitu silma keskmiselt tuleb kahel täringul kokku):*

$$EY = E(X_1 + X_2) = E(X_1) + E(X_2) = 2E(X_1) = 2 \cdot 3,5 = 7,$$

sest

$$E(X_1) = \sum_{k=1}^{6} k \cdot \frac{1}{6} = 3,5.$$

Võiksime aga püstitada järgnevaid küsimusi:

- *Millega võrdub summa Y keskväärtus (teiste sõnadega kahel täringul oodatav silmade arvude summa) juhul, kui on teada, et esimesel tuli 2 silma?*
- *Millega võrdub esimese täringu keskmine saadud silmade arv juhul kui on teada, et summa tuli 5?*

Ühesõnaga, me otsime juhusliku suuruse keskväärtust konkreetsel tingimusel. Info olemasolu muudab keskväärtust, mille arvutamisel tuleb kasutada tinglikke tõenäosusi.

Definitsioon 3 *Väärtuste hulgaga $\{x_i,\ i \in I\}$ juhusliku suuruse X tinglikuks jaotuseks tingimusel, et toimus sündmus B (kus $P(B) > 0$) nimetatakse paare $(x_i, p_{i|B}),\ i \in I$, kus*

$$p_{i|B} = P(\{X = x_i\} \mid B).$$

Definitsioon 4 *Keskväärtust omava diskreetse juhusliku suuruse X tinglikuks keskväärtuseks tingimusel, et sündmus B (mille korral $P(B) > 0$) toimus, nimetatakse arvu*

$$E(X \mid B) = \sum_{i \in I} x_i p_{i|B},$$

kus $(x_i, p_{i|B}),\ i \in I$ on juhusliku suuruse X tinglik jaotus tingimusel, et B toimus.

Näide 6 *Leiame vastused eelmises näites püstitatud küsimustele:*

$$E(Y|\{X_1=2\}) = \sum_y yP\left(\{Y=y\}|\{X_1=2\}\right) = \sum_{y=3}^{8} y\cdot\frac{1}{6} = 11/2.$$

$$\begin{aligned} E\left(X_1|\{Y=5\}\right) &= \sum_x xP\left(\{X_1=x\}|\{Y=5\}\right) \\ &= \sum_x x\cdot\frac{P\left(\{X_1=x\}\cap\{Y=5\}\right)}{P\{Y=5\}} = \sum_{x=1}^{4} x\frac{1/36}{4/36} = \frac{5}{2}. \end{aligned}$$

Tõenäosusteooria rakendustes läheb sageli vaja järgnevat täistõenäosuse valemi analoogi keskväärtuse arvutamiseks.

Teoreem 1 *Olgu $B_j,\ i\in J$ (kus J on kas lõplik või loenduv hulk) sündmuste täissüsteem. Siis kehtib võrdus*

$$E(X) = \sum_{j\in J} E(X\mid B_j)P(B_j).$$

Tõestus. Alustame võrdsuse paremast poolest:

$$\begin{aligned} \sum_{j\in J} E(X|B_j)P(B_j) &\overset{def.}{=} \sum_{j\in J}\sum_{i\in I} x_i P\left(\{X=x_i\}|\,B_j\right)P(B_j) \\ &= \sum_{i\in I} x_i \sum_{j\in J} P\left(\{X=x_i\}\cap B_j\right) \\ &= \sum_{i\in I} x_i P\left(\{X=x_i\}\cap(\cup_{j\in J}B_j)\right) \text{ P aditiivsus lõikumatute } B_j \text{ korral} \\ &= \sum_{i\in I} x_i P\{X=x_i\} \text{ (sest } \cup_{j\in J} B_j=\Omega) \\ &= EX.\square \end{aligned}$$

Sageli on sobivateks sündmusteks B_j mingi teise juhusliku suuruse Y väärtustele vastavad sündmused.

Näide 7 *Lasketiirus on võimalik valida 3 püssi vahel. Olgu tiiru tulnud laskuri puhul nende püssidega märki tabamise tõenäosused ühel lasul vastavalt 0,1, 0,3 ja 0,7. Laskur valib juhuslikult püssi ja laseb 10 lasku. Olgu X tabamuste arv. Leida EX. (Lahendus loengul tahvlil.)*

2.2.2 Diskreetne juhuslik vektor

Sageli määratakse ühes katses mitme juhusliku suuruse väärtused (näiteks autojuhi vanus ning auto vanus, mõlemad täisaastates). Sellisel juhul ei aita paljude huvipakkuvate sündmuste tõenäosuste arvutamiseks nende juhuslike suuruste jaotustest, vaid on vaja informatsiooni selle kohta, kuidas need juhuslikud suurused koos käituvad. Tuletame meelde, kuida vastavat informatsiooni kirja panna ja kasutada.

Definitsioon 5 *Juhuslikku vektorit (X,Y) nimetatakse* diskreetseks*, kui X ja Y on diskreetsed juhuslikud suurused. Diskreetse juhusliku suuruse korral nimetatakse kolmikuid $(x_i,y_j,p_{ij}),\ i\in I,\ j\in J$, kus $p_{ij}=P(\{X=x_i,Y=y_j\}$ ning $\{x_i:i\in I\}$ ja $\{y_j:j\in J\}$ on vastavalt juhuslike suuruste X ja Y väärtuste hulgad, juhusliku vektori (X,Y) jaotuseks ehk juhuslike suuruste X ja Y ühisjaotuseks.*

Märkus. *Lihtsuse mõttes on eelnev definitsioon toodud kahest juhuslikust suurusest koosneva vektori kohta, kuid üldistus n-mõõtmelisele juhule on arusaadav: vektori kõik komponendid peavad olema diskreetsed juhuslikud suurused ning jaotus koosneb siis $n+1$ arvust koosnevatest komplektidest, kus esimesed n arvu on vektori komponentide võimalikud väärtused ja viimane on neile vastava tulemuste vektori saamise tõenäosus.*

Näide 8 *Kaardipakist (52 kaarti) võetakse ilma tagasipanekuta 2 kaarti, juhusliku suuruse X väärtuseks on saadud potide arv ning Y väärtuseks on saadud musta masti kaartide arv. Leiame juhusliku vektori (X,Y) jaotuse. Kuna*

$$P(\{X=0,Y=0\})=\frac{C_{26}^2}{C_{52}^2}=\frac{25}{102},\ P(\{X=0,Y=1\})=\frac{26\cdot 13}{C_{52}^2}=\frac{26}{102},$$
$$P(\{X=0,Y=2\})=\frac{C_{13}^2}{C_{52}^2}=\frac{6}{102},\ P(\{X=1,Y=1\})=\frac{13\cdot 26}{C_{52}^2}=\frac{26}{102},$$
$$P(\{X=1,Y=2\})=\frac{13\cdot 13}{C_{52}^2}=\frac{13}{102},\ P(\{X=2,Y=2\})=\frac{C_{13}^2}{C_{52}^2}=\frac{6}{102}$$

ning kõikide ülejäänud paaride tõenäosused on nullid, siis on X ja Y ühisjaotus antud tabeliga

$X\backslash Y$	*0*	*1*	*2*
0	$\frac{25}{102}$	$\frac{26}{102}$	$\frac{6}{102}$
1	*0*	$\frac{26}{102}$	$\frac{13}{102}$
2	*0*	*0*	$\frac{6}{102}$

Lugeja võib leida sellest tabelist EX (keskmine potide arv kahe võetud kaardi hulgast), EY (mustade kaartide keskmine kahe hulgast) ning $E(X|\{Y=2\})$ (keskmine potide arv tingimusel, et võetud kaartide hulgast on kaks musta kaarti).

Diskreetse juhusliku vektori jaotusest on lihtne leida komponentideks olevate juhuslike suuruste jaotusi ning nagu ikka, õigete arvutuste tunnuseks on see, et tõenäosused summeeruvad üheks. Juhusliku vektori (X,Y) käsitlemisel nimetatakse juhuslike suuruste X ja Y jaotusi *marginaaljaotusteks*. Järgnev lemma tõestati kursuses „Tõenäosusteooria ja statistika I“:

Lemma 4 *Juhusliku vektori (X,Y) jaotuse $\{(x_i,y_j,p_{ij}) : i\in I,\ j\in J\}$ korral kehtivad võrdused*

$$\sum_{i\in I}p_{ij}=P(\{Y=y_j\}),\quad \sum_{j\in J}p_{ij}=P(\{X=x_i\}),\quad \sum_{i\in I,\,j\in J}p_{ij}=1.$$

Tähistus. *Edaspidi kasutatame ühisjaotuse marginaaljaotuste tõenäosuste puhul tähistusi*

$$p_{i\cdot}=\sum_{j\in J}p_{ij},\ p_{\cdot j}=\sum_{i\in I}p_{ij}.$$

Sageli on vaja arvutada keskväärtusi juhusliku vektori funktsioonidest. Selles osas on abiks järgmine tulemus.

Teoreem 2 *Olgu X ja Y juhuslikud suurused ühisjaotusega $\{(x_i,y_j,p_{ij}) : i\in I,\ j\in J\}$ ning olgu $g:\mathbb{R}^2\to\mathbb{R}$ selline funktsioon, et juhuslik suurus $g(X,Y)$ omab lõplikku keskväärtust. Sel juhul kehtib võrdus*

$$E[g(X,Y)]=\sum_{i\in I}\sum_{j\in J}g(x_i,y_j)p_{ij}.$$

Märkus. *Sarnane tulemus kehtib ka siis, kui vaatleme funktsiooni rohkem kui kahest diskreetsest juhuslikust suurusest.*

Samuti on meil varem defineeritud diskreetsete juhuslike suuruste sõltumatus, mis ühisjaotuse kaudu ümber sõnastades on selline:

Definitsioon 6 *Diskreetseid juhuslikke suuruseid X ja Y nimetatakse sõltumatuteks, kui iga $i \in I$ ja iga $j \in J$ korral kehtib võrdus*

$$p_{ij} = p_{i\cdot} p_{\cdot j}$$

Kuna statistika kasutamisel on tegemist tavaliselt rohkem kui kahe sõltumatu juhusliku suurusega, siis sõnastema sõltumatuse mõiste ka üldkujul.

Definitsioon 7 *Diskreetseid juhuslikke suuruseid $X_1, X_2, \ldots, X_n$ nimetatakse sõltumatuteks, kui sündmused $\{X_1 = x_1\}, \ldots, \{X_n = x_n\}$ on täielikult sõltumatud iga X_1 võimaliku väärtuse x_1, X_2 võimaliku väärtuse x_2, ..., iga X_n võimaliku väärtuse x_n korral.*

Loomulikult oleks hea siinkohal järgi vaadata, mida tähendab sündmuste täielik sõltumatus.

2.3 Juhusliku vektori jaotusfunktsioon. Pidevad juhuslikud vektorid

Üldisemate juhuslike vektorite jaotusi saab kirjeldada jaotusfunktsioonide abil.

Definitsioon 8 *Juhusliku vektori (X, Y) jaotusfunktsiooniks (ehk juhuslike suuruste X ja Y ühisjaotuse jaotusfunktsiooniks) nimetatakse funktsiooni*

$$F_{X,Y}(x, y) = P(\{X \leq x,\, Y \leq y\}), \quad x, y \in \mathbb{R}.$$

Lemma 5 *(Juhusliku vektori jaotusfunktsiooni omadused). Olgu (X, Y) juhuslik vektor jaotusfunktsiooniga $F_{X,Y}$. Siis kehtivad järgnevad omadused*

1. $0 \leq F_{X,Y}(x, y) \leq 1 \ \forall (x, y) \in \mathbb{R}^2$,
2. $F_{X,Y}$ *on kummagi muutuja järgi paremalt pidev igas punktis,*
3. $\lim\limits_{y \to \infty} F_{X,Y}(x, y) = F_X(x) \ \forall x \in \mathbb{R}$, $\lim\limits_{x \to \infty} F_{X,Y}(x, y) = F_Y(y) \ \forall y \in \mathbb{R}$,
4. $\lim\limits_{y \to -\infty} F_{X,Y}(x, y) = 0 \ \forall x \in \mathbb{R}$, $\lim\limits_{x \to -\infty} F_{X,Y}(x, y) = 0 \ \forall y \in \mathbb{R}$.
5. $P(\{a < X \leq b, c < Y \leq d\}) = F_{X,Y}(b, d) - F_{X,Y}(a, d) - F_{X,Y}(b, c) + F_{X,Y}(a, c)$.

Definitsioon 9 *Juhuslikku vektorit (X, Y) nimetatakse* pidevaks*, kui tema jaotusfunktsioon avaldub kujul*

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \left(\int_{-\infty}^{y} f_{X,Y}(u, v)\, dv \right) du, \quad x, y \in \mathbb{R}$$

mingi funktsiooni $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ korral. Funktsiooni $f_{X,Y}$ nimetatakse sel juhul juhusliku vektori (X, Y) tihedusfunktsiooniks *(ehk juhuslike suuruste X ja Y* ühistiheduseks*).*

Tuletame meelde sõltumatute juhuslike suuruste üldise definitsiooni, mis sobib suvalist tüüpi juhuslike suuruste korral.

Definitsioon 10 *Juhuslikke suurusi X ja Y nimetatakse sõltumatuteks, kui iga $x, y \in \mathrm{R}$ korral on sündmused $\{X \leq x\}$ ja $\{Y \leq y\}$ sõltumatud.*

Järeldus 1 *Juhuslikud suurused on sõltumatud parajasti siis, kui nende ühisjaotuse jaotusfunktsioon avaldub kujul*

$$F_{X,Y}(x, y) = F_X(x)F_Y(y) \ \forall x, y \in \mathbb{R},$$

kus F_X ja F_Y on vastavalt juhuslike suuruste X ja Y jaotusfunktsioonid.

2.3.1 Mitmemõõtmelised pidevad jaotused

Meeldetuletus: juhuslik vektor (X, Y) on pideva jaotusega, kui tema jaotusfunktsioon avaldub kujul

$$F_{X,Y}(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(s, t)\, ds\, dt.$$

Lemma 6 *(Tihedusfunktsiooni omadused) Olgu (X, Y) pidev juhuslik vektor jaotusfunktsiooniga $F_{X,Y}$ ja tihedusfunktsiooniga $f_{X,Y}$. Siis kehtivad järgmised omadused:*

1. *Funktsioon $f_{X,Y}$ on mittenegatiivne, st $f_{X,Y}(x, y) \geq 0 \ \forall (x, y) \in \mathbb{R}^2$;*
2. *kehtivad võrdused*

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy,$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx,$$
$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$$

3. *Kui $D \subset \mathbb{R}^2$ on esitatav loenduva arvu ristkülikute abil kasutades ühendeid, ühisosasid ja täiendeid (st Boreli σ-algebra suhtes mõõtuv hulk), siis*

$$P(\{(X, Y) \in D\}) = \iint_D f_{X,Y}(x, y)\, dx\, dy.$$

4. *Kui $g : \mathbb{R}^2 \to \mathbb{R}$ on "piisavalt heade omadustega" funktsioon (nt pidev või selline, mille valemit me oskame kirja panna) ning*

$$\iint_{\mathbb{R}^2} |g(x, y)| f_{X,Y}(x, y)\, dx\, dy < \infty,$$

siis

$$E(g(X, Y)) = \iint_{\mathbb{R}^2} g(x, y) f_{X,Y}(x, y)\, dx\, dy.$$

5. *Kui $F_{X,Y}$ on diferentseeruv punktis (x, y), siis*

$$f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}}{\partial x \partial y}(x, y)$$

Lisaks oskusele tõenäosusarvutusi teha, on tähtis osata tegeleda ka sõltumatutest pidevatest juhuslikest suurustest moodustatud juhuslike vektoritega.

Lemma 7 *Pidevad juhuslikud suurused X ja Y on sõltumatud parajasti siis, kui nende ühisjaotuse tihedusfunktsioon avaldub kujul*

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) \ \forall x, y \in \mathbb{R}.$$

Tõestus. Ühtepidi: olgu X ja Y sõltumatud, siis järelduse 1 kohaselt kehtib võrdus

$$F_{X,Y}(x, y) = F_X(x) F_Y(y).$$

Kasutades jaotusfunktsiooni omadust 5, saame

$$\begin{aligned} f_{X,Y}(x, y) &= \frac{\partial^2 F_{X,Y}}{\partial x \partial y}(x, y) \\ &= \frac{\partial}{\partial x}\left(\frac{\partial}{\partial y} F_X(x) F_Y(y)\right) = f_X(x) f_Y(y). \end{aligned}$$

Teistpidi: kehtigu võrdus $f_{X,Y}(x, y) = f_X(x) f_Y(y)$, siis pideva juhusliku vektori definitsiooni kohaselt

$$\begin{aligned} F_{X,Y}(x, y) &= \int_{-\infty}^{x} \left(\int_{-\infty}^{y} f_{X,Y}(s, t)\, dt \right) ds = \int_{-\infty}^{x} f_X(s)\, ds \int_{-\infty}^{y} f_Y(t)\, dt \\ &= F_X(x) F_Y(y), \end{aligned}$$

seega juhuslikud suurused on sõltumatud.□

Eelnev tulemus on kasutatav kahte moodi:

1. kahe juhusliku suuruse sõltumatuse kindlakstegemiseks;
2. sõltumatute juhuslike suuruste ühisjaotuse tihedusfunktsiooni leidmiseks.

Näide 9 *Olgu X ja Y sõltumatud standardse normaaljaotusega juhuslikud suurused. Leiame tõenäosuse, et punkt (X, Y) satub ühikringi. Viimase lemma põhjal teame, et (X, Y) tihedusfunktsioon juhuslike suuruste X ja Y tihedusfunktsioonide korrutis; tihedusfunktsiooni omaduste põhjal saame*

$$P(X^2 + Y^2 \leq 1) = \iint\limits_{x^2+y^2 \leq 1} f_{X,Y}(x, y)\, dx\, dy.$$

Seega (kasutades üleminekut polaarkoordinaatidele)

$$\begin{aligned} P(X^2 + Y^2 \leq 1) &= \frac{1}{2\pi} \iint\limits_{x^2+y^2 \leq 1} e^{-\frac{x^2+y^2}{2}}\, dx\, dy \\ &= \frac{1}{2\pi} \int_0^{2\pi} \int_0^1 r e^{-\frac{r^2}{2}}\, dr\, d\theta = 1 - e^{-\frac{1}{2}}. \end{aligned}$$

Näide 10 *Kahemõõtmelise pideva juhusliku suuruse tihedusfunktsioon on järgmine:*

$$f(x,y) = \begin{cases} x^2 + \frac{xy}{3},\ 0 \le x \le 1,\ 0 \le y \le 2, \\ 0,\ \textit{muidu.} \end{cases}$$

Veenduda, et $f(x,y)$ on tõepoolest kahemuutuja tihedusfunktsion (lahendus tahvlil). Leida $P(X \ge 1 - Y)$ (lahendus 2. praktikumil).

Sõltumatute juhuslike suuruste puhul on mitmesuguste keskväärtuste arvutamisel kasulik järgmine tulemus.

Lemma 8 *Kui $X_i,\ i = 1, 2, \ldots, n$ on sõltumatud juhuslikud suurused ja $f_i,\ i = 1, \ldots, n$ on sellised ühe muutuja funktsioonid, et $f_i(X_i)$ on lõplikku keskväärtust omavad juhuslikud suurused, siis*

$$E[f_1(X_1) \cdot f_2(X_2) \cdots f_n(X_n)] = E[f_1(X_1)]E[f_2(X_2)] \cdots E[f_n(X_n)].$$

Lemma tulemus kehtib ka üldiste sõltumatute juhuslike suuruste puhul, kuid siin kursusel tõestame selle esialgu ainult kahe diskreetse ja kahe pideva juhusliku suuruse puhul.

Tõestus. Olgu X diskreetne juhuslik suurus võimalike väärtustega $\{x_i,\ i \in I\}$ ning Y diskreetne juhuslik suurus võimalike väärtustega $\{y_j,\ j \in J\}$, olgu nende ühisjaotus $(x_i, y_j, p_{ij}),\ i \in I,\ j \in J$. Siis valemi (2) põhjal

$$\begin{aligned} E[f_1(X)f_2(Y)] &= \sum_{i\in I}\sum_{j\in J} f_1(x_i)f_2(y_j)p_{ij} \\ &\underset{\text{sõlt.}}{=} \sum_{i\in I}\sum_{j\in J} f_1(x_i)f_2(y_j)p_{i\cdot}p_{\cdot j} \\ &= \sum_{i\in I} f_1(x_i)P(\{X = x_i\}) \sum_{j\in J} f_2(y_j)P(\{Y = y_j\}) \\ &= E[f_1(X)] \cdot E[f_2(Y)]. \end{aligned}$$

Pidevate juhuslike suuruste X korral kasutame pideva juhusliku vektori tihedusfunktsiooni omadust 4:

$$\begin{aligned} E[f_1(X)f_2(Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_1(x)f_2(y)f_{X,Y}(x,y)\,dx\,dy \\ &\underset{\text{sõlt.}}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_1(x)f_2(y)f_X(x)f_Y(y)\,dx\,dy \\ &= \int_{-\infty}^{\infty} f_1(X)f_X(x)\,dx \int_{-\infty}^{\infty} f_2(y)f_Y(y)\,dy \\ &= E[f_1(X)] \cdot E[f_2(Y)].\square \end{aligned}$$

Eelnevast tulemusest järeldub lihtsalt momente genereeriva funktsiooni kasulik omadus.

Järeldus 2 *Olgu X ja Y sõltumatud juhuslikud suurused, mille momente genereerivad funktsioonid on lõplikud nullpunkti mingis ümbruses. Siis juhusliku suuruse $Z = X + Y$ momente genereeriv funktsioon on samuti lõplik nullpunkti ümbruses ning kehtib võrdus*

$$M_Z(t) = M_X(t)M_Y(t).$$

Tõestus.

$$M_Z(t) = E(e^{Zt}) = E\left(e^{(X+Y)t}\right) = E\left(e^{Xt}e^{Yt}\right) \quad = E\left(e^{Xt}\right)E\left(e^{Yt}\right) = M_X(t)M_Y(t).$$

□

Eelneva tulemuse abil saab lihtsalt tõestada väga tihti kasutatava tulemuse sõltumatute normaaljaotusega juhuslike suuruste summa kohta.

Lemma 9 *Kui $X \sim N(\mu_1, \sigma_1)$ ja $Y \sim N(\mu_2, \sigma_2)$ on sõltumatud juhuslikud suurused, siis $Z = X + Y$ on jaotusega $N(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$ juhuslik suurus.*

Tõestus. Tõestasime 2. praktikumis (ül. 7). □

Märkus. Üsna ilmne, et eelnevat lemmat on võimalik sõnastada ka üldisemalt. Olgu X_1, X_2,..., X_n ($n \geq 2$) normaaljaotusega sõltumatud juhuslikud suurused, kusjuures $X_i \sim N(\mu_i, \sigma_i)$, $i = 1, 2, ..., n$. Ja olgu juhuslik suurus $Z = \sum_{i=1}^{n} X_i$. Siis juhuslik suurus Z on samuti normaaljaotusega:

$$Z \sim N\left(\sum_{i=1}^{n} \mu_i, \sqrt{\sum_{i=1}^{n} \sigma_i^2}\right).$$

Näeme, et normaaljaotus rahuldab nn *reproduktiivsuse omadust*: kui liita kaks (või rohkem) kindla jaotusega juhuslikku suurust, siis tulemusena saadud juhuslik suurus on sama tüüpi jaotusega, mis liidetavad. Selliseid jaotusi on veel. Kodutöös 2 näitasime, et sama omadusega on ka Poissoni jaotus. Hiljem näitame, et ka hii-ruut-jaotusel on sama omadus.

2.3.2 Sõltumatute pidevate juhuslike suuruste summa ja jagatise jaotus

Sageli esineb praktikas situatsioon, kus huvipakkuv juhuslik suurus avaldub sõltumatute juhuslike suuruste summana. Osutub, et sel juhul avaldub summa tihedusfunktsioon liidetavate tihedusfunktsioonide konvolutsiooni kujul.

Lemma 10 *Olgu X ja Y pidevad sõltumatud juhuslikud suurused tihedusfunksioonidega f_X ja f_Y. Sel juhul juhusliku suuruse $Z = X + Y$ tihedusfunktsioon avaldub kujul*

$$f_Z(z) = \int_{-\infty}^{\infty} f_Y(y) f_X(z - y)\, dy, \quad z \in \mathbb{R}.$$

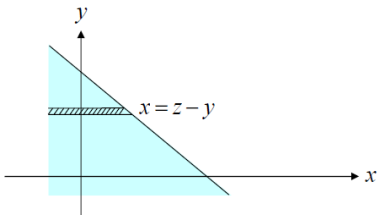
Tõestus. Meil on vaja leida juhusliku suuruse Z tihedusfunktsiooni $f_Z(z)$, mille on võimalik tuletada selle juhusliku suuruse jaotusfunktsioonist $F_Z(z)$. Leiame esmalt jaotusfunktsiooni:

$$F_Z(z) \overset{Def.}{=} P(Z \leq z) = P(X + Y \leq z) = P(X \leq z - Y)$$

Lemma (6) põhjal leiame

$$\begin{aligned} P(X \leq z - Y) &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{z-y} f_{X,Y}(x, y)\, dx \right) dy \\ &\underset{\text{sõlt.}}{=} \int_{-\infty}^{\infty} \left(\int_{-\infty}^{z-y} f_X(x) f_Y(y)\, dx \right) dy = \int_{-\infty}^{\infty} f_Y(y) \left(\int_{-\infty}^{z-y} f_X(x)\, dx \right) dy \\ &= \int_{-\infty}^{\infty} f_Y(y) F_X(z - y)\, dy \end{aligned}$$

Siin on integreerimisel abiks järgmine joonis:



Integraaliteooria tulemuste põhjal (näiteks Fubini-Tonelli teoreemi rakendusena) saab näidata, et eelnevat integraali võib diferentseerida z järgi integraalimärgi all, mistõttu saame

$$f_Z(z) = F'_Z(z) = \int_{-\infty}^{\infty} \frac{\partial}{\partial z} \left(f_Y(y) F_X(z-y)\right)\, dy$$
$$= \int_{-\infty}^{\infty} f_Y(y) f_X(z-y)\, dy$$

Sellega on lemma tõestatud. □

Märkus. Eelmises lemmas on võimalik kasutada alternatiivset viisi:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx, \quad z \in \mathbb{R}.$$

(*Harjutus lugejale*).

Näide 11 *Olgu $X \sim U(0,1)$ ja $Y \sim U(0,1)$ kaks sõltumatut juhusliku suurust. Leida juhusliku suuruse $Z = X + Y$ tihedusfunktsioon.*
Ühtlase jaotuse kohta teame, et kui $X \sim U(0,1)$, siis $f_X(x) = 1$ kui $0 \le x \le 1$ ja 0 vastasel juhul. Analoogiliselt ka juhusliku suurusega Y. Eelneva lemma põhjal saame:

$$f_Z(z) = \int_{-\infty}^{\infty} f_Y(y) f_X(z-y)\, dy$$
$$= \int_0^1 f_X(z-y)\, dy$$

Näeme, et integraali väärtus on 0 väljaspool lõiku $[0,\ 1]$ (juhusliku suuruse X definitsiooni tõttu), mille tõttu huvitume piirkonnast $0 \le z - y \le 1$, mis omakorda samaväärne piirkonnaga $z - 1 \le y \le z$.
Integraali leidmiseks vaatame erinevaid võimalusi z muutumiseks $y \in [0,1]$ suhtes.
1) Kui $0 \le z \le 1$, siis $f_Z(z) = \int_0^z dy = z$.
2) Kui $1 < z \le 2$, siis $f_Z(z) = \int_{z-1}^1 dy = 2 - z$.
3) Kui $z < 0$ või $z > 2$, siis $f_Z(z) = 0$.

Seega,

$$f_Z(z) = \begin{cases} z, & \text{kui } 0 \le z \le 1, \\ z - 2, & \text{kui } 1 < z \le 2, \\ 0, & \text{vastasel juhul.} \end{cases}$$

Lemma 11 *Olgu X ja Y sõltumatud pidevad juhuslikud suurused tihedusfunktsioonidega f_X ja f_Y. Siis juhusliku suuruse $Z = \frac{X}{Y}$ tihedusfunktsioon avaldub kujul*

$$f_Z(z) = \int_{-\infty}^{\infty} |y|\, f_X(zy) f_Y(y)\, dy.$$

Tõestus. Kuna X ja Y on sõltumatud pidevad juhuslikud suurused, siis vektor (X, Y) on pidev juhuslik vektor tihedusfunktsiooniga $f_{X,Y}(x, y) = f_X(x) f_Y(y)$. Leiame juhusliku suuruse Z jaotusfunktsiooni avaldise. Definitsiooni kohaselt

$$F_Z(z) = P\left(\left\{\frac{X}{Y} \le z\right\}\right).$$

Selleks, et leida vastavat tasandi piirkonda (x, y)-tasandil, on kasulik jagatisest lahti saada. Siin aga tuleb arvestada, et Y-ga läbi korrutades sõltub tekkiva võrratuse märk Y märgist, mistõttu saame kirjutada

$$\left\{\frac{X}{Y} \le z\right\} = (\{Y > 0\} \cap \{X \le zY\}) \cup (\{Y < 0\} \cap \{X \ge zY\}).$$

Kasutades nüüd tõenäosuse lõplik-aditiivsust (kuna ühend on leitud teineteist välistavatest sündmustest), saame

$$F_Z(z) = P(\{Y > 0\} \cap \{X \le zY\}) + P(\{Y < 0\} \cap \{X \ge zY\}).$$

Lemma (6) põhjal leiame

$$\begin{aligned} P(\{Y > 0\} \cap \{X \le zY\}) &= \int_0^{\infty} \left(\int_{-\infty}^{zy} f_{X,Y}(x, y)\, dx \right) dy \\ &\underset{\text{sõlt.}}{=} \int_0^{\infty} \left(\int_{-\infty}^{zy} f_X(x) f_Y(y)\, dx \right) dy = \int_0^{\infty} F_X(zy) f_Y(y)\, dy, \\ P(\{Y < 0\} \cap \{X \ge zY\}) &= \int_{-\infty}^{0} \left(\int_{zy}^{\infty} f_{X,Y}(x, y)\, dx \right) dy \\ &\underset{\text{sõlt.}}{=} \int_{-\infty}^{0} \left(\int_{zy}^{\infty} f_X(x) f_Y(y)\, dx \right) dy = \int_{-\infty}^{0} (1 - F_X(zy)) f_Y(y)\, dy. \end{aligned}$$

Integraaliteooria tulemuste põhjal (näiteks Fubini-Tonelli teoreemi rakendusena) saab näidata, et eelnevaid integraale võib diferentseerida z järgi integraalimärgi all, mistõttu saame

$$\begin{aligned} f_Z(z) = F_Z'(z) &= \int_0^{\infty} y f_X(zy) f_Y(y)\, dy - \int_{-\infty}^{0} y f_X(zy) f_Y(y)\, dy \\ &= \int_{-\infty}^{\infty} |y| f_X(zy) f_Y(y)\, dy. \end{aligned}$$

Sellega on lemma tõestatud. □

2.4 Täiendavaid teadmisi kovariatsioonidest ja korrelatsioonidest.

Mitme juhusliku suuruse korral pakub enamasti huvi nendevahelise seose olemasolu. Sageli on võimalik raskesti mõõdetavaid juhuslikke suuruseid teiste, lihtsamini või odavamalt

mõõdetavate abil prognoosida või siis erinevaid juhuslikke suurusi sobivalt kombineerides riske maandada.

Juhuslike suuruste vahel võib olla nii lineaarseid kui mittelineaarseid seoseid. Näiteks võivad juhuslikud suurused X, Y, Z olla omavahel seotud võrdusega

$$Z = X^2 \cdot \cos(Y),$$

mille korral on tegemist mittelineaarse seosega. Samas seos

$$Z = 0,4\,X - 0,6Y$$

on lineaarne seos. Lineaarseid seoseid on lihtsam uurida ning järgnevas vaatleme neid lähemalt.

Lineaarse seose olemasolu kindlakstegemise seisukohalt on tähtsad mõisted kovariatsioon ja korrelatsioonikordaja. Eelmisest kursusest teame, et lõplikke dispersioone omavate juhuslike suuruste X ja Y kovariatsioon on defineeritud võrdusega

$$\operatorname{cov}(X, Y) = E[(X - EX)(Y - EY)]$$

ning et seda saab arvutada ka kujul

$$\operatorname{cov}(X, Y) = E[XY] - EX \cdot EY.$$

Kovariatsioon sõltub aga ühikutest, milles me X ja Y väärtust väljendame. Näiteks kui me uurime inimese pikkuse ja kaalu vahelist seost, siis tuleb kovariatsioon erinev sõltuvalt sellest, kas kaalu mõõdetakse kilogrammides või grammides, samas mõõtühikud ei tohiks mõjutada juhuslike suuruste omavahelise sõltuvuse tugevust. Osutub, et üheks heade omadustega sõltuvuse mõõdikuks on Pearsoni korrelatsioonikordaja.

Definitsioon 11 *Pearsoni korrelatsioonikordajaks kahe lõpliku dispersiooniga juhusliku suuruse X ja Y vahel nimetatakse arvu*

$$\rho_{X,Y} = \frac{\operatorname{cov}(X, Y)}{\sqrt{DX\,DY}}.$$

Oluline on aru saada, et Pearsoni korrelatsioonikordaja mõõdab ainult lineaarset sõltuvust (kui hästi on Y väärtus lähendatav suurusega kujul $aX + b$ mingi a ja b korral) ning ei ütle midagi teistsuguste sõltuvuste kohta.

Näide 12 *Olgu $X \sim B(2, \frac{1}{2})$ ning olgu $Y = |X - 1|$. Otseste arvutustega on lihtne näha, et $\operatorname{cov}(X, Y) = 0$ ning seega ka $\rho_{X,Y} = 0$, kuid Y on selgelt sõltuv X-st (kui X väärtust teame, on ka Y väärtus teada).*

Kovariatsiooni ja korrelatsiooni tähtsamad omadused (koos varem tõestatutega) on kokku võetud järgnevas lemmas.

Lemma 12 *Olgu X, Y ja Z lõplikku dispersiooni omavad juhuslikud suurused. Siis kehtivad valemid*

1. $\operatorname{cov}(X, X) = DX$;
2. $\operatorname{cov}(X, Y) = E(XY) - EX \cdot EY$;

3. $D(X+Y) = DX + DY + 2\mathrm{cov}(X,Y)$;
   $D(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} DX_i + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathrm{cov}(X_i, X_j)$;
4. $\mathrm{cov}(X,Y) = \mathrm{cov}(Y,X)$.
5. $\mathrm{cov}(\alpha\, X + \beta\, Y, Z) = \alpha\, \mathrm{cov}(X,Z) + \beta\, \mathrm{cov}(Y,Z) \quad \forall \alpha, \beta \in \mathbb{R}$;
6. *kui* X *ja* Y *on sõltumatud, siis* $\mathrm{cov}(X,Y) = 0$ *ning* $D(X+Y) = DX + DY$.
7. $|\mathrm{cov}(X,Y)| \le \sqrt{DX \cdot DY}$
8. $-1 \le \rho_{X,Y} \le 1$
9. $\rho_{X,Y} = 1$, *kui* $Y = aX + b$ *mingite reaalarvude* a, b, *kus* $a > 0$ *korral; kui* $Y = aX + b$ *negatiivse* a *korral, siis* $\rho_{X,Y} = -1$.

Tõestus. Omadus 1 järeldub otse kovariatsiooni ja dispersiooni definitsioonidest. Omadused 2 ja 3 on tõestatud eelnevas kursuses.

$$\begin{aligned}\mathrm{cov}(X,Y) &= E[(X - EX)(Y - EY)] = E(XY) - E[X\,EY] - E[Y\,EX] + EX\,EY \\ &= E(XY) - EX\,EY.\end{aligned}$$

Omadus 3 tuleneb dispersiooni definitsioonist ja keskväärtuse lineaarsusest:

$$\begin{aligned}D(X+Y) &= E[(X + Y - EX - EY)^2] \\ &= E[(X - EX)^2 + 2(X - EX)(Y - EY) + (Y - EY)] \\ &= DX + 2\mathrm{cov}(X,Y) + DY.\end{aligned}$$

Omadus 4 tuleneb otse definitsioonist, omadus 5 keskväärtuse lineaarsusest. Omadus 6 on varem tõestatud.

Omaduse 7 näitamiseks kasutame teadmist, et dispersioon on alati mittenegatiivne. Seega kehtib iga reaalarvu a korral võrratus

$$D(X + aY) \ge 0.$$

Summa dispersiooni omadusest saame nüüd

$$DX + D(aY) + 2\mathrm{cov}(X, aY) \ge 0,$$

kust kovariatsiooni lineaarsust ja dispersiooni omadust $D(aY) = a^2 DY$ kasutades saame

$$DX + a^2 DY + 2a\mathrm{cov}(X,Y) \ge 0.$$

Kui $DY = 0$, siis on selle võrratuse vasakul pool tegemist a suhtes lineaarse funktsiooniga ja see saab olla alati mittenegatiivne ainult siis, kui a kordaja on null ning seega omadus 7 kehtib. Kui aga $DY \neq 0$, siis saame valida $a = -\frac{\mathrm{cov}(X,Y)}{DY}$ (mis minimiseerib võrratuse vasakul pool olevat avaldist a suhtes) ning saame

$$DX + \frac{\mathrm{cov}(X,Y)}{DY} - 2\frac{\mathrm{cov}(X,Y)^2}{DY} \ge 0,$$

kust järeldub

$$\mathrm{cov}(X,Y)^2 \le DX \cdot DY.$$

Sellest võrratusest ruutjuurt võttes saamegi nõutud võrratuse.

Omadus 8 tuleneb nüüd otse korrelatsiooni definitsioonist ja omadusest 7. Omaduse 9 tõestus jääb lugejale harjutuseks.□.

Seos 6 väidab, et sõltumatute juhuslike suuruste kovariatsioon on 0. Vastupidine üldiselt ei kehti: **kovariatsioon võib olla ka 0 siis, kui juhuslikud suurused on sõltuvad**.

Näide 13 *Olgu (X, Y) jaotustabel järgmine:*

$Y \setminus X$	-1	0	1
0	$\frac{1}{3}$	0	$\frac{1}{3}$
1	0	$\frac{1}{3}$	0

Veendu, et X ja Y on sõltuvad, $E(XY) = 0$ ja $EX = 0$. Seega, $cov(X, Y) = 0$.

Näide 14 *Olgu (X, Y) ühtlase jaotusega ringil raadiusega R, st ühistihedus on*

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{\pi R^2}, \text{ kui } x^2 + y^2 \leq R^2; \\ 0, \text{ mujal.} \end{cases}$$

Veendu, et $EX = EY = E(XY) = 0$, kuid X ja Y pole sõltumatud.

2.4.1 Juhusliku vektori keskväärtus ja kovariatsioonimaatriks

Eelmises kursuses defineerisime juhusliku valimi läbi juhusliku vektori $(X_1,\ X_2, ...,\ X_n)$ sõltumatute elementidega, kus iga element X_i vastab antud valimi ühele elemendile. Näiteks, kui uuritavaks tunnuseks on inimese palk, siis on X_5 valimisse sattunud 5. inimese palganäitaja. Kuna inimese sattumine/mittesattumine valimisse on juhuslik, siis ka palganäitaja on juhusliku loomuga. Viienda inimesena võib sattuda ükskõik milline inimene vaadeldavast üldkogumist ja seega ka viienda inimese palk X_5 on juhuslik suurus. Kui inimeste valik toimub üksteisest sõltumatult, siis ka komponendid X_i ja X_j, $i \neq j$ on sõltumatud.

Sageli uuritakse statistikas valimielemente ühe tervikuna, st moodustab vektor $\mathbf{X} = (X_1,\ X_2, ...,\ X_n)^T$ n-mõõtmelist juhuslikku suurus, kus sümbol T tähendab transponeerimist. Edaspidi defineerime juhusliku vektori keskväärtust ja kovariatsiooni(maatriksit).

Definitsioon 12 *Juhusliku vektori $\boldsymbol{X} = (X_1,\ X_2,\ ...,\ X_n)^T$ keskväärtus on vektor*

$$E\boldsymbol{X} = (EX_1,\ EX_2\ , ...,\ EX_n)^T.$$

Definitsioon 13 *Juhusliku vektori $\boldsymbol{X} = (X_1,\ X_2,\ ...,\ X_n)^T$ kovariatsioonimaatriks (ka dispersioonimaatriks) on järgmine $n \times n$ sümmeetriline maatriks:*

$$D(\boldsymbol{X}) = \begin{pmatrix} DX_1 & cov(X_1,\ X_2) & cov(X_1,\ X_3) & \dots & cov(X_1,\ X_n) \\ cov(X_2,\ X_1) & DX_2 & cov(X_2,\ X_3) & \dots & cov(X_2,\ X_n) \\ cov(X_3,\ X_1) & cov(X_3,\ X_2) & DX_3 & \dots & cov(X_3,\ X_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ cov(X_n,\ X_1) & cov(X_n,\ X_2) & cov(X_n,\ X_3) & \dots & DX_n \end{pmatrix}$$

Definitsioon 14 *Juhusliku vektori $\boldsymbol{X} = (X_1,\ X_2,\ ...,\ X_n)^T$ korrelatsioonimaatriks on järgmine $n \times n$ sümmeetriline maatriks:*

$$\rho(\boldsymbol{X}) = \begin{pmatrix} 1 & \rho(X_1,\ X_2) & \rho(X_1,\ X_3) & \dots & \rho(X_1,\ X_n) \\ \rho(X_2,\ X_1) & 1 & \rho(X_2,\ X_3) & \dots & \rho(X_2,\ X_n) \\ \rho(X_3,\ X_1) & \rho(X_3,\ X_2) & 1 & \dots & \rho(X_3,\ X_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho(X_n,\ X_1) & \rho(X_n,\ X_2) & \rho(X_n,\ X_3) & \dots & 1 \end{pmatrix}$$

Lemma 13 *Kovariatsiooni- ja korrelatsioonimaatriksil on järgmised omadused:*

1. *Kovariatsioonimaatriks ja korrelatsioonimaatriks on sümmeetrilised.*
2. *Kui vektori* $\boldsymbol{X} = (X_1,\ X_2,\ ...,\ X_n)^T$ *komponendid on sõltumatud, on kovariatsioonimaatriks diagonaalmaatriks ja korrelatsioonimaatriks ühikmaatriks.*
3. *Kui* $\boldsymbol{\Sigma}$ *on vektori* $\boldsymbol{X} = (X_1,\ X_2,\ ...,\ X_n)^T$ *kovariatsioonimaatriks ja* $\boldsymbol{a} = (a_1,\ a_2, ...,\ a_n)^T$ *on suvaline n-dimensionaalne konstantne vektor, siis*

$$\boldsymbol{a}^t\boldsymbol{\Sigma}\boldsymbol{a} = D(a_1X_1 + a_2X_2 + ... + a_nX_n) \geq 0.$$

Tõestus. Omadused 1 ja 2 on ilmsed. Kolmanda väite tõestamisel kasutame Lemma 12 väiteid 1 ja 3:

$$\boldsymbol{a}^T\boldsymbol{\Sigma}\boldsymbol{a} = (a_1,\ a_2, ...,\ a_n)\begin{pmatrix} 1 & \rho(X_1,\ X_2) & \rho(X_1,\ X_3) & \dots & \rho(X_1,\ X_n) \\ \rho(X_2,\ X_1) & 1 & \rho(X_2,\ X_3) & \dots & \rho(X_2,\ X_n) \\ \rho(X_3,\ X_1) & \rho(X_3,\ X_2) & 1 & \dots & \rho(X_3,\ X_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho(X_n,\ X_1) & \rho(X_n,\ X_2) & \rho(X_n,\ X_3) & \dots & 1 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \\ \dots \\ a_n \end{pmatrix}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} a_ia_j\ cov(X_i,\ X_j) = \sum_{i=1}^{n} a_i^2DX_i + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_ia_j\ \mathrm{cov}(X_i, X_j) = D(\sum_{i=1}^{n} a_iX_i),$$

mis on tõepoolest suurem või võrdne nulliga. □

Näide 15 *Olgu juhusliku vektori* $(X,\ Y)$ *jaotustabel järgmine*

$X \setminus Y$	*0*	*1*	*2*
0	$\frac{1}{9}$	$\frac{1}{3}$	$\frac{1}{4}$
1	$\frac{1}{9}$	$\frac{1}{6}$	*0*
2	$\frac{1}{36}$	*0*	*0*

Veenduda, et $X \sim B(2, \frac{1}{6})$ *ja* $Y \sim B(2, \frac{1}{2})$. *Leiame vektori* $(X,\ Y)^T$ *keskväärtuse ja kovariatsioonimaatriksi.*

Selleks peame teadma $cov(X,\ Y) = E(XY) - EX\ EY$, *kusjuures*

$$E(XY) = \sum_{i=0}^{2}\sum_{j=0}^{2} i \cdot j \cdot p_{ij} = \frac{1}{6}.$$

Binoomjaotuse keskväärtus on teada eelmisest kursusest, $EX = n \cdot p = \frac{1}{3}$ *ja* $EY = 1$, *millest*

$$cov(X,\ Y) = \frac{1}{6} - \frac{1}{3} \cdot 1 = -\frac{1}{6}.$$

Samuti $DX = np(1-p) = \frac{1}{3}\frac{5}{6} = \frac{5}{18}$ *ning* $DY = \frac{1}{2}$. *Järelikult, vektori* $(X,\ Y)^T$ *keskväärtus on* $(\frac{1}{3},\ 1)^T$ *ja kovariatsioonimaatriks on*

$$\begin{pmatrix} \frac{5}{18} & -\frac{1}{6} \\ -\frac{1}{6} & \frac{1}{2} \end{pmatrix}$$

Leiame $D(X - Y)$ *kovariatsioonimaatriksi abil:*

$$D(X - Y) = (1,\ -1)\begin{pmatrix} \frac{5}{18} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{1}{2} \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = \frac{5}{18} + \frac{1}{3} + \frac{1}{3} + \frac{1}{2} = \frac{13}{9}.$$

2.4.2 Mitmemõõtmeline normaaljaotus

Lisaks mitmesuguste juhuslike vektorite ühisjaotustele on praktikas väga sageli kasutatavaks jaotuseks mitmemõõtmeline normaaljaotus. Toome ära mitmemõõtmelise normaaljaotuse definitsiooni.

Definitsioon 15 *Juhuslik vektor* $\boldsymbol{X} = (X_1,\ X_2,\ ...,\ X_n)^T$ *on mitmemõõtmelise normaaljaotusega, kui tema tihedusfunktsioon avaldub kujul*

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}} exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right),$$

kus $\boldsymbol{x} = (x_1, ..., x_n) \in \mathcal{R}^n$, $\boldsymbol{\mu} \in \mathcal{R}^n$ *ja* $\boldsymbol{\Sigma}$ *on* $n \times n$ *sümmeetriline positiivselt poolmääratud maatriks. Vektor* $\boldsymbol{\mu}$ *ja maatriks* $\boldsymbol{\Sigma}$ *on mitmemõõtmelise normaaljaotuse parameetrid ja nende parameetritega mitmemõõtmelist normaaljaotust tähistatakse* $N(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$.

Märkus. Ühemõõtmelisel juhul oleme normaaljaotuse teiseks parameetriks kasutanud juhusliku suuruse standardhälvet, näiteks $X \sim N(\mu, \sigma)$ (õpikus Pärna, 2013). Sageli kasutatakse teise parameetrina juhusliku suuruse dispersiooni $X \sim N(\mu, \sigma^2)$ (õpikus Traat, 2006). Mitmemõõtmelise normaaljaotuse tähistuses on siiski levinum teine variant, ehk teise parameetri rollis on kovariatsioonimaatriks $\boldsymbol{\Sigma}$.

Näide 16 ***Kahemõõtmeline normaaljaotus***. *Vaatleme kahemõõtmelist juhusliku vektorit* $(X,\ Y)^T$. *Olgu* $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ *ja*

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix},$$

kus $\sigma_x^2 = DX$, $\sigma_y^2 = DY$ *ja* $\sigma_{xy} = cov(X,\ Y)$. *Siis*

$$|\boldsymbol{\Sigma}| = \sigma_x^2\sigma_y^2 - (\sigma_{xy})^2 = \sigma_x^2\sigma_y^2\left(1 - \frac{(\sigma_{xy})^2}{\sigma_x^2\sigma_y^2}\right) = \sigma_x^2\sigma_y^2(1-\rho^2),$$

kus $\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$ *on korrelatsioonikordaja* X *ja* Y *vahel. Leiame veel pöördmaatriksi* $\boldsymbol{\Sigma}^{-1}$:

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_x^2\sigma_y^2 - (\sigma_{xy})^2}\begin{pmatrix} \sigma_y^2 & -\sigma_{xy} \\ -\sigma_{xy} & \sigma_x^2 \end{pmatrix} = \frac{1}{1-\rho^2}\begin{pmatrix} \frac{1}{\sigma_x^2} & -\frac{\rho}{\sigma_x\sigma_y} \\ -\frac{\rho}{\sigma_x\sigma_y} & \frac{1}{\sigma_y^2} \end{pmatrix}$$

Seega kahemõõtmelise normaaljaotuse tihedusfunktsioon on

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_x\sigma_y} + \frac{(y-\mu_2)^2}{\sigma_y^2}\right)\right]. \tag{2.1}$$

Eelnevast teame, et kui kovariatsioon kahe juhusliku suuruse vahel on 0, siis ei tähenda see veel juhuslike suuruste sõltumatust. Küll aga kehtib vastupidine seos: X ja Y on sõltumatud, siis $cov(X,\ Y) = 0$. Mitmemõõtmeline normaaljaotus on selles mõttes eriline: väide kehtib mõlemas suunas (vt. järgmist lemmat).

Lemma 14 *Mitmemõõtmelise normaaljaotusega vektori* $\boldsymbol{X} = (X_1, X_2, ..., X_N)^T$ *komponendid on sõltumatud parajasti siis, kui nendevahelised kovariatsioonid* $(cov(X_i,\ X_j),\ j \neq j)$ *on kõik nullid. Kovariatsioonimaatriks on sel juhul diagonaalne (väljaspool peadiagonaali on kõik nullid).*

Tõestame loengul kahemõõtmelisel juhul.

Mitmemõõtmelise normaaljaotuse lisaomadused:

1. Sõltumatud normaaljaotusega juhuslikud suurused moodustavad mitmemõõtmelise normaaljaotusega vektori (kovariatsioonimaatriks on diagonaalne).

2. Olgu $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ja $\mathbf{D}: \ l \times n$ maatriks astakuga $l \leq n$ ning olgu $\mathbf{Y} = \mathbf{D}\ \mathbf{X}: \ l \times 1$ (juhuslik vektor, mis on saadud elementide X_i lineaarsete kombinatsioonide abil). Siis $\mathbf{Y} \sim N(\mathbf{D}\boldsymbol{\mu},\ \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^T)$.
   Sellest omadusest järldub:

   (a) Kui $\mathbf{D} = \boldsymbol{d} = (d_1, d_2, ..., d_n): \ 1 \times n$ reavektor, siis $\mathbf{Y} = \boldsymbol{d}\mathbf{X} = \sum_{i=1}^{n} d_i X_i$. See tähendab, et mitmemõõtmelise normaaljaotuse korral vektori elementide lineaarne kombinatsioon on samuti normaaljaotusega.

   (b) Kui $\mathbf{D} = \boldsymbol{d} = (d_1, d_2, ..., d_k, 0, 0, ..., 0): \ 1 \times n$, mille esimest k komponenti erineb nullist ja ülejäänud on nullid, siis $\mathbf{Y} = \boldsymbol{d}\mathbf{X} = \sum_{i=1}^{k} d_i X_i$. See tähendab, et mitmemõõtmelise normaaljaotuse korral vektori elementide suvalise alamhulga lineaarne kombinatsioon on samuti normaaljaotusega.

   (c) Saame alati valida $\mathbf{D}$ nii, et see "võtaks"vektorist $\mathbf{X}$ suvalise alamhulga elemente, näiteks kui $\mathbf{X} = (X_1, X_2, X_3)^T$ ja

$$\mathbf{D} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix},$$

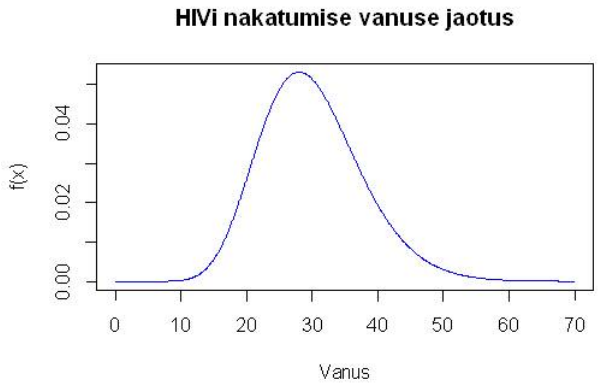
   siis $\mathbf{Y} = \mathbf{D}\ \mathbf{X} = (X_1, X_2)^T$. See tähendab, et mitmemõõtmelise normaaljaotusega vektori suvaline alamhulk on samuti normaaljaotusega.

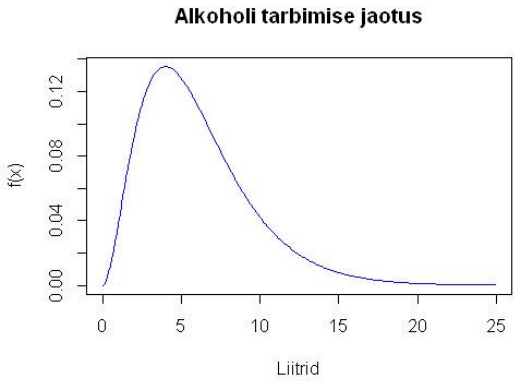
2.5 Kolm tähtsat pidevat jaotust statistikas ja seosed nende vahel

Järgnev peatükk põhineb õpikul Traat (2006) ja käsitleb χ^2-, t- ning F-jaotust. Nendel jaotustel on tähtis roll vahemikhinnangute leidmisel ja hüpoteeside kontrollimisel. Seetõttu uurime siin neid jaotusi põhjalikumalt kasutades eelnevalt saadud teadmisi pidevate jaotuste kohta.

2.5.1 χ^2-jaotus (Hii-ruut-jaotus)

χ^2-jaotust kasutatakse vahemikhinnangute ja hüpoteeside kontrolli ülesannetes, mis on seotud üldkogumi dispersiooni- või standardhälbega. Samuti on sel tähtis roll nn t-jaotuse moodustamisel (vt. järgmine alampeatükk). Osutub, et ka mõned nähtused on samuti hii-ruut jaotusega:



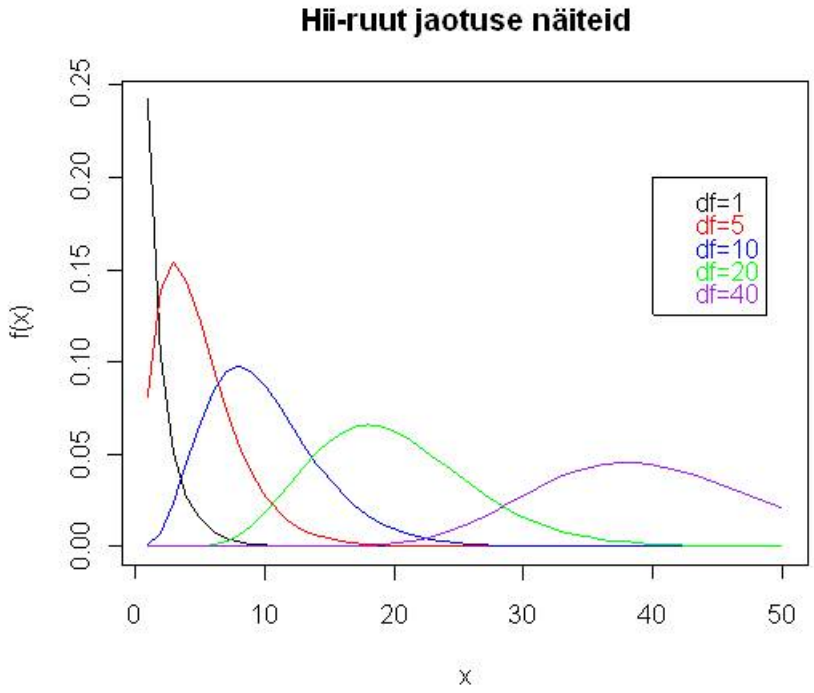


Definitsioon 16 *Juhuslik suurus X on hii-ruut jaotusega parameetriga (ka vabadusastmete arvuga) f, $X \sim \chi^2(f)$, kui tema tihedusfunktsioon avaldub seosega*

$$f(x) = kx^{\frac{f}{2}-1}e^{-\frac{x}{2}}, \quad x \geq 0, \tag{2.2}$$

kus $k = \frac{2^{-\frac{f}{2}}}{\Gamma(\frac{f}{2})}$ on normeeriv konstant, vabadusastmete arv $f \in \mathbb{N}$ on jaotuse parameeter ja $\Gamma(y) = \int_0^\infty t^{y-1}e^{-t}dt, y \in \mathbb{R}^+$ on gammafunktsioon.

Teadmiseks, et $\Gamma(y) = (y-1)\Gamma(y-1), y > 1$ ning $\Gamma(1) = 1$ ja $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.
Järgmisel joonisel on toodud χ^2-jaotuse tihedusfunktsioonid erinevate parameetri f väärtuste korral. Paneme tähele, et parameetri väärtuse kasvades muutub tihedusfunktsioon "sümmeetrilisemaks".



Lemma 15 *Erijuhul, kui $f = 1$, siis on hii-ruut jaotuse tihedusfunktsiooniks:*

$$f(x) = \frac{1}{\sqrt{2\pi x}} e^{-\frac{x}{2}} \tag{2.3}$$

Tõestus on harjutus lugejale.

Lemma 16 *(Hii-ruut-jaotuse m.g.f. ja esimesed momendid.) Olgu juhuslik suurus X hii-ruut-jaotusega, $X \sim \chi^2(f)$, $f > 0$. Siis juhusliku suuruse X momente genereeriv funktsioon on*

$$M(t) = (1-2t)^{-f/2}, \quad t < \frac{1}{2}. \tag{2.4}$$

ning keskväärtus ja dispersioon on vastavalt

$$EX = f, \quad DX = 2f.$$

Tõestus. Juhusliku suuruse m.g.f. tuletuskäik on 4. praktikumi ülesanne. Jaotuse momendid on leitavad valemist:

$$EX^k = \frac{d^k M(t)}{dt^k}|_{t=0}.$$

Keskväärtus:

$$EX = M'(t)\big|_{t=0} = -\frac{f}{2}(1-2t)^{-\frac{f}{2}-1}(-2)\Big|_{t=0} = f(1-2t)^{-\frac{f}{2}-1}\Big|_{t=0} = f.$$

Dispersiooni saame avaldisest $DX = EX^2 - (EX)^2$,

$$EX^2 = M''(t)\big|_{t=0} = f\left(-\frac{f}{2}-1\right)(1-2t)^{-f/2-2}(-2)\Big|_{t=0} = f^2 + 2f,$$

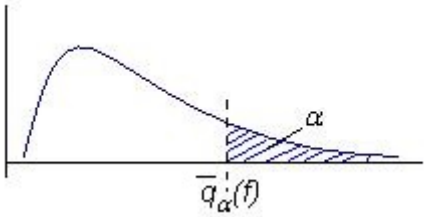
millest $DX = f^2 + 2f - f^2 = 2f$. □

Rakendusülesannetes läheb edaspidi vaja hii-ruut-jaotusega juhusliku suuruse täiendkvantiili väärtuseid. Tuletame meelde siinkohal pideva juhusliku suuruse täiendkvantiili mõistet:

Definitsioon 17 *Arvu $\bar{q}_\alpha$ nimetatakse pideva juhusliku suuruse α-täiendkvantiiliks kui kehtib järgmine seos:*

$$P(X \geq \bar{q}_\alpha) = \alpha. \tag{2.5}$$

Täiendkvantiili mõistet iseloomustab ka järgmine joonis:



Lisas A on ära toodud hii-ruut-jaotuse täiendkvantiilide tabel levinute α väärtuste korral.

Lemma 17 *Suure f korral on hii-ruut jaotus lähendatav normaaljaotusega:*

$$\chi^2(f) \approx N(f, \sqrt{2f}).$$

Teoreem 3 *(Hii-ruut-jaotuse aditiivsus.) Olgu $X_1, X_2, ..., X_n$ sõltumatud juhuslikud suurused jaotusega vastavalt $\chi^2(f_1), \chi^2(f_2), \ldots, \chi^2(f_n)$. Siis*

$$Y = \sum_{i=1}^{n} X_i \sim \chi^2(\sum_{i=1}^{n} f_i).$$

Tõestus Kasutame Y momente genereerivat funktsiooni, millest sõltumatust arvestades saame

$$M_Y(t) = Ee^{tY} = Ee^{t\sum X_i} = Ee^{tX_1} \cdot Ee^{tX_2} \cdot \ldots \cdot Ee^{tX_n}.$$

Kuna

$$X_i \sim \chi^2(f_i) \Rightarrow Ee^{tX_i} = (1-2t)^{-f_i/2}, \quad t < \frac{1}{2},$$

siis

$$M_Y(t) = (1-2t)^{-\sum f_i/2}, \quad t < \frac{1}{2}.$$

Saime $\chi^2(\sum f_i)$ momente genereeriva funktsiooni, seega

$$Y \sim \chi^2(\sum_{i=1}^{n} f_i). \ \square$$

Teoreem 4 *(Hii-ruut-jaotuse seos standardse normaaljaotusega.) Kui $X_1, X_2, ..., X_n$ on sõltumatud juhuslikud suurused, kus $X_i \sim N(0,1),\ i = 1, \ldots, n$, siis*

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n), \tag{2.6}$$

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 \sim \chi^2(n-1), \tag{2.7}$$

kus $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Tõestus. Esmalt leiame ühe komponendi X_i^2 jaotuse. Alustades jaotusfunktsioonist, saame

$$F_{X_i^2}(x) = P(X_i^2 \le x) = P(-\sqrt{x} \le X_i \le \sqrt{x}) = \Phi(\sqrt{x}) - \Phi(-\sqrt{x}),\ x \ge 0$$

millest

$$F_{X_i^2}(x) = 2\Phi(\sqrt{x}) - 1,$$

kus Φ on normaaljaotuse $N(0,1)$ jaotusfunktsioon. Teame, et tihedusfunktsioon on jaotusfunktsiooni tuletis. Pidades silmas liitfunktsiooni diferentseerimise reegleid, saame

$$f_{X_i^2}(x) = \frac{dF_{X_i^2}(x)}{dx} = 2\phi(\sqrt{x})\frac{1}{2\sqrt{x}} = \frac{1}{\sqrt{2\pi x}}\, e^{-x/2}, \quad x > 0,$$

kus ϕ on normaaljaotuse $N(0,1)$ tihedusfunktsioon. Seose (2.3) abil veendume, et tulemus on hii-ruut jaotuse tihedusfunktsioon vabadusastmete arvuga 1,

$$X_i^2 \sim \chi^2(1).$$

Teoreemile 3 toetudes olemegi tõestanud esimese väite (2.6).

Vaatame nüüd teist väidet kujul

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2. \tag{2.8}$$

Kuna $E(\sqrt{n}\bar{X}) = 0$ ja $D(\sqrt{n}\bar{X}) = 1$, siis

$$\sqrt{n}\bar{X} \sim N(0,1).$$

Tuginedes eelnevale saame

$$n\bar{X}^2 \sim \chi(1).$$

Nüüd on avaldises (2.8) kahe hii-ruut jaotusega juhusliku suuruse vahe, mis aga ise ei pruugi olla hii-ruut jaotusega. Vaatame sõltumatute komponentidega vektorit

$$\mathbf{X} = (X_1, X_2, \ldots, X_n)^T \sim N(\mathbf{0}_n,\ \mathbf{I}_n),$$

kus $\mathbf{0}_n = E\mathbf{X}$ on vektori $\mathbf{X}$ keskväärtus ja $\mathbf{I}_n = D\mathbf{X}$ on vektori $\mathbf{X}$ kovariatsiooni maatriks. Sellest moodustame uue n-vektori $\mathbf{Y} = \mathbf{C}\mathbf{X}$, kus $\mathbf{C}$ on ortogonaalne maatriks, $\mathbf{C}\mathbf{C}^T = \mathbf{I}$. Vektori $\mathbf{Y}$ komponendid Y_i on normaaljaotusega, sest tegu on vektori $\mathbf{X}$ lineaarkombinatsiooniga. Vastavaks keskväärtusvektoriks saame:

$$E\mathbf{Y} = E(\mathbf{C}\mathbf{X}) = \mathbf{C}E\mathbf{X} = \mathbf{0}\ (vektor),$$

ja kovariatsioonimaatriksiks, mis defineeritakse seosega $D\mathbf{Y} = \mathbf{C}\ \mathbf{I}_n\ \mathbf{C}^T = \mathbf{I}$. Saime, et $DY_i = 1$. Seega $Y_i \sim N(0,1),\ \forall i$, veelgi enam, nad on sõltumatud juhuslikud suurused, sest kovariatsioonid on nullid (tuletame meelde, et mitmemõõtmelise normaaljaotuse korral tähendab nulliline kovariatsioon juhuslike suuruste sõltumatust). Valime $\mathbf{C}$ nii, et

$$\mathbf{C} = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots \frac{1}{\sqrt{n}} \\ suvalised \end{pmatrix}.$$

(Maatriksi 2×2 korral on see harjutuseks lugejale.) Nüüd

$$\mathbf{Y} = \mathbf{C}\mathbf{X} = \begin{pmatrix} \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \cdot & \cdots & \cdot \end{pmatrix} \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{n}} \sum X_i \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

Seega $Y_1 = \frac{1}{\sqrt{n}} \sum X_i = \sqrt{n}\bar{X}$. Vaatame summat:

$$\sum_{i=1}^{n} Y_i^2 = Y^T Y = X^T C^T C X = \sum_{i=1}^{n} X_i^2.$$

Kirjutame võrduse (2.8) suuruste Y_i kaudu:

$$\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 = \sum_{i=1}^{n} Y_i^2 - Y_1^2 = \sum_{i=2}^{n} Y_i^2.$$

Kuna $Y_1, Y_2, ..., Y_n$ on sõltumatud $N(0,1)$ juhuslikud suurused, siis kasutades seost (2.6) saame,

$$\sum_{i=2}^{n} Y_i^2 \sim \chi^2(n-1),$$

millega oleme ka seose (2.7) tõestanud. □

Lemma 18 *(Valimikeskmise ja dispersiooni sõltumatus normaaljaotuse korral.) Kui* X_i, $i = 1, 2, \ldots, n$ *on sõltumatud normaaljaotusega* $N(0, 1)$ *juhuslikud suurused, siis* $\sum_{i=1}^{n}(X_i - \bar{X})^2$ *ja* $\bar{X}$ *on sõltumatud juhuslikud suurused.*

Tõestus. Viimasest teoreemist järeldus, et Y_1 on sõltumatu suurusest Y_i, $i = 2, 3, \ldots, n$ ning seega $\sqrt{n}\bar{X}$ on sõltumatu suurusest

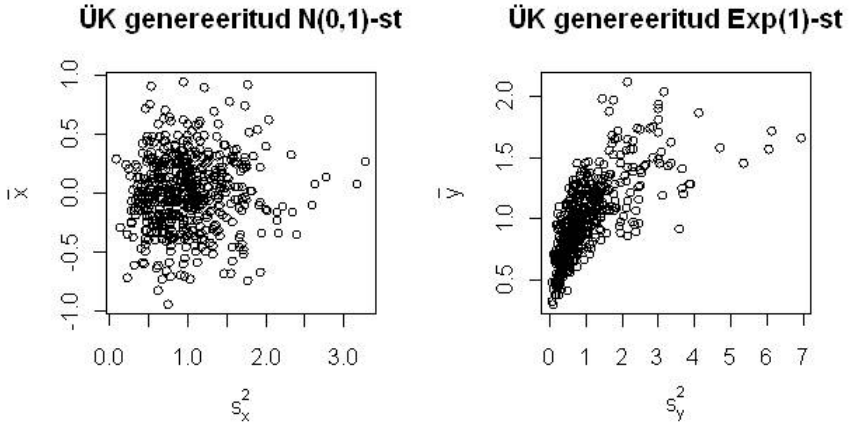
$$\sum_{i=2}^{n} Y_i^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2. \ \square$$

Järeldus ütleb ka, et valimikeskmine $\bar{X}$ ja valimidispersioon s^2 on sõltumatud normaaljaotusega üldkogumi korral. On näidatud, et teiste üldkogumijaotuste korral see omadus üldiselt ei kehti.

Näide 17 *Illustreerimi simuleerimisnäite põhjal, et* s^2 *ja* $\bar{X}$ *on tõepoolest sõltumatud jaotuse* $N(0, 1)$ *korral. Kontrnäitena kasutame jaotust* $Exp(1)$.

Algoritm:

1. *Genereerida üks valim mahuga* $n = 10$ *jaotusest* $N(0, 1)$ *ja teine valim jaotusest* $Exp(1)$.
2. *Arvutada valimite põhjal keskmised* $\bar{x}$, $\bar{y}$ *ja dispersioonid* s_x^2, s_y^2.
3. *Kanda punkt* $(\bar{x}, s_x^2)$ *ühele graafikule ja punkt* $(\bar{y}, s_y^2)$ *teisele.*
4. *Korrata sammud 1-3* $k = 500$ *korda.*



Näeme, et normaaljaotusega üldkogumi korral näitab punktipilv sõltumatuse mustrit, eksponentjaotuse korral aga mittelineaarset sõltuvust valimikeskmise ja -dispersiooni vahel.

Järeldus 3 *(Hii-ruut jaotuse seos jaotusega* $N(\mu, \sigma)$*.) Sõltumatute* $X_i \sim N(\mu, \sigma)$ *korral kehtib:*

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

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2(n-1).$$

Tõestus. Kuna $\frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\mu)^2 = \sum_{i=1}^n (\frac{X_i-\mu}{\sigma})^2$ ja $\frac{X_i-\mu}{\sigma} \sim N(0,1)$, siis omaduse 3 põhjal on järelduse 1. seos tõestatud.
Edasi

$$\frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\bar{X})^2 = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\mu-(\bar{X}-\mu))^2 = \sum_{i=1}^n (Z_i-\bar{Z})^2,$$

kus $Z_i = (X_i-\mu)/\sigma \sim N(0,1)$ ja $\bar{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$. Rakendades om. 3 saamegi järelduse 2. väidet samutigi tõestatud. □

2.5.2 Studenti t-jaotus

Aines "Tõenäosusteooria ja statistika I"tutvusime juba selle jaotusega ning kasutasime t-jaotuse täiendkvantiile usaldusintervallide leidmisel ning hüpoteeside kontrollimisel. Siin anname jaotuse täpse definitsiooni ning uurime t-jaotuse seoseid teiste tuntud jaotustega.

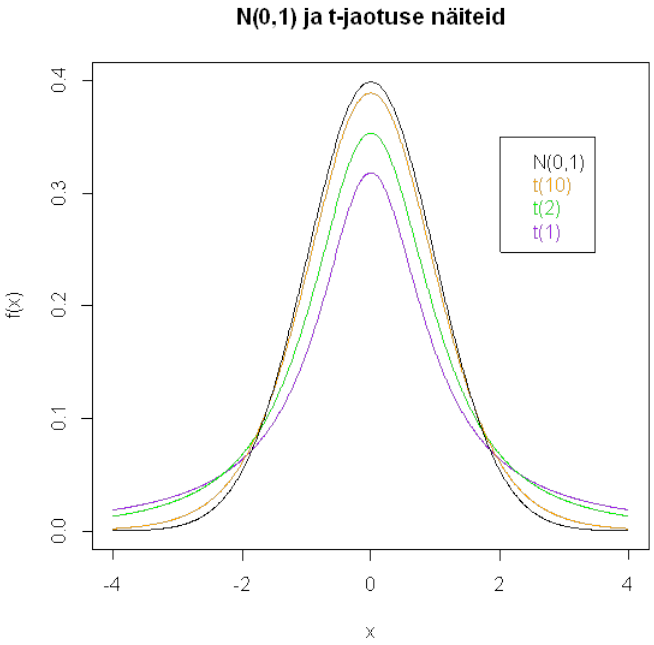
Definitsioon 18 *Juhuslik suurus X on t-jaotusega vabadusastmete arvuga f, $X \sim t(f)$, kui tema tihedusfunktsioon avaldub kujul*

$$f(x) = k(1+\frac{x^2}{f})^{-\frac{f+1}{2}}, \quad -\infty < x < \infty,$$

kus

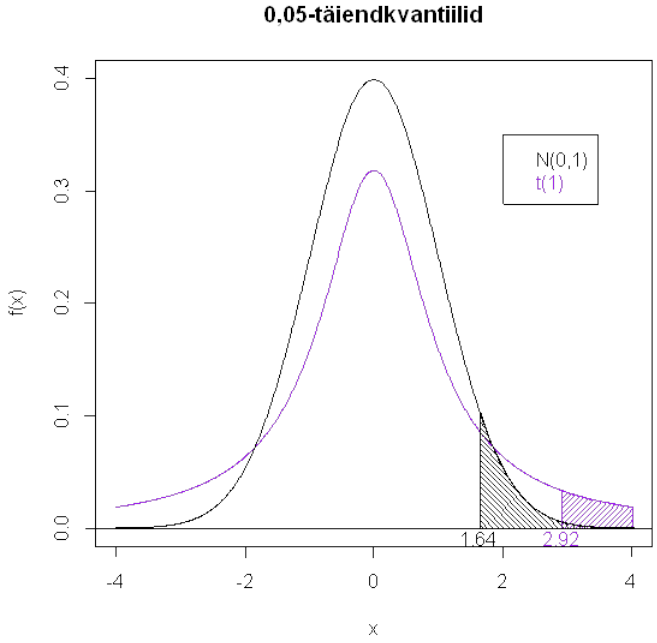
$$k = \frac{\Gamma(\frac{f+1}{2})}{\sqrt{f\pi}\Gamma(\frac{f}{2})}, \quad f \in \{1,2,\ldots\}.$$

Jaotus on sümmeetriline 0-punkti suhtes. Vabadusastmete arvu f kasvades $t(f) \to N(0,1)$. Jaotus on eriline $f = 1$ korral. Vastavat jaotust nimetatakse Cauchy jaotuseks ja sellel ei leidu momente. Muudel juhtudel on $EX = 0$. Dispersioon $DX = f/(f-2)$, ja see leidub $f > 2$ korral. Erinevate vabadusastmetega t-jaotusi näeb järgmiselt jooniselt. Märkame ka lähenemist jaotusele $N(0,1)$, kui f kasvab.



Jaotuse α-täiendkvantiili tähistame $t_\alpha(f)$ (vt. järgmist joonist), mis märgib väärtust, mille korral

$$P(X > t_\alpha(f)) = \alpha.$$



Paneme tähele, et $\lambda_\alpha \leq t_\alpha(f))$, kus λ_α on jaotuse $N(0,1)$ α-täiendkvantiil. Suure f korral $\lambda_\alpha \approx t_\alpha(f))$. t-jaotuse täiendkvantiilid on tabuleeritud (vt. Lisa B).

Järgmises teoreemis tõestame, et et t-jaotus tekib standardse normaaljaotuse ning hii-ruut jagatise käigus.

Teoreem 5 *(Seos normaal- ja χ^2- jaotusega.) Kui juhuslik suurus X on normaaljaotusega $N(0,1)$ ning juhuslik suurus Y on hii-ruut jaotusega vabadusastmete arvuga f, kusjuures X ja Y on sõltumatud, siis*

$$Z = \frac{X}{\sqrt{\frac{Y}{f}}} \sim t(f).$$

Tõestus. Tõestuse idee põhineb juba tuttaval avaldisel kahe pideva juhusliku suuruse jagatise tihedusfunktsiooni jaoks. Kui $Z = \frac{X}{V}$, kusjuures X ja V on sõltumatud ning V on mittenegatiivne juhuslik suurus, siis

$$f_Z(x) = \int_0^\infty v f_X(vx) f_V(v) dv. \tag{2.9}$$

Teame juhusliku suuruse X tihedusfunktsiooni. Siin leiame $V = \sqrt{\frac{Y}{f}}$ tihedusfunktsiooni. Esiteks V^2 jaoks saame:

$$F_{V^2}(x) = P(\frac{Y}{f} \leq x) = P(Y \leq fx) = F_Y(fx),$$

millest diferentseerimisel argumendi järgi saame tihedusfunktsiooni:

$$f_{V^2}(x) = \frac{dF_{V^2}(x)}{dx} = \frac{dF_Y(fx)}{dx} = f_Y(fx)f.$$

Kuna $Y \sim \chi^2(f)$, siis asendades hii-ruudu tiheduse avaldises (2.2) argumendi x korrutisega fx, saame

$$f_{V^2}(x) = k_1 x^{f/2-1} e^{-fx/2},$$

kus k_1 on konstant. Nüüd V jaoks:

$$F_V(x) = P(V \le x) = P(V^2 \le x^2) = F_{V^2}(x^2),\ V > 0,$$
$$f_V(x) = \frac{d}{dx} F_{V^2}(x^2) = f_{V^2}(x^2)2x = k_1 x^{2(f/2-1)} e^{-fx^2/2} 2x.$$

Asendades teadaolevad tihedused avaldisse (2.9) saamegi integreerimise abil teoreemi tõestatud. □

Statistikas omab tähtsust antud teoreemi järgmine rakendus.

Teoreem 6 *(Keskväärtuse ja standardhälbe jagatisest.) Olgu $X_i \sim N(\mu, \sigma)$ sõltumatud juhuslikud suurused, $i = 1, 2, \ldots, n$, siis*

$$\frac{\bar{X} - \mu}{\boldsymbol{s}/\sqrt{n}} \sim t(n-1),$$

kus $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ ja $\boldsymbol{s} = (\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2)^{1/2}$.

Tõestus. Anname avaldisele teise kuju:

$$\frac{\bar{X} - \mu}{\boldsymbol{s}/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{\frac{1}{n-1}\sum (X_i - \bar{X})^2}} = \frac{\sqrt{n}(\frac{\bar{X}-\mu}{\sigma})}{\sqrt{\frac{\frac{1}{\sigma^2}\sum (X_i - \bar{X})^2}{n-1}}}.$$

Kuna

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

ja teoreemi 4 põhjal

$$\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi^2(n-1),$$

siis teoreemi 5 põhjal saame

$$\frac{\bar{X} - \mu}{\boldsymbol{s}/\sqrt{n}} \sim t(n-1).$$

□

2.5.3 F-jaotus

Veel üks statistikas tähtis jaotus on F-jaotus.

Definitsioon 19 *Juhuslik suurus X on $F-$jaotusega vabadusastmete arvudega f_1 ja f_2, $X \sim F(f_1, f_2)$, kui tema tihedusfunktsioon avaldub seosega*

$$f(x) = k\ x^{1/2(f_1-2)}(f_2 + f_1 x)^{-1/2(f_1+f_2)},\quad x > 0,$$

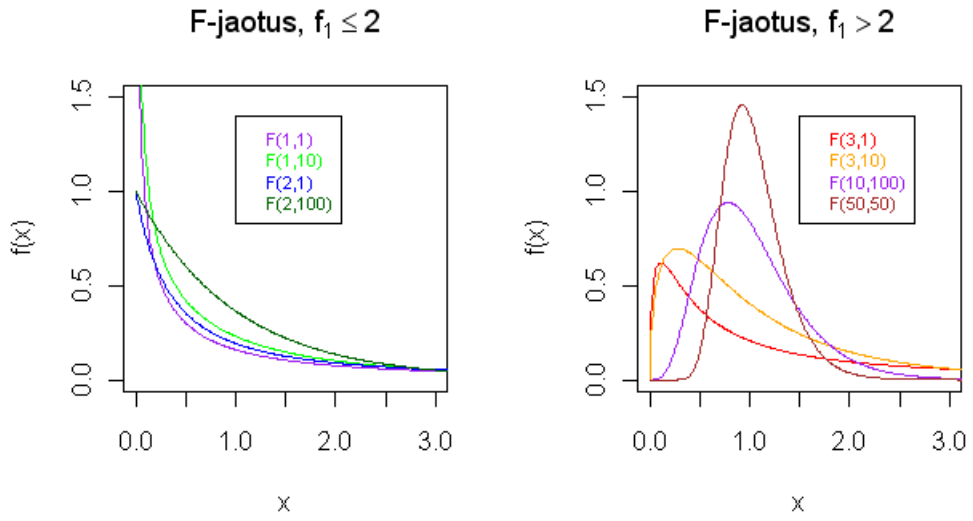
kus normeeriv konstant omab kuju

$$k = \frac{f_1^{f_1/2} f_2^{f_2/2} \Gamma(\frac{f_1+f_2}{2})}{\Gamma(\frac{f_1}{2})\Gamma(\frac{f_2}{2})}.$$

Jaotust kasutame statistiliste otsustuste tegemisel ÜK dispersiooni/-de kohta:

- vahemikhinnang ÜK dispersioonile ja standardhälbele
- hüpoteeside kontroll kahe ÜK dispersiooni võrdsuse kohta
- faktoranalüüs: hüpotees faktori mõju kohta

Järgmisel joonisel on toodud F-jaotus erinevate parameetrite korral.



Graafikult näeme, et

- tihedusfunktsiooni kuju on langev, kui $f_1 \leq 2$;
- kui $f_1 > 2$, siis on tegemist ühemodaalse ebasümmeetrilise jaotusega;
- keskväärtus eksisteerib, kui $f_2 > 2$: $EX = \frac{f_2}{f_2-2}$;
- dispersioon eksisteerib, kui $f_2 > 4$: $DX = \frac{2f_2^2(f_1+f_2-2)}{f_1(f_2-2)^2(f_2-4)}$.

Jaotuse täiendkvantiilid on tabuleeritud (vt. Lisa C). Osutub, et F-jaotus on vahetult seotud hii-ruut jaotusega.

Teoreem 7 *(Seos hii-ruut jaotusega.) Kui juhuslik suurus* $U \sim \chi^2(f_1)$ *ja* $V \sim \chi^2(f_2)$ *ning* U *ja* V *on sõltumatud, siis juhuslik suurus* X *on* F*- jaotusega:*

$$X = \frac{U/f_1}{V/f_2} \sim F(f_1, f_2).$$

Tõestuse idee. Siin anname ainult tõestuse idee.

- Leiame $\frac{U}{f_1}$ tihedusfunktsiooni:

$$P(\frac{U}{f_1} \leq x) = P(U \leq f_1 x) = F_U(f_1 x), \quad f_{U/f_1}(x) = f_U(f_1 x) f_1.$$

- Analoogiliselt leiame $\frac{V}{f_2}$ tihedusfunktsiooni.

- Nüüd kasutame teadmist, et kui $Z = \frac{X}{Y},\ X \perp Y$, siis

$$f_Z(z) = \int_0^\infty y f_X(yz) f_Y(y)\, dy.$$

  Integreerimise tulemuseks on F-jaotuse tihedusfunktsioon.

□

Järeldus 4 *Teoreemist järeldub, et kui $X \sim F(f_1, f_2)$, siis $\frac{1}{X} \sim F(f_2, f_1)$.*

Statistikas on tähtsal kohal teoreemi rakendus valimidispersioonidele.

Teoreem 8 *(Kahe valimidispersiooni jagatisest.) Olgu antud juhuslik valim $x_1, x_2, ..., x_{n_1}$ jaotusest $N(\mu_1, \sigma_1)$ ja sellest sõltumatu valim $y_1, y_2, ..., y_{n_2}$ jaotusest $N(\mu_2, \sigma_2)$. Vastavad valimite dispersioonid olgu s_1 ja s_2. Siis*

$$\frac{\boldsymbol{s}_1^2/\sigma_1^2}{\boldsymbol{s}_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1),$$

kus $\boldsymbol{s}_i^2$, $i = 1, 2$ on valimidispersioonile $s_i^2, i = 1, 2$ vastav hinnangufunktsioon.

Tõestus on harjutuseks lugejale.

Peatükk 3

Punkthinnang

Järgnevas peatükis keskendume statistilisele ülesandele, milleks on üldkogumi parameetrite hindamine. Kordame hinnangu omadusi, mis võimaldavad võrrelda ühele ja samale parameetrile pakutud hinnanguid omavahel. Nii saame põhjendatult valida parima nendest. Lihtsamate parameetrite korral (nagu näiteks üldkogumi keskmine) pole keeruline välja pakkuda hinnanguid intuitiivselt. Kuid kuidas seda teha näiteks Gini kordaja korral? Siinkohal tutvume kolme meetodiga, mis võimaldavad tuletada hinnangut lähtuvalt jaotusest või jaotuse momentidest. See tähendab, et üldkogumi parameetrite hindamisel seame uuritavale tunnusele vastavusse mudelit, ehk jaotust F. Jaotuse F kuju võib olla teada või mitte, kuid jaotuse parameetrid on enamasti tundmatud ja neid soovitakse hinnata juhusliku valimi abil.

3.1 Punkthinnang ja hinnangufunktsioon

Meid huvitab tunnus jaotusega F , mis sõltub tundmatust parameetrist θ, $F = F(\theta)$. Olgu antud juhuslik valim $\boldsymbol{x}$ üldkogumijaotusest F:

$$\begin{aligned}\boldsymbol{x} &= (\ x_1, x_2, ..., x_n),\\ &\quad\ \uparrow\ \uparrow \qquad \uparrow\\ \boldsymbol{X} &= (X_1, X_2, ..., X_n)\ ,\ \ X_i \sim F, \text{ sõltumatud.}\end{aligned}$$

Definitsioon 20 *Punkthinnanguks parameetrile θ nimetatakse väärtust, mis arvutatakse juhusliku valimi funktsioonina $\hat{\theta} = \hat{\theta}(\mathbf{x})$. Sama funktsiooni teoreetilisest valimist, $\hat{\theta}(\mathbf{X})$, nimetatakse hinnangufunktsiooniks.*

Inglise keeles on head iseloomulikud sõnad nende kahe mõiste eristamiseks: *estimate* – punkthinnang, *estimator* – hinnangufunktsioon. Üldjuhul nimetame teoreetilise valimi funktsiooni statistikuks. Hinnangufunktsioon on selline statistik, mida kasutatakse parameetri hindamise eesmärgil.

Punkthinnang on arv. Hinnangufunktsioon (statistik) on juhuslik suurus. Punkthinnang on hinnangufunktsiooni realisatsioon antud valimil:
$\mathbf{X} \to \mathbf{x}$,
$\hat{\theta}(\boldsymbol{X}) \to \hat{\theta}(\boldsymbol{x})$.
Lühiduse mõttes jätame hinnangufunktsiooni argumendi sageli kirjutamata, tema juhuslikule olemusele viitame rasvase kirjapildiga: $\hat{\theta}(\boldsymbol{X}) = \hat{\boldsymbol{\theta}}$.

Näide 18 *Tootja soovib hinnata üht tüüpi kohukeste keskmist kaalu, hindamaks kas tööliin on õigesti kalibreeritud.*
Selleks kaalutakse 20 kohukest ja saadakse andmed:

28.87	*25.61*	*30.88*	*27.98*	*26.66*	*27.15*	*29.50*	*27.54*	*27.74*	*27.94*
26.42	*28.04*	*28.28*	*28.49*	*28.50*	*24.46*	*29.11*	*29.13*	*27.31*	*26.25*

Kui eeldada andmetele normaaljaotust $N(\mu, \sigma)$, siis μ on huvipakkuv keskmine kaal, mis pole teada ja mida üritame saadud valimi põhjal hinnata.

Normaaljaotuse keskväärtuse hindamiseks võib intuitiivselt välja pakkuda järgmisi hinnanguid:

1. *valimikeskmine* $\hat{\mu}_1 = \bar{x} = 555.86/20 = 27.793$*;*
2. *mediaan* $\hat{\mu}_2 = (x_{(10)} + x_{(11)})/2 = (27.94 + 27.98)/2 = 27.960$*;*
3. *ekstreemsete väärtuste poolsumma* $\hat{\mu}_3 = (x_{(1)}+x_{(20)})/2 = (24.46+30.88)/2 = 27.670$*;*
4. *kärbitud keskmine (jaotuse sabad (10%) on välja jäetud)* $\hat{\mu}_4 = \bar{x}_{karb} = (555.86 - 24.46 - 25.61 - 29.50 - 30.88)/16 = 27.838$.

Vastavad hinnangufunktsioonid on $\hat{\boldsymbol{\mu}}_1 = \bar{X}$, $\hat{\boldsymbol{\mu}}_2 = (X_{(10)}+X_{(11)})/2$, $\hat{\boldsymbol{\mu}}_3 = (X_{(1)}+X_{(20)})/2$ *ja* $\hat{\boldsymbol{\mu}}_4 = \bar{x}_{karb}$*. Need on teoreetilised suurused, juhusliku loomuga. Nende abil saame uurida hinnangute omadusi.*

3.2 Hinnangu omadused

Hinnangu omadused on kirjeldatud vastava hinnangufunktsiooni jaotusega. Jaotuse leidmiseks on mitu võimalust.

1. Analüütiliselt ja täpselt (täpne tihedus- ja jaotusfunktsiooni avaldise kuju). Puudus: saab rakendada üksnes lihtsatel juhtudel.
2. Ligikaudselt, kasutades asümptootilisi tulemusi (paljude hinnangute korral on tõestatud, et valimimahu kasvades nende jaotus läheneb teatud piirjaotusele, nt. normaaljaotusele). Tuletame siinkohal tsentraalse piirteoreemi:

   Olgu $X_1, X_2, ...$ *sõltumatud sama jaotusega juhuslikud suurused, kus* $EX_i = \mu$, $DX_i = \sigma$ *(lõplik). Olgu* $Y_n = X_1 + X_2 + ... + X_n$*, siis* $n \to \infty$ *kehtib, et*

$$\frac{Y_n - EY_n}{\sqrt{DY_n}} = \frac{Y_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{D} N(0, 1),$$

mis ütleb, et $Y_n \sim AsN(n\mu, \sqrt{n}\sigma) \Rightarrow \frac{Y_n}{n} \sim AsN\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

3. Statistilise simulatsiooni teel. Oletame, et tahetakse teada $\hat{\boldsymbol{\theta}}$ jaotust. Fikseeritakse jaotus $F(\theta)$ ja sellest genereeritakse juhuslik valim. Valimi põhjal leitakse punkthinnang $\hat{\theta}^{(1)}$. Protseduuri korratakse R (suur) korda. Saadakse punkthinnangute valim $\hat{\theta}^{(1)}, ..., \hat{\theta}^{(R)}$, mille histogramm hindabki hinnangufunktsiooni $\hat{\theta}$ jaotust, valimikarakteristikud aga jaotuse parameetreid. Meetodi puuduseks on see, et tulemused kehtivad ainult konkreetse ülesandepüstituse korral.

Näide 19 *Olgu* $x_1, x_2, ..., x_n$ *juhuslik valim jaotusest* $U(0, \theta)$*, kus* θ *on tundmatu. Selle parameetri hindamiseks pakutakse järgmine hinnang:*

$$\hat{\theta} = max(x_1, x_2, ..., x_n).$$

Leiame $\hat{\boldsymbol{\theta}}$ *jaotuse (täpsemini öeldes tihedusfunktsiooni) nii analüütiliselt kui ka simulatsiooni teel.*

Kõigepealt hinnangufunktsioon: $\hat{\boldsymbol{\theta}} = max(X_1, X_2, ..., X_n)$*. Antud statistiku jaotust on lihtne analüütiliselt tuletada:*

$$\begin{aligned} F_{X_{max}}(x) &= P(X_{max} \leq x) = P(X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x) \\ &= \prod_{i=1}^{n} P(X_i \leq x) = \{F(x)\}^n . \end{aligned}$$

Diferentseerides saame tihedusfunktsiooni,

$$f_{X_{max}}(x) = \{F_{X_{max}}(x)\}' = n\{F(x)\}^{n-1} f(x), \tag{3.1}$$

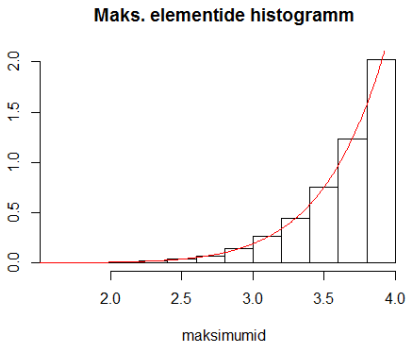
kus $f(x) = \frac{1}{\theta}$ *on jaotuse* $U(0, \theta)$ *tihedusfunktsioon,* $x \in [0, \theta]$*. Jaotusfunktsioon* $F(x)$ *on seega*

$$F_{X_{max}}(x) = \int_0^x \frac{1}{\theta} dt = \frac{t}{\theta}\Big|_0^x = \frac{x}{\theta}.$$

Kokkuvõttes saame täpse valemi hinnangufunktsiooni $\hat{\boldsymbol{\theta}}$ *tihedusfunktsiooni jaoks:*

$$f_{X_{max}}(x) = n\left(\frac{x^{n-1}}{\theta^n}\right). \tag{3.2}$$

Järgmisel joonisel on punase joonega kantud funktsioon $f_{X_{max}}(x)$ *juhul, kui valimimaht* $n = 10$ *ja* $\theta = 4$*. Histogramm vastab simuleerimistulemustele, kus* $r = 10000$ *korda genereeritakse 10-elemendiline valim jaotusest* $U(0,\ 4)$ *ja arvutatakse saadud valimi maksimum.*



Näeme, et simuleeritud jaotus on sama kujuga, mis sai leitud analüütiliselt.

Kui hinnangufunktsiooni jaotus on teada, siis saab hakata uurima selle hinnangu omadusi.

Definitsioon 21 *Hinnang on nihketa kui kehtib:* $E\hat{\boldsymbol{\theta}} = \theta$*, vastasel juhul nihkega, kus nihe on defineeritud kui* $B = E\hat{\boldsymbol{\theta}} - \theta$.

Inglise keeles: nihketa = *unbiased.*

Näide 20 *Tõestame, et kui* $x_1, ..., x_n$ *on juhuslik valim jaotusest* $U(0, \theta)$*, siis* $\hat{\theta}_1 = max(x_1, ..., x_n)$ *on nihkega ja* $\hat{\theta}_2 = 2 \cdot \bar{x}$ *on nihketa hinnang parameetrile* θ.

Alustame hinnangust $\hat{\theta}_2$*. Sellele vastav hinnangufunktsioon on* $\hat{\boldsymbol{\theta}}_2 = 2\bar{X} = \frac{2}{n}\sum_{i=1}^{n} X_i$*. Arvestades, et* $X_i \sim U(0,\ \theta)$ *ja järelikult* $EX_i = \theta/2,\ \forall i = 1, 2, ..., n$ *leiame hinnangu* $\hat{\boldsymbol{\theta}}_2$ *keskväärtuse:*

$$E\hat{\boldsymbol{\theta}}_2 = E\left(\frac{2}{n}\sum_{i=1}^{n} X_i\right) = \frac{2}{n}\sum_{i=1}^{n} EX_i = \frac{2}{n}\sum_{i=1}^{n}\frac{\theta}{2} = \theta,$$

ehk hinnang $\hat{\theta}_2$ *on nihketa parameetri* θ *jaoks.*

Nüüd vaatleme hinnangut $\hat{\theta}_1$*. Vastav hinnangufunktsioon on* $\hat{\boldsymbol{\theta}}_1 = max(X_1, ..., X_n)$*. Kuna* X_i *on pidevad juhuslikud suurused, on ka maksmiaalne element pidev. Keskväärtuse saame pideva juhusliku suuruse keskväärtuse abil:*

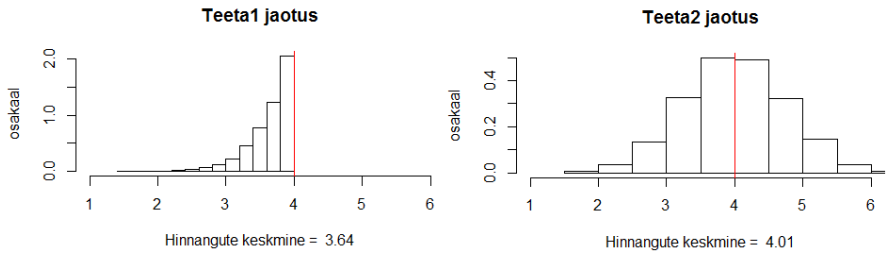
$$E\hat{\boldsymbol{\theta}}_1 = \int_{-\infty}^{\infty} x \cdot f_{\hat{\boldsymbol{\theta}}_1}(x)dx,$$

kus avaldis $f_{\hat{\boldsymbol{\theta}}_1}(x)$ *jaoks on toodud valemi (3.2) abil. Lihtsa integreerimise teel saame, et*

$$E\hat{\boldsymbol{\theta}}_1 = \frac{n\theta}{\theta + 1} \neq \theta.$$

Seega, on hinnang $\hat{\theta}_1$ *nihkega hinnang parameetrile* θ*. Ta hindab tegeliku parameetrit alla.*

Järgmisel joonisel on toodud simuleerimistulemused kahe hinnangu jaoks: kõigepealt fikseeriti parameetri θ *väärtus,* $\theta = 4$ *(simuleerimisülesandes on enamasti tegelik parameeter teada) ja seejärel võeti 10 000 valimit mahuga 10 jaotusest* $U(0,\ 4)$*. Iga valimi korral leiti* $\hat{\theta}_1$ *ja* $\hat{\theta}_2$ *väärtused (kokku kaks komplekti pikkusega 10 000 väärtust). Nende histogrammid peegeldavad* $\hat{\boldsymbol{\theta}}_1$ *ja* $\hat{\boldsymbol{\theta}}_2$ *jaotust. Punane vertikaaljoon vastab tegelikule* $\theta = 4$ *väärtusele ning seda parameetrit hindavad leitud* $\hat{\theta}_1$ *ja* $\hat{\theta}_2$ *väärtused igal simulatsiooni sammul.*



Vasakul joonisel näeme, et ükski väärtus ei ületa parameetrit $\theta = 4$, enamus on sellest väiksemad. Teoreetiliselt saime samuti, et tegemist on alahinnanguga. Paremal joonisel aga on $\hat{\theta}_2$ väärtused ja need on hajutatud tegeliku väärtuse ümbruses ligikaudu võrdse osakaaluga, mis kinnitab teoreetilist tulemust, et hinnang $\hat{\theta}_2$ on nihketa.

Hinnangu iseloomustab ka selle varieeruvus ehk dispersioon. Järgmine mõiste võtab kokku nihke ja dispersiooni.

Definitsioon 22 *Hinnangu $\hat{\theta}$ ruutkeskmiseks veaks (MSE=Mean Square Error) nimetatakse suurust*

$$MSE(\hat{\boldsymbol{\theta}}) = E(\hat{\boldsymbol{\theta}} - \theta)^2.$$

Lemma 19 *Ruutkeskmist viga on võimalik esitada järgmisel alternatiivsel kujul:*

$$MSE(\hat{\boldsymbol{\theta}}) = D\hat{\boldsymbol{\theta}} + \Big[E(\hat{\boldsymbol{\theta}}) - \theta\Big]^2 = \textit{Hinnangu varieeruvus} + (\textit{nihe})^2.$$

Tõestuses tuleb lähtuda MSE definitsioonist ja on harjutuseks lugejale.

Näide 21 *Leiame eelmises näites mõlema hinnangu MSE. Kuna hinnanng $\hat{\theta}_2$ on nihketa, siis*

$$MSE(\hat{\boldsymbol{\theta}}_2) = D(\hat{\boldsymbol{\theta}}_2).$$

Arvestades, et $X_i \sim U(0,\ \theta)$ korral $DX_i = \frac{\theta^2}{12}\ \forall i = 1, ..., n$ ja et X_i on sõltumatud juhuslikud suurused, saame

$$D(\hat{\boldsymbol{\theta}}_2) = D\left(\frac{2}{n}\sum_{i=1}^{n} X_i\right) = \frac{4}{n^2}\sum_{i=1}^{n}\frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$

Hinnangu $\hat{\theta}_2$ MSE leiame pideva juhusliku suuruse funktsiooni $g(x)$ keskväärtusena $E(g(x))$, kus $g(x) = (x-\theta)^2$:

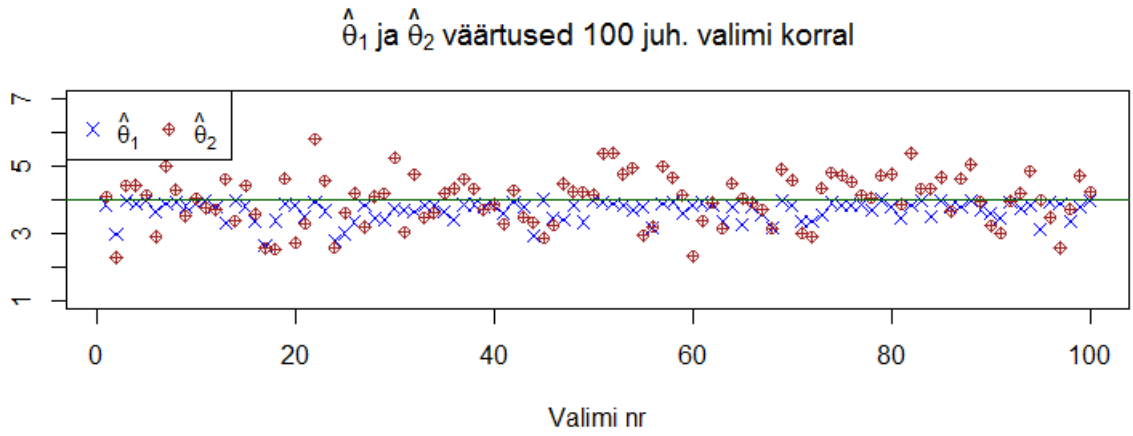
$$MSE(\hat{\boldsymbol{\theta}}_1) = E[g(x)] = \int_0^{\theta}(x-\theta)^2 f_{\hat{\boldsymbol{\theta}}_2}(x)dx = \int_0^{\theta}(x-\theta)^2 \cdot \frac{nx^{n-1}}{\theta^n}dx = \frac{2\theta^2}{(n+2)(n+1)}.$$

Näeme, et valimimahu kasvades koondub $MSE(\hat{\boldsymbol{\theta}}_2)$ nulliks kiirusega n ja $MSE(\hat{\boldsymbol{\theta}}_1)$ kiirusega n^2, ehk kiiremini. Kõikide valimite korral, kus $n > 1$ kehtib

$$MSE(\hat{\boldsymbol{\theta}}_1) < MSE(\hat{\boldsymbol{\theta}}_2).$$

See tähendab, et kuigi nihkega on hinnang $\hat{\theta}_1$ väiksema ruutkeskmise veaga.

Järgnev joonis iseloomustab MSE mõlema hinnangu korral. Simuleerimise algoritm on sama, mis eelmises näites, kuid nüüd on esimest 100 hinnangut mõlema komplekti korral kantud ühele joonisele. Tegelikule parameetrile $\theta = 4$ vastab horisontaalne joon. Näeme, et sinised ristid ei ületa horisontaalset joont ühelegi valimi korral (sest tegemist on alahinnanguga), kuid punased ringid on hajutatud horisontaalse joone ümber (keskmiselt annavad tegeliku parameetrit $\theta = 4$). Siniste ristide varieeruvus on väiksem kui punaste ringide oma. Seda tulemust saime ka analüütiliselt.



Kui ühele ja samale parameetrile on pakutud kaks (või rohkem) hinnangufunktsiooni, siis võrreldakse neid sageli järgmise omaduse abil.

Definitsioon 23 *Öeldakse, et nihketa hinnang $\hat{\theta}_1$ on **efektiivsem** kui nihketa hinnang $\hat{\theta}_2$ kui kehtib: $D\hat{\boldsymbol{\theta}}_1 \leq D\hat{\boldsymbol{\theta}}_2$ range võrratusega vähemalt ühe hinnangufunktsiooni väärtuse korral vastavast parameeterruumist.*

NB! Nihkega hinnangute korral kasutatakse võrdluseks ruutkeskmist viga.

Eelmises näites on $\hat{\theta}_1 = max(x_1, ..., x_n)$ efektiivsem kui $\hat{\theta}_2 = 2\bar{x}$ vaatamata nihkele.
Tutvume siinkohal veel ühe tähtsa hinnangu omadusega.

Definitsioon 24 *Öeldakse, et hinnang $\hat{\theta}$ on **mõjus**, kui $\forall \theta \in A$ ja $\forall \varepsilon > 0$ korral*

$$P\left(|\hat{\boldsymbol{\theta}} - \theta| > \varepsilon\right) \xrightarrow{n\to\infty} 0.$$

Mõjusa hinnangu korral kontsentreerub vastava hinnangufunktsiooni jaotus valimimahu kasvades üha lähemale õigele väärtusele θ, mis kindlustab, et konkreetse valimi põhjal arvutatud punkthinnang on õige väärtuse lähedal. Siiski jääb eelnev definitsioon üsna teoreetiliseks ning praktikas kasutatakse efektiivsuse kontrollimiseks alternatiivset esitust.

Lemma 20 *Kui kehtivad järgmised tingimused:*

1. $\lim_{n\to\infty} E\hat{\boldsymbol{\theta}} = \theta$,
2. $\lim_{n\to\infty} D\hat{\boldsymbol{\theta}} = 0$,

siis hinnang $\hat{\theta}$ on mõjus hinnang.

Tõestus. Kehtigu tingimused (1) ja (2). Suvalise juhusliku suuruse Z korral (lõpliku dispersiooniga) kehtib Tšebõševi võrratus:

$$P(|\ Z - EZ\ | > \varepsilon) \leq \frac{DZ}{\varepsilon^2}, \quad \forall \varepsilon > 0.$$

Kui nüüd võtta $Z = \hat{\boldsymbol{\theta}}$, saame

$$P(|\ \hat{\boldsymbol{\theta}} - E\hat{\boldsymbol{\theta}}\ | > \varepsilon) \leq \frac{D\hat{\boldsymbol{\theta}}}{\varepsilon^2}.$$

Arvestades tingimusi (1) ja (2), saame

$$P(|\ \hat{\boldsymbol{\theta}} - \theta\ | > \varepsilon) \to \frac{0}{\varepsilon^2} = 0, n \to \infty,$$

mis ongi mõjususe definitsioon. Seega $\hat{\theta}$ on mõjus hinnang. □

Märkus. Lemma 20 tingimused 1-2 saab kokku võtta ka järgmise tingimuse abil:

$$\lim_{n\to\infty} MSE(\hat{\boldsymbol{\theta}}) = 0.$$

Näide 22 *Tõestame, et eelmise kolme näite hinnangud $\hat{\theta}_1$ kui ka $\hat{\theta}_2$ on mõjusad hinnangud. Selleks kasutame eelmises näites leitud MSE:*

$$\begin{aligned} MSE(\hat{\boldsymbol{\theta}}_2) &= \frac{\theta^2}{3n} \to 0,\ n \to \infty; \\ MSE(\hat{\boldsymbol{\theta}}_1) &= \frac{2\theta^2}{(n+2)(n+1)} \to 0,\ n \to \infty. \end{aligned}$$

Sellest järeldub, et mõlemad hinnangud on efektiivsed hinnangud.

Kui mõni hinnang on leitud, st arvutatud valimi väärtuste põhjal, siis on heaks tavaks leida sellele ka mõni täpsuse näitaja, mis iseloomustaks leitud hinnangu varieeruvust. Väga levinud on järgmised täpsuse näitajad:

- hinnangu standardviga: $\sqrt{\hat{D}\hat{\boldsymbol{\theta}}}$
- hinnangu suhteline viga: $\sqrt{\hat{D}\hat{\boldsymbol{\theta}}}\Big/\hat{\theta}$. Hea hinnangu korral jääb tavaliselt suhteline viga alla 0,2.

Näide 23 *Näites 18 on leitud, et $\hat{\mu}_1 = \bar{x} = 27,793$. Leiame selle hinnangu suhtelise vea. Selleks peame teadma hinnangu dispersiooni,*

$$D\hat{\boldsymbol{\mu}}_1 = D\bar{X} = \frac{\sigma^2}{n}.$$

Antud juhul andmejaotuse dispersioon $DX_i = \sigma^2$ pole teada, seetõttu hindame σ^2 valimi põhjal kasutades vastavat nihketa hinnangut,

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{20}(x_i - \bar{x})^2 \approx 2,137.$$

See, et hinnang s^2 on nihketa parameetrile σ^2, on näidatud aines 'Tõenäosusteooria ja statistika I'.
Kokkuvõttes, $\hat{D}\hat{\boldsymbol{\mu}}_1 = \frac{2,137}{20} \approx 0,1069$, millest saame hinnangu suhtelise vea:

$$\frac{\sqrt{\hat{D}\hat{\boldsymbol{\mu}}_1}}{\hat{\boldsymbol{\mu}}_1} = \frac{\sqrt{0,1069}}{27,793} \approx 0,0117,$$

ehk tegemist on üsnagi täpse hinnanguga.

3.3 Taasvaliku meetodid hinnangu standardvea leidmiseks

Alati ei õnnestu leida hinnangu $\hat{\theta}$ standardvea $\sqrt{\hat{D}(\hat{\boldsymbol{\theta}})}$ analüütilist kuju. Praktikas on levinud taasvaliku meetodid standardvea hindamiseks. Meetodid on ligikaudsed, kuid siiski sageli annavad küllaltki täpset standardvea väärtust. Siin vaatleme parameetrilist ja mitteparameetrilist taasvaliku (= bootstrap) meetodeid.

3.3.1 Parameetriline bootstrap

Olgu valimi $x_1, x_2, ..., x_n$ kohta teada, et see on pärit jaotusest $F(\theta)$, kus θ on tundmatu. Valimi põhjal leitakse sellele parameetrile θ mõni hinnang $\hat{\theta}$, näiteks $\hat{\theta} = 27,96$ (näites 18 mediaanil põhinev hinnang).

Edasi kasutades arvutit, genereeritakse nn bootstrap valimeid jaotusest $F(27,96)$ sama mahuga n ning iga bootstrap valimi põhjal arvutatakse bootstrap hinnang $\hat{\theta}^*$:

1. bootstrap valim: $x_1^*, x_2^*, ..., x_n^*$, punkhinnang $\hat{\theta}_1^*$
2. bootstrap valim: $x_1^*, x_2^*, ..., x_n^*$, punkhinnang $\hat{\theta}_2^*$
...
B. bootstrap valim: $x_1^*, x_2^*, ..., x_n^*$, punkhinnang $\hat{\theta}_B^*$.

B on tavaliselt suur (näiteks 1000 ja rohkem).
Tähistame $\bar{\theta}^* = \sum_{i=1}^{B} \hat{\theta}_i^*/B$ - bootstrap hinnangute keskmine. Siis hinnangu $\hat{\theta}$ standardviga bootstrap meetodil on

$$\sqrt{\hat{D}\hat{\boldsymbol{\theta}}}_{BS} = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\theta}_i^* - \bar{\theta}^*\right)^2}.$$

Näide 24 *Eelmises näites leidsime, et* $\sqrt{\hat{D}(\bar{X})} = \sqrt{0,1069} \approx 0,327$. *Mida oskame öelda teiste hinnangute standardvigade kohta? Kuna nende kuju pole enam lineaarne, siis analüütiline tuletuskäik võtab aega (pole siiski võimatu!).*

Leiame kõigi nelja hinnangu standardvead parameetrilise boostrap-meetodi abil. Allpool on toodud R-i kood valimikeskmise standardvea hindamiseks. hinnangute standardvigu saab leida analoogilisel teel.

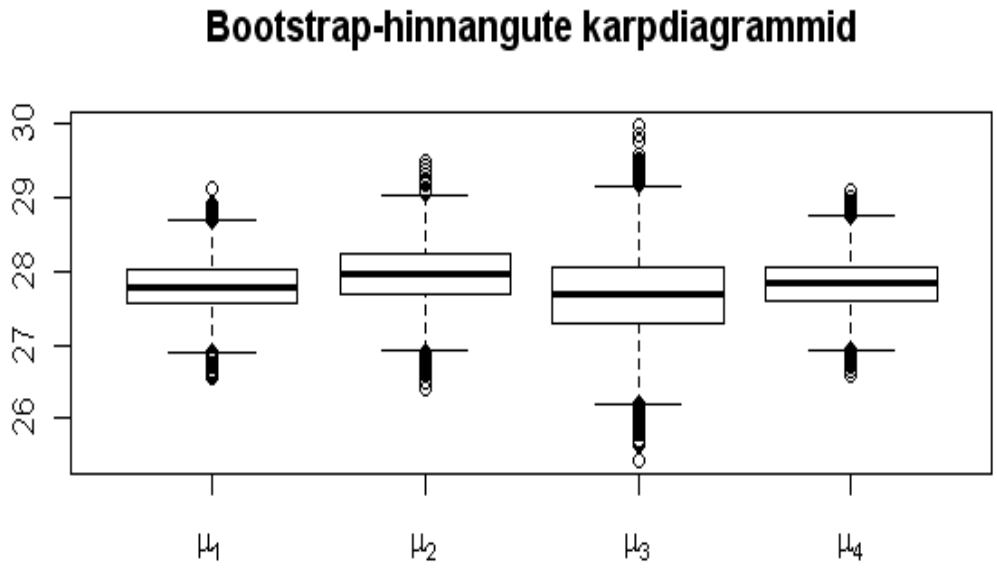
```
x=c(24.46,25.61,26.25,26.42,26.66,27.15,27.31,27.54,27.74,27.94,27.98,
28.04,28.28,28.49,28.50,28.87,29.11,29.13,29.50,30.88)
 k=10000

 # Valimikeskmine:

mu1=mean(x)
bt_mu1=rep(NA,k)
for(i in 1:k){
valim=rnorm(20,mu1,sd(x))
bt_mu1[i]=mean(valim)
}
sd(bt_mu1)
```

Programmitöö tulemuseks on `0,328`*, mis on teoreetilisele tulemusele üsna lähedal. Hinnangute* $\hat{\mu}_2, \hat{\mu}_3$ *ja* $\hat{\mu}_4$ *bootstrap-hinnangute leidmiseks võib kasutada vastavalt:* `median`*,* `(min(x)+max(x))/2` *ja* `mean(x,trim=0.1)`*.*

Kõigi nelja hinnangu (ligikaudsed) tulemused on:

$$\sqrt{\hat{D}\hat{\mu}_1} = 0,328; \quad \sqrt{\hat{D}\hat{\mu}_2} = 0,398; \quad \sqrt{\hat{D}\hat{\mu}_3} = 0,554; \quad \sqrt{\hat{D}\hat{\mu}_4} = 0,334.$$

Järgneval joonisel on hinnangute komplektid iseloomustatud karpdiagrammide abil, millest on näha, et tõepoolest hinnang $\hat{\mu}_4$ *on kõige suurema varieeruvusega.:*



3.3.2 Mitteparameetriline bootstrap

See meetod erineb parameetrilisest bootstrapist selle poolest, et ei kasuta X jaotust F, seda polegi vaja teada ega eeldada. On olemas valim mahuga n, mille kohta ei pea teadma, mis jaotuse klassist see pärit on. Endiselt huvipakkuvaks ÜK parameetriks on θ.

Arvuti abil võetakse olemasolevast valimist bootstrap valimeid mahuga n kasutades lihtsat juhuvalikut **tagasipanekuga**.

Iga bootstrap valimi põhjal arvutatakse bootstrap hinnang $\hat{\theta}^*$.

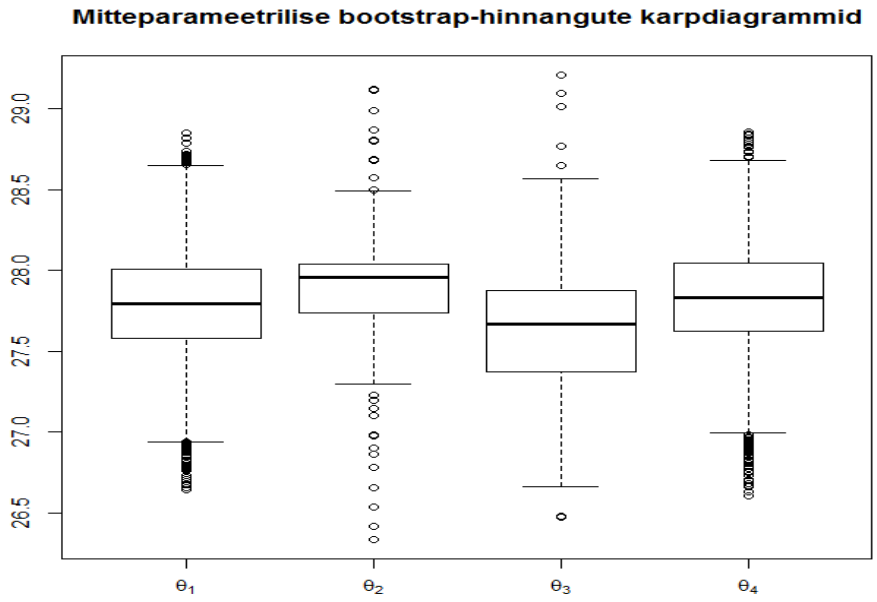
Hinnangu $\hat{\theta}$ standardviga leitakse analoogiliselt eelmise versiooniga,

$$\sqrt{\hat{D}\hat{\boldsymbol{\theta}}}_{BS} = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\theta}_i^* - \bar{\theta}^*\right)^2}.$$

Ei tohi unustada, et taasvaliku meetodi abil saab leida vaid ligikaudset standardvea väärtust.

Näide 25 *Lahendame eelmist näidet mitteparameetrilise bootsrap meetodi abil. Allpool on toodud vastav R-kood ja karpdiagrammid:*

```
x=c(24.46,25.61,26.25,26.42,26.66,27.15,27.31,27.54,27.74,27.94,27.98,
28.04,28.28,28.49,28.50,28.87,29.11,29.13,29.50,30.88)
 k=10000

 teeta1=rep(NA,k) #valimikeskmine

teeta2=rep(NA,k) #mediaan
teeta3=rep(NA,k) #(Maximum+Minimum)/2
teeta4=rep(NA,k) #Kärbitud keskmine (2 min ja 2 max elementi on ära jäetud)
 for(i in 1:k){

valim=sample(x, 20, replace=TRUE) #valik TGA mahuga 20 esialgsest valimist
teeta1[i]=mean(valim)
teeta2[i]=median(valim)
teeta3[i]=(min(valim)+max(valim))/2
teeta4[i]=mean(valim,trim=0.1)
} boxplot(teeta1, teeta2, teeta3, teeta4,
names=c(expression(theta[1]),expression(theta[2]),expression(theta[3]),
expression(theta[4])),
main="Mitteparameetrilise bootstrap-hinnangute karpdiagrammid")
```



3.3.3 Taylori ritta arendus

See meetod on veel üheks alternatiiviks hinnangu standardvea leidmiseks. Eriti on see abiks hinnangu omaduste uurimisel teoreetiliselt. Taasvaliku meetodid annavad ühte arvulist väärtust hinnangu standardvea kohta, Taylori ritta arenduse abil on võimalik saada analüütuilist kuju hinnangu keskväärtusele ja dispersioonile, kuigi ligikaudselt. Meetod on tuntud ka **Delta meetod**i nime all.

Oletame, et huvitume funktsiooni $g(\theta)$ hinnangust $g(\hat{\theta})$. Näiteks meditsiinis on laialt ka-

sutatav nn šansside suhe $g(p) = \frac{p}{1-p}$, kus p on omaduse/haiguse A osakaal üldkogumis (tundmatu). Valimi põhjal saame küll leida nihketa hinnangu parameetrile p, aga mida oskame sel juhul öelda hinnangust $g(\hat{p}) = \frac{\hat{p}}{1-\hat{p}}$? Kas see on nihketa? Mis on selle standardviga? Tegemist on mittelineaarse hinnanguga ning teadaolevaid keskväärtuse ja dispersiooni omadusi rakendada pole võimalik.

Olgu $g(\hat{\boldsymbol{\theta}})$ selline mittelineaarne hinnang, kus $\hat{\theta}$ ise on mõjus hinnang keskväärtusega $E\hat{\boldsymbol{\theta}} = \theta$ ja dispersiooniga $D\hat{\boldsymbol{\theta}}$. Idee põhineb hinnangu $g(\hat{\boldsymbol{\theta}})$ arendamisel Taylori ritta tegeliku parameetri θ ümbruses ning selle lineaarosa kasutamises.

Lemma 21 *Olgu $g(\hat{\boldsymbol{\theta}})$ differentseeruv ja $g'(\theta) \neq 0$. Lisaks, eksisteerigu mõjus hinnang parameetrile θ, olgu $\hat{\theta}$. Siis funktsiooni $g(\hat{\boldsymbol{\theta}})$ ligikaudne keskväärtus ja dispersioon avalduvad järgmiselt:*

$$E[g(\hat{\boldsymbol{\theta}})] \approx g(\theta), \quad D[g(\hat{\boldsymbol{\theta}})] \approx g'(\theta)^2 D[\hat{\boldsymbol{\theta}}].$$

***Tõestus**. Arendame funktsiooni $g(\hat{\boldsymbol{\theta}})$ Taylori ritta punkti θ ümbruses ja võtame sellest ainult lineaarse liikme:*

$$g(\hat{\boldsymbol{\theta}}) \approx g(\theta) + g'(\theta)(\hat{\boldsymbol{\theta}} - \theta).$$

Leiame lineaarliikme keskväärtuse:

$$E[g(\hat{\boldsymbol{\theta}})] \approx E[g(\theta)] + g'(\theta)(E\hat{\boldsymbol{\theta}} - \theta) = g(\theta)$$

ja dispersiooni:

$$D[g(\hat{\boldsymbol{\theta}})] \approx D[g(\theta)] + D[g'(\theta)(\hat{\boldsymbol{\theta}} - \theta)] = g'(\theta)^2 D[\hat{\boldsymbol{\theta}} - \theta] = g'(\theta)^2 D[\hat{\boldsymbol{\theta}}].$$

□

Näide 26 *Olgu antud suur valim $x_1, ..., x_n$ jaotusest $Exp(\lambda)$. Varasemast teame, et*

$$EX_i = \frac{1}{\lambda} \text{ ja } DX_i = \frac{1}{\lambda^2}, \ i = 1, ..., n.$$

Huvitume parameetri λ hindamisest.

Kuna $\frac{1}{\lambda}$ on jaotuse keskväärtus, siis selle hindamiseks sobib valimikeskmine (tegemist on mõjusa hinnanguga):

$$\frac{1}{\hat{\lambda}} = \bar{x},$$

millest

$$\hat{\lambda} = \frac{1}{\bar{x}}.$$

Hinnangu $\hat{\lambda}$ keskväärtuse ja dispersiooni leidmiseks rakendame Taylori ritta arendust. Siin on $\theta = EX$, $\hat{\theta} = \bar{X}$, $g(\hat{\theta}) = 1/\bar{X}$. Peame teadma veel $E\hat{\theta}$ ja $D\hat{\theta}$:

$$E\hat{\theta} = E\bar{X} = ... = \lambda^{-1} \text{ ja } D\hat{\theta} = D\bar{X} = ... = (n\lambda^2)^{-1}. \ (ise!)$$

Eelnevast lemmast saame keskväärtuse,

$$E[g(\hat{\theta})] \approx g(\theta) = \frac{1}{1/\lambda} = \lambda,$$

järelikult tegemist on ligikaudselt nihketa hinnanguga.

Ligikaudse dispersiooni saamiseks kirjutame esmalt välja $g'(\theta)$:

$$g'(\theta) = g'(\hat{\theta})\Big|_{\hat{\theta}=\theta} = \left(\bar{X}^{-1}\right)'\Big|_{\bar{X}=1/\lambda} = -\bar{X}^{-2}\big|_{\bar{X}=1/\lambda} = -\lambda^2.$$

Eelneva lemma järgi $D[g(\hat{\theta})] \approx g'(\theta)^2 D[\hat{\theta}]$, millest

$$D(\bar{X}^{-1}) \approx \lambda^4 \cdot \frac{1}{n\lambda^2} = \frac{\lambda^2}{n}.$$

Seega, hinnang $\hat{\lambda} = \frac{1}{\bar{X}}$ on ligikaudselt mõjus hinnang parameetrile λ.

3.4 Hinnangu leidmise meetodid

Siiani hinnangu omaduste uurimisel hinnang ise oli ette antud. Näiteks, normaaljaotuse keskväärtuse μ hinnanguks kasutamise valimikeskmist $\bar{x}$, aga ka mediaani ning muid hinnanguid. Sageli on võimalik hinnanguid intuitiivselt, kuid on olukordi, kus seda pole võimalik teha. Kuidas sel juhul toimida? Siin krsuses vaatleme kolme meetodit hinnangute saamiseks.

3.4.1 Suurima tõepära meetod

Suurima tõepära meetod, inglise keeles *maximum likelihood method*, on üks kasulikumaid ja matemaatiliselt ilusamaid meetodeid hinnangute leidmiseks. Samuti on neil hinnangutel mitmeid häid omadusi.

Definitsioon 25 *Olgu antud valim $x_1, x_2, ..., x_n$ jaotusest $F(x;\theta)$, mis võib olla kas pidev või diskreetne. Tõepärafunktsiooniks nimetame avaldist:*

$$L(\theta) = \begin{cases} f(x_1;\theta) \cdot f(x_2;\theta) \cdot \ldots \cdot f(x_n;\theta), & \textit{pideval juhul}, \\ p(x_1;\theta) \cdot p(x_2;\theta) \cdot \ldots \cdot p(x_n;\theta), & \textit{diskreetsel juhul}, \end{cases}$$

kus $f(x;\theta)$ on jaotuse F tihedusfunktsioon (pideval juhul) ja $p(x;\theta)$ on jaotuse F tõenäosusfunktsioon (diskreetsel juhul), $\theta \in A$.

Olgu meil $\mathbf{X} = (X_1, X_2, ..., X_n)$, X_i sõltumatud, $X_i \sim F(x;\theta)$, siis $L(\theta)$ on valimi $\mathbf{x}$= $(x_1, x_2, ..., x_n)$ saamise tõenäosus (diskreetsel juhul) või juhusliku vektori $\mathbf{X}$ tihedusfunktsiooni väärtus punktis $\mathbf{x}$ (pideval juhul) antud θ korral.

Realiseerunud valimi $\mathbf{x}$ korral on suurused $x_1, x_2, ..., x_n$ teadaolevad arvud ja $L(\theta)$ on üksnes parameetri θ funktsioon. Eesmärgiks on leida niisugune θ väärtus parameeterruumist A, et $L(\theta)$ oleks maksimaalne. Me ütleme, et vastav θ väärtus on tõepäraseim antud valimi jaoks (st. ka vastav üldkogumi jaotus on tõepäraseim antud valimi jaoks).

Suurima tõepära printsiip – tõepäraseima üldkogumijaotuse määramine antud valimi jaoks.

Definitsioon 26 *Väärtust $\hat{\theta}$ parameeterruumis A, mille korral $L(\theta)$ saavutab maksimaalse väärtuse, nimetatakse parameetri θ suurima tõepära hinnanguks:*

$$L(\hat{\theta}) = \max_{\theta \in A} L(\theta).$$

Suurima tõepära hinnangu praktilisel leidmisel on sageli lihtsam kasutada tõepärafunktsiooni logaritmi. Tänu logaritmi monotoonsusele saavutavad $L(\theta)$ ja $\ln L(\theta)$ maksimumi samas punktis, st määravad sama suurima tõepära hinnangu.

Definitsioon 27 *Logaritmiline tõepärafunktsioon on*

$$l(\theta) = \ln L(\theta) = \begin{cases} \sum_{i=1}^{n} \ln f(x_i;\theta), & \textit{pideval juhul,} \\ \sum_{i=1}^{n} \ln p(x_i;\theta), & \textit{diskreetsel juhul.} \end{cases}$$

Näide 27 Mündivise. Üldkogumijaotuseks on mündi visketulemuse (vapp, kiri) jaotus, kus vapi tulemise tõenäosuseks on p. Olgu eelnevalt teada, et $p \in \{\frac{1}{2}; \frac{1}{4}\}$. Olgu meil kaks vaatlust: $x_1 = vapp$ ja $x_2 = vapp$. Kumb on tõepärasem hinnang parameetrile p, kas $\frac{1}{2}$ või $\frac{1}{4}$?
Kirjutame välja tõepärafunktsiooni:

$$L(p) = P(X = x_1) \cdot P(X = x_2) = p^2,$$

millest $L(\frac{1}{2}) = \frac{1}{4}$, $L(\frac{1}{4}) = \frac{1}{16}$.
Kuna $L(\frac{1}{2}) > L(\frac{1}{4})$, siis $\hat{p} = \frac{1}{2}$ on suurima tõepära hinnang p-le. □

Näide 28 *Olgu üldkogumijaotuseks eksponentjaotus $Exp(\lambda)$ ja olgu $\lambda = \frac{1}{\theta}$. Vastav tihedusfunktsioon on*

$$f(x;\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}},\ x \geq 0.$$

Parameeter θ olgu tundmatu. Pole raske kontrollida, et θ on antud jaotuse keskväärtus. Olgu meil $n = 4$ vaatlust jaotusest:

$$0.322,\ 0.879,\ 0.222,\ 0.012.$$

Leiame valimi tõepärafunktsiooni

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \frac{1}{\theta^n} e^{-\frac{\sum_{i=1}^{n}}{\theta}} = \frac{1}{\theta^4} e^{\frac{-1.435}{\theta}}$$

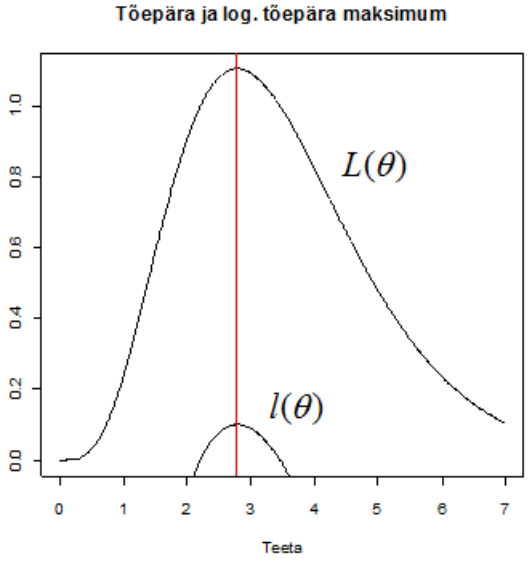
ja logaritmilise tõepärafunktsiooni

$$l(\theta) = \ln L(\theta) = -4 \ln \theta - \frac{1.435}{\theta}.$$

Näeme, et tõepärafunktsioonid on θ funktsioonid. Mõlemad funktsioonid saavutavad maksimumi samal kohal, sest logaritm on monotoonselt kasvav funktsioon (vt joonist).
Maksimumi leidmiseks leiame tuletise,

$$\frac{d}{d\theta} l(\theta) = -\frac{4}{\theta} + \frac{1.435}{\theta^2}.$$

Võrdsustades tuletise nulliga saame $l(\theta)$ maksimumpunkti, mis on ühtlasi parameetri θ suurima tõepära hinnanguks, $\hat{\theta} = 0.358$.



Et eksponentjaotus sõltub ainult ühest parameetrist, siis oleme koos θ hindamisega hinnanud ka üldkogumijaotuse

$$f(x;\hat{\theta}) = \frac{1}{0.358} e^{-\frac{x}{0.358}},\ x \geq 0.$$

Antud ülesande 4 vaatlust olid tegelikult genereeritud eksponentjaotusest parameetriga $\theta = 0.5$. Hinnang $\hat{\theta} = 0.358$ ei ole eriti täpne, sest vaatlusi oli vähe. □

3.4.2 Vähimruutude meetod

Seda meetodit kasutatakse parameetri hindamiseks siis, kui üldkogumi jaotuse tihedusfunktsioon (tõenäosusfunktsioon) ei ole teada. Teame vaid, et $x_1, x_2, ..., x_n$ on jaotusest, mille keskväärtus on parameetri θ funktsioon, $\theta \in A$ on tundmatu. Parameetri θ hindamiseks vähimruutude meetodil vaadeldakse vaatluste hälbeid keskväärtusest $\mu(\theta)$ ja moodustatakse hälvete ruutude summa:

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

Definitsioon 28 *Parameetri θ vähimruutude hinnanguks nimetatake väärtust $\hat{\theta}$ parameeterruumis A, mille korral $Q(\theta)$ omandab vähima väärtuse*

$$Q(\hat{\theta}) = \min_{\theta \in A} Q(\theta).$$

Näide 29 *Kiirendus vabal langemisel. Eset lastakse langeda n korda teatud punktist. Mõõdetakse aja t jooksul läbitud teepikkused $x_1, x_2, ..., x_n$. Olgu tundmatuks parameetriks θ vabalangemise kiirendus. Teame, et $s = \frac{\theta t^2}{2}$ on teoreetiline teepikkus ehk keskmine teepikkus. Moodustame*

$$Q(\theta) = \sum_{i=1}^{n} (x_i - \frac{\theta t^2}{2})^2.$$

Funktsiooni $Q(\theta)$ miinimumpunkti leiame diferentseerimise abil:

$$\frac{dQ(\theta)}{d\theta} = \sum_{i=1}^{n} 2(x_i - \frac{\theta t^2}{2})(-\frac{t^2}{2}).$$

Võrdsustades tuletise nulliga saame

$$\sum_{i=1}^{n} x_i - \frac{\theta n t^2}{2} = 0 \;\Rightarrow\; \theta = \frac{2\sum_{i=1}^{n} x_i}{nt^2} = \frac{2\bar{x}}{t^2}.$$

Seega vähimruutude hinnang θ-le on

$$\hat{\theta} = \frac{2\bar{x}}{t^2}. \tag{3.3}$$

□

Mida oleks vaja teada antud ülesande lahendamiseks suurima tõepära meetodil?

Vähimruutude meetodi üldistus. Olgu $x_1, x_2, ..., x_n$ erinevate jaotustega juhuslike suuruste $X_1, X_2, ..., X_n$ vaatlused, kus $EX_i = \mu_i(\theta)$, $DX_i = \sigma_i^2$. Nüüd vaatame kaalutud hälvete ruutude summat

$$Q(\theta) = \sum_{i=1}^{n} \lambda_i (x_i - \mu_i(\theta))^2,$$

kus $\lambda_i = \frac{1}{\sigma_i^2}$ on kaalud. Siin laseme suurema dispersiooniga juhusliku suuruse vaatlusel mõjuda väiksema kaaluga ja vastupidi. Vähimruutude hinnang parameetrile θ leitakse nii nagu varemgi $Q(\theta)$ minimiseerimisel.

Sageli pole σ_i^2 teada. Kui aga vaatluste dispersioonid avalduvad näiteks ühe tundmatu σ^2 kordsetena, $k_i\sigma^2$, kus k_i on teada arvud, siis saadav vähimruutude hinnang sisaldab üksnes teadaolevaid väärtusi. Veendu!

Näide 30 *Kiirendus vabal langemisel. Vaatame ajavahemikke $t_1, t_2, \ldots, t_n$, mille jooksul ese läbis vahemaad $x_1, x_2, ..., x_n$. Oletame, et vaatluste dispersioonid on võrdsed, siis $\lambda_i \equiv 1$. Keskmine teepikkus on igal katsel erinev,*

$$EX_i = \frac{\theta t_i^2}{2}.$$

Minimiseerides

$$Q(\theta) = \sum_{i=1}^{n} (x_i - \frac{\theta t_i^2}{2})^2,$$

tuletise leidmise ja nulliga võrdsustamise abil,

$$\frac{dQ(\theta)}{d\theta} = \sum_{i=1}^{n} 2(x_i - \frac{\theta t_i^2}{2})(-\frac{t_i^2}{2}),$$

$$\sum_{i=1}^{n} x_i t_i^2 - \frac{\theta}{2} \sum_{i=1}^{n} t_i^4 = 0,$$

saame kiirenduse vähimruutude hinnanguks:

$$\hat{\theta} = 2\frac{\sum_{i=1}^{n} x_i t_i^2}{\sum_{i=1}^{n} t_i^4}.$$

Erijuhul $t_i \equiv t$ järeldub siit hinnang (3.3). Lihtne on kontrollida, et saime nihketa hinnangu vabalangemise kiirendusele:

$$E\hat{\boldsymbol{\theta}} = \frac{2\sum_{i=1}^{n} t_i^2 EX_i}{\sum_{i=1}^{n} t_i^4} = \frac{2\sum_{i=1}^{n} t_i^2 \frac{\theta t_i^2}{2}}{\sum_{i=1}^{n} t_i^4} = \theta.$$

□

3.4.3 Momentide meetod

Olgu $X \sim F(\theta)$ üldkogumi jaotus ja θ tundmatu parameeter. Üldisemalt, olgu θ vektorparameeter $(\theta_1, \theta_2, \ldots, \theta_l)$. Üldkogumijaotuse kõik karakteristikud (keskväärtus, mediaan, kvantiilid, momendid jm) sõltuvad samuti parameetrist θ. Seega üldkogumijaotuse k-ndat järku moment on θ funktsioon

$$EX^k = \mu_k(\theta).$$

Parameetri θ leidmiseks momentide meetodil koostatakse võrrandisüsteem,

$$\mu_k(\theta) = m_k, \quad k = 1, 2, \ldots, l, \tag{3.4}$$

kus

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

on valimi k-ndat järku moment.

Definitsioon 29 *Võrrandisüsteemi* (3.4) *lahendit* $(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_l)$ *nimetatakse parameetri* $\theta = (\theta_1, \theta_2, \ldots, \theta_l)$ *hinnanguks momentide meetodil.*

Lihtne on näha, et m_k on nihketa hinnanguks $\mu_k(\theta)$-le. Teades nihketa hinnangut mõnele teisele jaotuskarakteristikule, võime ka selle abil võrrandi koostada. Sageli võrdsustatakse näiteks üldkogumi ja valimi dispersioonid, $E(X - EX)^2 = s^2$, et siit tundmatu parameeter avaldada.

Näide 31 *Olgu üldkogumijaotuseks* $U(0, \theta)$*, st ühtlane jaotus lõigul* $[0, \theta]$*, kus lõigu otspunkt* θ *on tundmatu. Jaotuse* $U(0, \theta)$ *keskväärtuseks on* $\theta/2$*. Olgu antud valim* $x_1, x_2, \ldots, x_n$ *sellest jaotusest. Valimikeskmise abil moodustame võrrandi*

$$\bar{x} = \frac{\theta}{2},$$

millest $\hat{\theta} = 2\bar{x}$ *on* θ *hinnanguks momentide meetodil.* □

Näide 32 *Olgu vaatlused* $x_1, x_2, ..., x_n$ *binoomjaotusest* $X \sim B(m, p)$*, kus mõlemad parameetrid on tundmatud. Sellist mudelit kasutatakse avastatud kuritegude arvu kirjeldamiseks kriminalistikas. Siis on* m *kuritegude tegelik arv (näiteks kuus, lihtsuse mõttes konstantne eri kuudel),* p *kuriteo avastamise tõenäosus ehk kuritegude avastamise määr) ja* x_i *on avastatud kuritegude arv kuul* i*. Teame, et binoomjaotuse keskväärtus ja dispersioon avalduvad valemitega* $EX = mp$*,* $DX = mp(1 - p)$*. Parameetrite hindamiseks võrdsustame* EX *ja* DX *nende nihketa hinnangutega valimist (valimimomentidega):*

$$\begin{cases} mp = \bar{x}, \\ mp(1-p) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2. \end{cases}$$

Asendades esimese võrrandi teise, leiame $\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \bar{x}(1 - p)$*, millest saame* p *hinnangu ja seejärel esimesest võrrandist ka* m *hinnangu:*

$$\begin{aligned} \hat{p} &= \frac{\bar{x} - \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}{\bar{x}} \\ \hat{m} &= \frac{\bar{x}}{\hat{p}}. \end{aligned}$$

□

Ühe ja sama parameetri hindamiseks võime erinevate meetoditega saada mõnikord ühesugused, aga mõnikord erinevad hinnangud. Üldjuhul on suurima tõepära hinnangud efektiivsemad kui teiste meetoditega leitud hinnangud.

Näide 33 *Olgu üldkogumijaotus antud tihedusfunktsiooniga*

$$X \sim f(x;\theta) = \theta(1+x)^{-\theta-1},\ x \geq 0.$$

Olgu $x_1, x_2, ..., x_n$ valim sellest jaotusest. Siis valimi logaritmiline tõepärafunktsioon on

$$l(\theta) = \sum_{i=1}^{n} \ln(\theta(1+x_i)^{-\theta-1}] = n\ln\theta - (\theta+1)\sum_{i=1}^{n}\ln(1+x_i).$$

Diferentseerides saame

$$\frac{d}{d\theta}l(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n}\ln(1+x_i),$$

millest nulliga võrdsustamisel avaldame suurima tõepära hinnangu

$$\hat{\theta}_{ST} = \frac{1}{n}\sum_{i=1}^{n}\ln(1+x_i).$$

Vähimruutude hinnangu leidmiseks on vaja teada, kuidas jaotuse keskväärtus sõltub parameetrist θ:

$$EX = \int_{-\infty}^{\infty} xf(x;\theta)dx = \theta\int_{0}^{\infty} x(1+x)^{-\theta-1}dx.$$

Muutujavahetusega $y = 1+x$, saame

$$EX = \theta\int_{1}^{\infty}(y-1)y^{-\theta-1}dy = \theta(\int_{1}^{\infty} y^{-\theta}dy - \int_{1}^{\infty} y^{-\theta-1}dy),$$

millest

$$EX = \frac{1}{\theta-1}.$$

Nüüd saame kirja panna hälvete ruutude summa:

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

Diferentseerides jõuame võrrandini:

$$\sum_{i=1}^{n} 2(x_i - \frac{1}{\theta-1})\,\frac{1}{(\theta-1)^2} = 0,$$

mille lahend on $Q(\theta)$ miinimumpunkt ja ühtlasi vähimruutude hinnang parameetrile θ:

$$\hat{\theta}_{VR} = \frac{n + \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i} = \frac{1+\bar{x}}{\bar{x}},$$

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

Momentide meetodi hinnangu leiame, kui võrdsustame valimi keskväärtuse üldkogumi keskväärtusega:

$$\bar{x} = \frac{1}{\theta-1}.$$

Siit, lahendades θ suhtes, saame

$$\hat{\theta}_{MM} = \frac{1+\bar{x}}{\bar{x}}.$$

Lõppkokkuvõttes, kahe meetodiga jõudsime sama hinnanguni, $\hat{\theta}_{VR} = \hat{\theta}_{MM}$. Suurima tõepära meetod andis nendest erineva hinnangu $\hat{\theta}_{ST}$. Üldjuhul on suurima tõepära hinnangud efektiivsemad teiste meetoditega leitud hinnangutest.□

Peatükk 4

Vahemikhinnang

Oleme õppinud, et valimi andmetelt leitud punkthinnang on mingisugune arv, mis on leitud hindamaks tundmatut üldkogumi (või jautuse) parameetrit. Valimi väärtuste varieeruvuse tõttu ei lange see üldjuhul kokku üldkogumi või jaotuse tegeliku parameetriga. Kuid oskame iseloomustada punkthinnangu täpsust kasutades selleks standard- ja suhtelist viga. Alternatiivne viis hinnangu esitamiseks on vahemikhinnang, mille abil saame edastada infot nii punkthinnangu kui ka selle varieeruvuse kohta.

Definitsioon 30 *Vahemikku I_θ, mis tõenäosusega $1-\alpha$ katab parameetrit θ, nimetatakse* ***vahemikhinnanguks*** *(ka usaldusintervalliks) parameetrile θ usaldusnivool $1-\alpha$. Vahemiku otspunkte $a_1(\mathbf{x})$, $a_2(\mathbf{x})$ (valimifunktsioonid) nimetatakse usalduspiirideks.*

Ingl. keeles: *confidence interval* (*CI*).

Kuidas vahemikhinnangut leida. Illustreerime põhiideed järgmise näite abil.

Näide 34 *Tuletame meelde näidet 18, kus tootja soovis hinnata kohukeste keskmist kaalu, hindamaks kas tööliin on õigesti kalibreeritud. Oletame, et tööliini juhendist leidis ta, et lubatud kaalude varieeruvuseks on $\sigma = 1,5g$. Juhuslikult võetud 20 kohukese kaalud on järgmised:*

28.87	*25.61*	*30.88*	*27.98*	*26.66*	*27.15*	*29.50*	*27.54*	*27.74*	*27.94*
26.42	*28.04*	*28.28*	*28.49*	*28.50*	*24.46*	*29.11*	*29.13*	*27.31*	*26.25*

Oletame, et saadud andmed on realisatsioonid juh. suurustest $X_i \sim N(\mu, \sigma),\ i = 1, ..., 20$. Kasutades valimikeskmist on leitud, et $\hat{\mu} = \bar{x} = 27,793g$. Sellele punkthinnangule vastav hinnangufunktsioon on $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma/\sqrt{n})$.

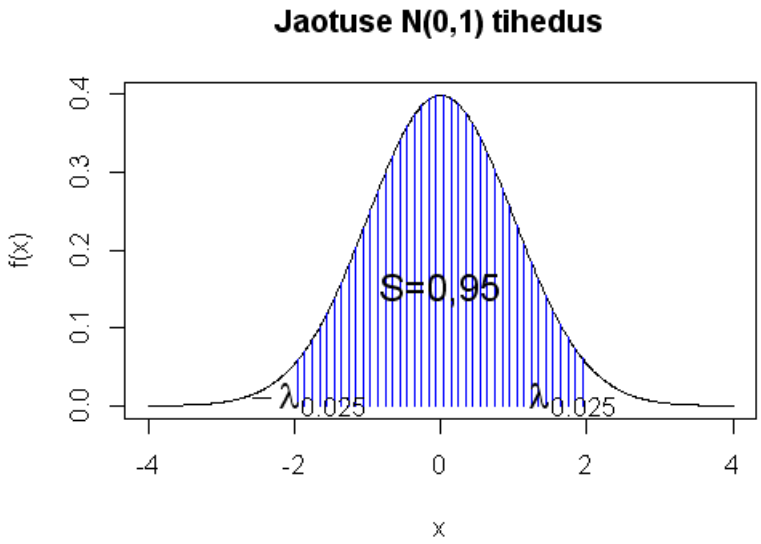
Vahemikhinnangu leidmiseks standardiseerime $\bar{X}$:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Standardsest normaaljaotusest juhusliku suuruse $Z \sim N(0, 1)$ kohta teame, et kehtib (veenduda!)

$$P(-\lambda_{\alpha/2} < Z < \lambda_{\alpha/2}) = 1 - \alpha,$$

kus sümboliga $\lambda_{\alpha/2}$ on tähistatud jaotuse $N(0,\ 1)$ $\alpha/2$-täiendkvantiil. Väidet iseloomustab ka järgmine joonis, kus täiendkvantiili $\lambda_{\alpha/2}$ väärtuse abil moodustatud sinine ala vastabki tõenäosusele $1-\alpha$.



Normaaljaotuse tabelist leiame, et $\lambda_{0,025} = 1,96$ *(veenduda!). Edasi saame:*

$$\begin{aligned}
0,95 &= P(-1,96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1,96) = \\
&= P(-1,96 \cdot \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < 1,96 \cdot \frac{\sigma}{\sqrt{n}}) = \\
&= P(-1,96 \cdot \frac{\sigma}{\sqrt{n}} - \bar{X} < -\mu < 1,96 \cdot \frac{\sigma}{\sqrt{n}} - \bar{X}) = \\
&= P(1,96 \cdot \frac{\sigma}{\sqrt{n}} + \bar{X} > \mu > -1,96 \cdot \frac{\sigma}{\sqrt{n}} + \bar{X}) = \\
&= P(\bar{X} - 1,96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1,96 \cdot \frac{\sigma}{\sqrt{n}}).
\end{aligned}$$

Leidsime ***juhusliku*** *intervalli (sõltub juhuslikust suurusest* $\bar{X}$*). Asendades* $\bar{X}$ *valimi põhjal arvutatud väärtusega, saame selle intervalli realisatsiooni, mida nimetamegi vahemikhinnanguks.*

Kohukese näite korral oleks vahemikhinnang keskmisele kaalule μ *usaldusnivool 95% selline:*

$$\begin{aligned}
I_\mu &= (\bar{x} - 1,96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1,96 \cdot \frac{\sigma}{\sqrt{n}}) = \\
&= (27,793 - 1,96 \cdot \frac{1,5}{\sqrt{20}}; 27,793 + 1,96 \cdot \frac{1,5}{\sqrt{20}}) = \\
&= (27,1356; 28,4504).
\end{aligned}$$

Paneme tähele, et antud vahemiku korral

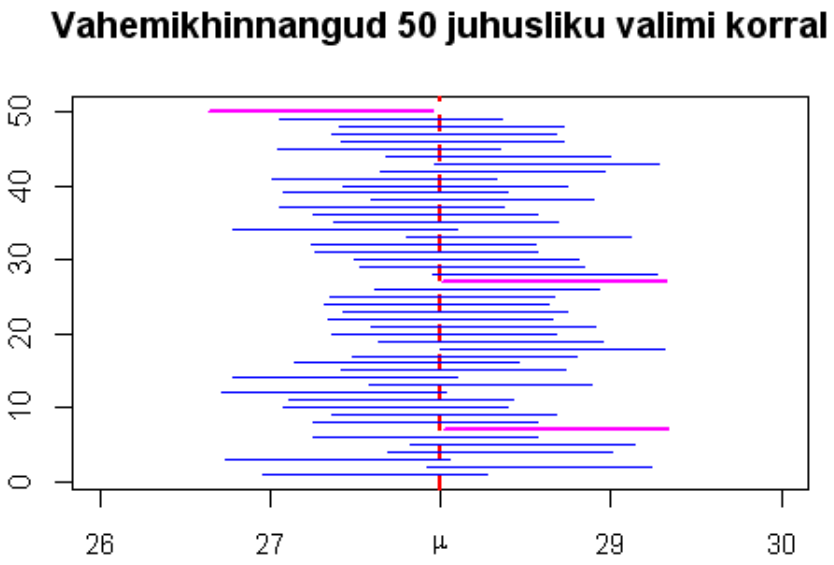
- keskpunktiks on $\bar{X}$, mis on **juhuslik**;
- vahemiku saame siis, kui liidame ja lahutame sellele $\lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, mis on **fikseeritud** üldkogumi standardhälbe- ja valimimahuga;
- vahemikhinnangu laius $c = 2 \cdot \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ on samuti sel juhul konstantne (erinevate valimite korral (sama mahuga) see ei muutu, muutub vaid vahemiku keskkoht).

Üldiselt võib ka laius c olla juhusliku loomuga, kuid seda vaatleme järgnevates peatükkides.

4.1 Üldist vahemikhinnangutest

Iga vahemikhinnang annab väärtuste hulka mingisugusele tundmatule parameetrile ja see väärtuste hulk katab tundmatut parameetrit teatud (üsna suure) tõenäosusega, mida omakordanimetatakse usaldusnivooks $1-\alpha$. Üks võimalik tõlgendus usaldusnivoole on järgmine.

Kui oleks võimalik võtta jaotusest/ üldkogumist suure arvu B erinevat juhuslikku valimit ja arvutada nendelt B vahemikhinnangut, siis oleksid vahemikhinnangud tabanud tundmatut üldkogumi parameetrit ligikaudu $1-\alpha$ juhtudel. Järgmisel joonisel on jaotuse parameeter μ teada (sest tegemist on simuleerimisülesandega) ning jaotusest on genereeritud 50 valimit. Nende põhjal on leitud 50 vahemikhinnangut μ-le ühe ja sama eeskirja ning sama valimimahu korral. Näeme, et enamus nendest tabab parameetri μ väärtust (punane punktiirjoon), kuid kolm roosat vahemiku seda ei tee. Järelikult, on antud näites usaldusnivoo $1-\alpha \approx 3/50 = 0,6$.



Paneme tähele, et vahemikhinnangu laius (c = *ülemine piir* - *alumine piir*, näites 34 oli selleks $c = 2\lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}$) sõltub üldiselt järgmistest näitajatest:

1. täiendkvantiili väärtusest (mida **kõrgem** on usaldusnivoo $1-\alpha$, ehk täpsus, seda **suurem** on täiendkvantiili väärtus ja seega seda **laiem** on vahemikhinnang);
2. uuritava tunnuse dispersioonist (mida **rohkem** andmed varieeruvad, seda **suurem** on dispersioon ja seda **laiem** on vahemik);
3. valimimahust n (mida **vähem** andmeid on valimis, seda **laiem** on vahemikhinnang).

Juhul kui vahemiku laius c on ette antud, nivoo on fikseeritud ning jaotuse dispersioon on teada, saab leida vajaliku valimimahu. Näites 34 näem valem järgmiselt välja:

$$n = \left(2\lambda_{\alpha/2} \cdot \frac{\sigma}{c}\right)^2.$$

Vahemikhinnangu leidmise üldine algoritm:

- Olgu θ tundmatu parameeter ning $x_1, x_2, ..., x_n$ juhuslik valim mingist jaotusest $F(\theta)$.
- Leiame esmalt punkthinnangu θ-le, $\hat{\theta}(\mathbf{x})$ – valimi funtsioonina.
- Paneme kirja vastava hinnangu funktsiooni $\hat{\theta}(\mathbf{X})$ ning leiame selle jaotuse (jaotus sõltub θst).
- Jaotust kasutades leiame punktid $b_1(\theta),\ b_2(\theta)$, nii et

$$P(b_1(\theta) \leq \hat{\theta}(\mathbf{X}) \leq b_2(\theta)) = 1 - \alpha, \tag{4.1}$$

  kus α on ette antud.
- Kui $b_1(\theta)$, $b_2(\theta)$ on rangelt monotoonsed funktsioonid, siis neil leiduvad pöördfunktsioonid, mille abil saab tõenäosusavaldise (4.1) teisendada kujule

$$P\left(a_1(\mathbf{X}) \leq \theta \leq a_2(\mathbf{X})\right) = 1 - \alpha. \tag{4.2}$$

- Asendades teoreetilise valimi $\mathbf{X}$ realiseerunud valimiga $\mathbf{x}$, saame, et

$$I_\theta = (a_1(\mathbf{x}),\ a_2(\mathbf{x}))$$

  on vahemikhinnang parameetrile θ usaldusnivool $1 - \alpha$.

Märkus. Matemaatiliselt korrektsem on öelda, et usaldusvahemik katab parameetrit tõenäosusega $1 - \alpha$, mitte et parameeter kuulub sinna vahemikku (nii rõhutame usaldusvahemiku juhuslikkust). Praktilistes rakendustes ei räägita ometi abstraktsest parameetrist ega selle katmisest. Väljendutakse sisukeskselt, näiteks, 95% tõenäosusega on huvipakkuv kaal $24,14$ kuni $28,45$g.

4.2 Vahemikhinnang normaaljaotuse keskväärtusele

Olgu antud valim normaaljaotusest, $x_1, x_2, \ldots, x_n \leftarrow N(\mu, \sigma)$, kus μ on tundmatu. Tahame leida μ vahemikhinnangut I_μ.

Teoreem 9 (vahemikhinnang jaotuse $N(\mu, \sigma)$ keskväärtusele) *Kui valim on normaaljaotusest $N(\mu, \sigma)$, siis kahepoolne usaldusvahemik parameetrile μ usaldusnivool $1 - \alpha$ on*

$$I_\mu = \bar{x} \pm \lambda_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \qquad \sigma \textit{ teada}, \tag{4.3}$$

$$I_\mu = \bar{x} \pm t_{\alpha/2}(f) \frac{s}{\sqrt{n}}, \qquad \sigma \textit{ tundmatu}, \tag{4.4}$$

kus s on valimi standardhälve, $\lambda_{\alpha/2}$ ja $t_{\alpha/2}(f)$ on vastavalt jaotuste $N(0,1)$ ja $t(f)$ $\alpha/2$-täiendkvantiilid, $f = n - 1$.

Tõestus.
Märgime, et $t_\alpha(f) \to \lambda_\alpha$ protsessis $f \to \infty$, sest t-jaotus läheneb jaotusele $N(0,1)$.
Vahemikhinnangu saame üldisest algoritmist, mis on kirjeldatud punktis 4.1.

Võtame punkthinnanguks valimikeskmise $\hat{\mu} = \bar{x}$. Vastava statistiku korral kehtib

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n}),$$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1). \tag{4.5}$$

Leiame jaotuse $N(0,1)$ täiendkvantiili $\lambda_{\alpha/2}$ (nt tabelist), siis kehtib

$$P(-\lambda_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \lambda_{\alpha/2}) = 1 - \alpha, \tag{4.6}$$

millest sulgude sees teisendades (tõenäosus ei muutu) saame

$$\begin{aligned} 1-\alpha &= P(-\lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}) \\ &= P(\bar{X} - \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}). \end{aligned}$$

Saime juhuslikud usalduspiirid, mis katavad parameetrit μ tõenäosusega $1-\alpha$. Asendades $\bar{X} = \bar{x}$, saame usaldusvahemiku I_μ kujul (4.3).

Teise seose tõestus on analoogiline. Võtame $\hat{\mu} = \bar{x}$. Kuna $\bar{X}$-statistiku normeeritud kujus (4.5) on σ tundmatu, kasutame tema hinnangut $\boldsymbol{s} = (\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2)^{1/2}$. Saadud juhusliku suuruse jaotus on meil teada Teoreemist 6,

$$\frac{\bar{X} - \mu}{\boldsymbol{s}/\sqrt{n}} \sim t(f), \quad f = n-1. \tag{4.7}$$

Leiame tabelist $t_{\alpha/2}(f)$, mille abil saame kirjutada:

$$P(-t_{\alpha/2}(f) < \frac{\bar{X} - \mu}{\boldsymbol{s}/\sqrt{n}} < t_{\alpha/2}(f)) = 1 - \alpha. \tag{4.8}$$

Teisendades sulgude sees tõenäosust muutmata saame

$$1 - \alpha = P(-t_{\alpha/2}(n-1)\frac{\boldsymbol{s}}{\sqrt{n}} < \bar{X} - \mu < t_{\alpha/2}(n-1)\frac{\boldsymbol{s}}{\sqrt{n}}) =$$

$$= P(\bar{X} - t_{\alpha/2}(n-1)\frac{\boldsymbol{s}}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}(n-1)\frac{\boldsymbol{s}}{\sqrt{n}}).$$

Asendame statistikud $\boldsymbol{s}$ ja $\bar{X}$ väärtustega valimist (s ja $\bar{X}$), saame usaldusvahemiku I_μ kujul (4.4).□

4.2.1 Ühepoolsed vahemikhinnangud

Praktikas tekib mõnikord vajadus **ühepoolsete** vahemikhinnangute järele ($a_1 = -\infty$ või $a_2 = \infty$).

Tõenäosuslike väidete kirjapanekul piirdutakse sel juhul vaid ühepoolsete võrratustega, mis toob kaasa $\alpha/2$-täiendkvantiili asendamise α-täiendkvantiiliga.

Näiteks väitest

$$P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \lambda_\alpha\right) = 1 - \alpha,$$

saame usaldusnivool $1-\alpha$ ühepoolse usaldusvahemiku (*näidata iseseisvalt!*)

$$I_\mu = (\bar{x} - \lambda_\alpha \frac{\sigma}{\sqrt{n}},\ \infty).$$

Juhul kui $I_\mu = (27,65; \infty)$ on leitud usaldusnivool $1-\alpha = 0,99$, siis võib seda tõlgendada järgmiselt: tõenäosusega 99% on keskmine kohukese kaal suurem kui $27,65g$.

4.3 Vahemikhinnang normaaljaotuse standardhälbele ja dispersioonile

Analoogiliselt keksväärtusega saame konstrueerida vahemikhinnangut dispersioonile ja standardhälbele. Järgmine teoreem annab vastavat eeskirja normaaljaotuse korral.

Teoreem 10 (vahemikhinnang jaotuse $N(\mu,\sigma)$ dispersioonile ja standardhälbele) *Olgu $x_1, x_2, ..., x_n$ juhuslik valim normaaljaotusest $N(\mu,\sigma^2)$. Siis dispersiooni σ^2 usaldusvahemik usaldusnivool $1-\alpha$ on*

$$I_{\sigma^2} = (k_1^2 s^2, k_2^2 s^2)$$

ja standardhälbe σ usaldusvahemik on

$$I_\sigma = (k_1 s, k_2 s),$$

kus

$$k_1^2 = \frac{f}{q_{\alpha/2}(f)},\quad k_2^2 = \frac{f}{q_{1-\alpha/2}(f)},\quad f = n-1,$$

ja $q_\alpha(t)$ on $\chi^2(f)$-jaotuse α-täiendkvantiil.

Tõestus. Võtame σ^2 punkthinnanguks $\hat{\sigma}^2 = s^2$ ning uurime vastavat statistikut.

$$\boldsymbol{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

kus $X_i \sim N(\mu,\sigma^2)$ on sõltumatud juhuslikud suurused. Arvestades Järelduse 3 tulemust,

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2(f),\quad f = n-1,$$

saame

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

Kasutades jaotuse täiendkvantiile $q_{\alpha/2}(f)$ ja $q_{1-\alpha/2}(f)$ kehtib väide

$$P(q_{1-\alpha/2}(f) < \frac{f}{\sigma^2}\boldsymbol{s}^2 < q_{\alpha/2}(f)) = 1-\alpha,$$

millest

$$P(f\boldsymbol{s}^2/q_{\alpha/2}(f) < \sigma^2 < f\boldsymbol{s}^2/q_{1-\alpha/2}(f)) = 1-\alpha.$$

Asendades $\boldsymbol{s}^2$ tema arvulise väärtusega s^2, saame usalduspiirid dispersioonile. Võttes ruutjuure võrratuse kõikidest pooltest tõenäosus ei muutu. Nii saame usalduspiirid ka standardhälbele (samal usaldusnivool $1-\alpha$). □

Paneme tähele, et standardhälbe σ korral oskasime esmalt leida usaldusvahemiku tema funktsioonile σ^2, millest saime otsitava usaldusvahemiku σ jaoks. Märgime, et kui n on väike, siis I_{σ^2} on väga lai.

4.4 Vahemikhinnang normaaljaotuse keskväärtuste vahele

Sageli tahetakse võrrelda erinevate gruppide keskmisi. Näiteks, kas keskmine viljasaak hektari kohta ühes maakonnas erineb keskmisest teises maakonnas? Leitakse juhuslikult valitud farmide keskmised viljasaagid $\bar{x}$ ja $\bar{y}$. Kui need erinevad, kas see viitab siis tegelikule maakondade erinevusele või on erinevus tingitud valimi juhuslikkusest? Ülesandeks on hinnata vahet $\mu_1 - \mu_2$. Kui $I_{\mu_1-\mu_2}$ asub tervenisti reaaltelje positiivsel poolel, on $\mu_1 > \mu_2$. Usaldusvahemik annab võimaliku erinevuse suuruse (usaldusnivool $1-\alpha$). Kui $I_{\mu_1-\mu_2}$ asub tervenisti reaaltelje negatiivsel poolel, saame väita, et $\mu_1 < \mu_2$. Kui $I_{\mu_1-\mu_2}$ sisaldab 0, ei saa väita, kumb väärtustest μ_1 või μ_2 on suurem.

Siin tuletame vahemikhinnang keskväärtuste vahele normaaljaotuse korral. Esmalt aga tuletame hinnangu normaaljaotuse dispersioonile suurima tõepära meetodil (läheb hiljem vaja).

Lemma 22 *Olgu antud kaks sõltumatut valimit $x_1, x_2, \ldots, x_{n_1}$ ja $y_1, y_2, \ldots, y_{n_2}$ vastavalt normaaljaotustest $N(\mu_1, \sigma)$ ja $N(\mu_2, \sigma)$, ehk jaotuste standardhälbed on võrdsed. Sel juhul nihketa hinnang jaotuse dispersioonile σ^2 kahe valimi põhjal suurima tõepära meetodil on kujul*

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}. \tag{4.9}$$

Tõestus. Kahe valimi ühiseks tõepärafunktsiooniks on

$$L(\mu_1, \mu_2, \sigma^2) = L(\mu_1, \sigma^2) \cdot L(\mu_2, \sigma^2),$$

mis normaaljaotuse tihedusfunktsiooni valemit kasutades saab kuju:

$$L(\mu_1, \mu_2, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{(n_1+n_2)/2}} \cdot e^{-\frac{1}{2\sigma^2}\ Q(\mu_1,\mu_2)}, \tag{4.10}$$

kus

$$Q(\mu_1, \mu_2) = \Big[\sum_{i=1}^{n_1}(x_i-\mu_1)^2 + \sum_{i=1}^{n_2}(y_i-\mu_2)^2\Big].$$

Seosest (4.10) saame logaritmilise tõepärafunktsiooni

$$l(\mu_1, \mu_2, \sigma^2) = -\frac{n_1+n_2}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\ Q(\mu_1, \mu_2). \tag{4.11}$$

Viimase diferentseerimisel ja seejärel nulliga võrdsustamisel saame võrrandisüsteemi:

$$\begin{aligned}
\frac{\partial l}{\partial \mu_1} &: -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(x_i-\mu_1)(-1) = 0,\\
\frac{\partial l}{\partial \mu_2} &: -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(y_i-\mu_2)(-1) = 0,\\
\frac{\partial l}{\partial \sigma^2} &: -\frac{n_1+n_2}{2}\frac{1}{2\pi\sigma^2}2\pi + \frac{1}{2\sigma^4}\ Q(\mu_1,\mu_2) = 0.
\end{aligned}$$

Võrrandisüsteemi lahendid on parameetrite μ_1, μ_2, σ^2 suurima tõepära hinnanguteks:

$$\hat{\mu}_1 = \bar{x},$$

$$\hat{\mu}_2 = \bar{y},$$

$$\hat{\sigma}^2 = \frac{Q(\mu_1,\mu_2)}{n_1+n_2} = \frac{1}{n_1+n_2}\Big(\sum_{i=1}^{n_1}(x_i-\mu_1)^2 + \sum_{i=1}^{n_2}(y_i-\mu_2)^2\Big). \tag{4.12}$$

Arvestades võrdusi

$$s_1^2 = \tfrac{1}{n_1-1}\textstyle\sum(x_i-\bar{x}^2), \tag{4.13}$$

$$s_2^2 = \tfrac{1}{n_2-1}\textstyle\sum(y_i-\bar{y}^2), \tag{4.14}$$

saame $\hat{\sigma}^2$ esitada valimidispersioonide abil:

$$\hat{\sigma}^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2}.$$

Kuna $E\boldsymbol{s}_i^2 = \sigma^2$, siis

$$E\hat{\boldsymbol{\sigma}}^2 = \frac{n_1+n_2-2}{n_1+n_2}\sigma^2.$$

Eelnevat arvestades saame nihketa hinnangu dispersioonile σ^2 kujul:

$$s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}, \tag{4.15}$$

mis ongi teoreemi väide (4.15). □

Saadud tulemuse abil on nüüd lihtne näidata järgmise teoreemii kehtivust.

Teoreem 11 ($I_{\mu_1-\mu_2}$ + kindlate eeldustega dispersioonide kohta) *Olgu $x_1, x_2, \ldots, x_{n_1}$ valim normaaljaotusest $N(\mu_1,\sigma_1)$ ja sellest sõltumatu valim $y_1, y_2, \ldots, y_{n_2}$ normaaljaotusest $N(\mu_2,\sigma_2)$, siis usalduspiirideks keskväärtuste vahele on*

$$I_{\mu_1-\mu_2} = \bar{x}-\bar{y} \pm \lambda_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},\ \textit{kui } \sigma_1^2,\ \sigma_2^2 \textit{ on teada} \tag{4.16}$$

ja

$$I_{\mu_1-\mu_2} = \bar{x}-\bar{y} \pm t_{\alpha/2}(f)\ s\ \sqrt{\frac{1}{n_1}+\frac{1}{n_2}},\quad \textit{kui } \sigma_1^2=\sigma_2^2=\sigma^2 \textit{ on tundmatu,} \tag{4.17}$$

kus

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_i-\bar{x})^2+\sum_{i=1}^{n_2}(y_i-\bar{y})^2}{n_1+n_2-2}, \qquad f = n_1+n_2-2.$$

Tõestus. Võtame $\mu_1-\mu_2$ punkthinnanguks $\bar{x}-\bar{y}$. Vaatame statistikut $\bar{X}-\bar{Y}$. Siin

$$\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i,\ \textit{kus } X_i \sim N(\mu_1,\sigma_1^2),$$

$$\bar{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i,\ \textit{kus } Y_i \sim N(\mu_2,\sigma_2^2),$$

$$E(\bar{X}-\bar{Y}) = \mu_1 - \mu_2,$$
$$D(\bar{X}-\bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Tähistades $d = \sigma_1^2/n_1 + \sigma_2^2/n_2$, vaatame järgmist normeeritud statistikut

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{d}} \sim N(0,1).$$

Seega

$$P\left(-\lambda_{\alpha/2} < \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{d}} < \lambda_{\alpha/2}\right) = 1-\alpha,$$

millest

$$P\Big(\bar{X}-\bar{Y}-\lambda_{\alpha/2}\sqrt{d} < \mu_1-\mu_2 < \bar{X}-\bar{Y}+\lambda_{\alpha/2}\sqrt{d}\Big) = 1-\alpha,$$

kust saame usaldusvahemiku $I_{\mu_1-\mu_2}$ teoreemi 11 esimese juhu jaoks.

Kui $\sigma_1^2 = \sigma_2^2 = \sigma^2$, kuid tundmatu, siis kasutame Lemmas 22 saadud nihketa hinnangut σ^2-le kahe valimi põhjal,

$$\boldsymbol{s} = \sqrt{\frac{\sum_{i=1}^{n_1}(X_i-\bar{X})^2 + \sum_{i=1}^{n_2}(Y_i-\bar{Y})^2}{n_1+n_2-2}}.$$

Seega statistiku kuju tuleb järgmine:

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\boldsymbol{s}\,\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}. \tag{4.18}$$

Jagades statistiku (4.18) lugejat ja nimetajat standarthälbega σ, saame selle viia kujule:

$$\frac{\Big(\bar{X}-\bar{Y}-(\mu_1-\mu_2)\Big)\Big/\Big(\sigma\,\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\Big)}{\boldsymbol{s}/\sigma},$$

kus nüüd lugeja on normaaljaotusega $N(0,1)$. Uurime nimetaja ruudu jaotust

$$\frac{\boldsymbol{s}^2}{\sigma^2} = \frac{\frac{1}{\sigma^2}\sum_{i=1}^{n_1}(X_i-\bar{X})^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n_2}(Y_i-\bar{Y})^2}{n_1+n_2-2}.$$

Vastavalt Järeldusele 3 teame, et lugeja liidetavad on vastavalt $\chi^2(n_1-1)$ ja $\chi^2(n_2-1)$ jaotusega. Teoreemi 3 järgi on lugeja $\chi^2(n_1+n_2-2)$ jaotusega ning Teoreemi 6 põhjal on suurus (4.18) t-jaotusega parameetriga n_1+n_2-2. Saame

$$P\left(-t_{\alpha/2}(f) < \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} < t_{\alpha/2}(f)\right) = 1-\alpha,$$

kus $f = n_1+n_2-2$. Sellest järeldub vahetult usaldusvahemik $I_{\mu_1-\mu_2}$ teisel juhul. □

Eelmise teoreemi väited sõltuvad eeldustest σ_1 ja σ_2 kohta. Sageli on need praktikas tundmatud ja on raske otsustada, kas need on üldkogumis võrdsed või mitte. Siis on abiks järgmine tulemus, mis ei sõltu eeldustest normaaljaotuse dispersioonide kohta. Selle miinuseks on see, et saadud valem on ligikaudne.

Teoreem 12 ($I_{\mu_1-\mu_2}$ ilma eeldusteta dispersioonide kohta) *Olgu $x_1, x_2, \ldots, x_{n_1}$ juhuslik valim normaaljaotusest $N(\mu_1, \sigma_1)$ ja sellest sõltumatu juhuslik valim $y_1, y_2, \ldots, y_{n_2}$ normaaljaotusest $N(\mu_2, \sigma_2)$, kus mõlema üldkogumi dispersioonid on tundmatud. Siis usalduspiirideks keskväärtuste vahele usaldusnivool $(1-\alpha)$ on ligikaudselt*

$$I_{\mu_1-\mu_2} \approx \bar{x} - \bar{y} \pm t_{\alpha/2}(f)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \tag{4.19}$$

kus s_1^2 ja s_2^2 on valimite dispersioonid ja

$$f = \left\lfloor \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}} \right\rfloor. \tag{4.20}$$

Tõestus. Analoogiliselt eelmisele teoreemile moodustame statistiku

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\boldsymbol{s}_1^2}{n_1} + \frac{\boldsymbol{s}_2^2}{n_2}}},$$

kus $\boldsymbol{s}_1^2$ on punkthinnangule (4.13) ja $\boldsymbol{s}_2^2$ on punkthinnangule (4.14) vastavad hinnangufunktsioonid. Huvitume statistiku T jaotusest.

Selleks kirjutame T teisel kujul jagades nimetajat ja lugejat suurusega $d = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$,

$$T = \frac{\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{d}}{\frac{\sqrt{\frac{\boldsymbol{s}_1^2}{n_1} + \frac{\boldsymbol{s}_2^2}{n_2}}}{d}}. \tag{4.21}$$

Paneme tähele, et avaldise (4.21) lugeja on jaotusega $N(0,\ 1)$. Uurime, kas on võimalik viia nimetaja kujule $\sqrt{\frac{Y}{f}}$, kus $Y \sim \chi^2(f)$. Selleks vaatleme nimetaja ruutu,

$$\frac{Y}{f} = \frac{\frac{\boldsymbol{s}_1^2}{n_1} + \frac{\boldsymbol{s}_2^2}{n_2}}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}. \tag{4.22}$$

Hii-ruut jaotuse kohta teame, et kui $Y \sim \chi^2(f)$, siis $EY = f$ ja $DY = 2f$, millest

$$E\left(\frac{Y}{f}\right) = 1 \text{ ja } D\left(\frac{Y}{f}\right) = \frac{2}{f}. \tag{4.23}$$

Tuletame konstandi f väärtuse lähtudes nendest võrranditest. Alustame keskväärtusest:

$$E\left(\frac{Y}{f}\right) = E\left(\frac{\frac{\boldsymbol{s}_1^2}{n_1} + \frac{\boldsymbol{s}_2^2}{n_2}}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = \frac{1}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} E\left(\frac{\boldsymbol{s}_1^2}{n_1} + \frac{\boldsymbol{s}_2^2}{n_2}\right) = \frac{1}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \left(\frac{1}{n_1} E(\boldsymbol{s}_1^2) + \frac{1}{n_2} E(\boldsymbol{s}_2^2)\right).$$

Arvestades, et nihketa hinnangute tõttu on $E(\boldsymbol{s}_1^2) = \sigma_1^2$ ja $E(\boldsymbol{s}_2^2) = \sigma_2^2$, võrdub keskväärtus $E\left(\frac{Y}{f}\right)$ ühega. See kinnitab küll hii-ruudu olemasolu, kuid ei anna vastust sellele, millega võrdub Y ja millega võrdub f. Uurime dispersiooni:

$$D\left(\frac{Y}{f}\right) = D\left(\frac{\frac{\boldsymbol{s}_1^2}{n_1} + \frac{\boldsymbol{s}_2^2}{n_2}}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = \frac{1}{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2} D\left(\frac{\boldsymbol{s}_1^2}{n_1} + \frac{\boldsymbol{s}_2^2}{n_2}\right). \tag{4.24}$$

Kuna valimid $x_1, ..., x_{n_1}$ ja $y_1, ..., y_{n_2}$ on omavahel sõltumatud, siis on ka sõltumatud hinnangufunktsioonid $\boldsymbol{s}_1^2$ ja $\boldsymbol{s}_2^2$. Peame teadma veel $D(\boldsymbol{s}_1^2)$. Kirjutame $\boldsymbol{s}_1^2$ teisel kujul:

$$\boldsymbol{s}_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2 = \sigma_1^2 \frac{1}{n_1 - 1} Z_1,$$

kus $Z_1 = \frac{1}{\sigma_1^2} \sum_{i=1}^{n_1} (X_i - \bar{X})^2 \sim \chi^2(n_1 - 1)$ Järelduse 3 kohaselt. Kasutades hii-ruut jaotuse dispersiooni $(2(n_1 - 1))$, saame

$$D(\boldsymbol{s}_1^2) = D\left(\sigma_1^2 \frac{1}{n_1 - 1} Z_1\right) = \frac{\sigma_1^4}{(n_1 - 1)^2} DZ_1 = \frac{\sigma_1^4}{(n_1 - 1)^2} \cdot 2(n_1 - 1) = \frac{2\sigma_1^4}{(n_1 - 1)}.$$

Asendades saadud tulemused võrrandisee (4.24) saame dispersiooni jaoks kuju:

$$D\left(\frac{Y}{f}\right) = \frac{1}{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2} \left(\frac{\sigma_1^4}{n_1^2(n_1 - 1)} + \frac{\sigma_2^4}{n_2^2(n_2 - 1)}\right) \tag{4.25}$$

Kasutades tulemust hii-ruut jaotuse dispersiooni kohta (4.23), saame et avaldis (4.25) peab võrduma $2/f$. Sellest saame tuletada f:

$$f = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{\sigma_1^4}{n_1^2(n_1 - 1)} + \frac{\sigma_2^4}{n_2^2(n_2 - 1)}} \tag{4.26}$$

Asendades nüüd saadud f avaldise (4.26) võrrandisse (4.22) on võimalik huvilistel tuletada avaldise juhusliku suuruse Y jaoks.

Saadud avaldis (4.26) hii-ruut jaotuse parameetri f jaoks võib anda reaalarvu. Teame aga, et hii-ruut jaotuse parameeter peab olema täisarv. Sellisel juhul on võimalik ümardada saadud f väärtust täisarvuni. Siiski väiksema f väärtuse korral saame laiema vahemikhinnangu $I_{\mu_1 - \mu_2}$, seega ümardamisele on eelistatavam avaldise (4.26) täisosa võtmine. Mõlemad manipulatsioonid parameetriga f muudavad juhusliku suuruse Y jaotust. Kuid on näidatud, et juhuslikk suuruse Y ligikaudseks jaotuseks jääb siiski $\chi^2(f)$.

Tõestuse alustasime statistiku T (4.21) jaotuse otsimisega. Leidsime, et selle lugeja on standardse normaaljaotusega. Nimetaja on kujul $\sqrt{(Y/f)}$, kus $Y \approx \chi^2(f)$. Kasutades Teoreemi 5 tulemus saame, et $T \approx t(f)$, kus f on avaldise (4.26) täisosa. Vahemikhinnangu edasine tuletuskäik on analoogiline eelmistele teoreemidele. □

Märkused viimasele teoreemile:

- Sulud $\lfloor a \rfloor$ tähistavad arvu a täisosa, $a \in \mathcal{R}$.
- Eelmine teoreem kehtib mõlemas olukorras: $\sigma_1^2 \neq \sigma_2^2$ ja $\sigma_1^2 = \sigma_2^2$, seega saab teda rakendada juhul, kui meil puudub mingisugune teave ÜK dispersioonide kohta.
- Meetodit f avaldamiseks tuntakse Welch-Satterthwaite nime all.
- R-is saab leida vahemikhinnangut järgmise käsu abil:

  ```
  t.test(valim1,valim2,var.equal=FALSE)
  ```

  või lihtsalt

  ```
  t.test(valim1,valim2)
  ```

Kirjandus

[1] Kursuse „Tõenäosusteooria ja statistika I“ lemmad ja teoreemid. Kättesaadav kursuse „Tõenäosusteooria ja statistika II“ kodulehelt Moodle's (moodule.ut.ee)

[2] P. L. Meyer (1970) Introductory probability and statistical applications. Addison-Wesley Publishing Company

[3] K. Pärna (2013) Tõenäosusteooria algkursus. Õpik kõrgkoolidele. Tartu Ülikooli Kirjastus

[4] I. Traat (2006) Matemaatilise statistika põhikursus. Tartu Ülikooli Kirjastus

Lisa A. χ^2-jaotuse täiendkvantiilid

α	0.1	0.05	0.01
1	2.71	3.84	6.63
2	4.61	5.99	9.21
3	6.25	7.81	11.34
4	7.78	9.49	13.28
5	9.24	11.07	15.09
6	10.64	12.59	16.81
7	12.02	14.07	18.48
8	13.36	15.51	20.09
9	14.68	16.92	21.67
10	15.99	18.31	23.21
11	17.28	19.68	24.72
12	18.55	21.03	26.22
13	19.81	22.36	27.69
14	21.06	23.68	29.14
15	22.31	25.00	30.58
16	23.54	26.30	32.00
17	24.77	27.59	33.41
18	25.99	28.87	34.81
19	27.20	30.14	36.19
20	28.41	31.41	37.57
21	29.62	32.67	38.93
22	30.81	33.92	40.29
23	32.01	35.17	41.64
24	33.20	36.42	42.98
25	34.38	37.65	44.31
26	35.56	38.89	45.64
27	36.74	40.11	46.96
28	37.92	41.34	48.28
29	39.09	42.56	49.59
30	40.26	43.77	50.89
31	41.42	44.99	52.19
32	42.58	46.19	53.49
33	43.75	47.40	54.78
34	44.90	48.60	56.06
35	46.06	49.80	57.34

Lisa A (jätkub). χ^2-jaotuse täiendkvantiilid

α	0.1	0.05	0.01
36	47.21	51.00	58.62
37	48.36	52.19	59.89
38	49.51	53.38	61.16
39	50.66	54.57	62.43
40	51.81	55.76	63.69
41	52.95	56.94	64.95
42	54.09	58.12	66.21
43	55.23	59.30	67.46
44	56.37	60.48	68.71
45	57.51	61.66	69.96
46	58.64	62.83	71.20
47	59.77	64.00	72.44
48	60.91	65.17	73.68
49	62.04	66.34	74.92
50	63.17	67.50	76.15
51	64.30	68.67	77.39
52	65.42	69.83	78.62
53	66.55	70.99	79.84
54	67.67	72.15	81.07
55	68.80	73.31	82.29
56	69.92	74.47	83.51
57	71.04	75.62	84.73
58	72.16	76.78	85.95
59	73.28	77.93	87.17
60	74.40	79.08	88.38
61	75.51	80.23	89.59
62	76.63	81.38	90.80
63	77.75	82.53	92.01
64	78.86	83.68	93.22
65	79.97	84.82	94.42
66	81.09	85.96	95.63
67	82.20	87.11	96.83
68	83.31	88.25	98.03
69	84.42	89.39	99.23
70	85.53	90.53	100.43