Tartu Ülikool
Matemaatika-informaatikateaduskond
Matemaatika ja statistika instituut

Natalja Lepik

Tõenäosusteooria ja statistika II

Tartu, 2017

Sisukord

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.

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*

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

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:*

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

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 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 3 *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 analoogne: vektori kõik komponendid peavad olema diskreetsed juhuslikud suurused ning jaotus koosneb siis* $n+1$ *arvust koosnevast komplektist, kus esimesed* n *arvu on vektori komponentide võimalikud väärtused ja viimane on neile vastava tulemuste vektori saamise tõenäosus.*

Näide 5 *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 meeldetuletuseks näiteks EX *(keskmine potide arv kahe võetud kaardi hulgast) ja* EY *(mustade kaartide keskmine kahe hulgast).*

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*

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

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

Teoreem 1 *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 4 *Diskreetseid juhuslikke suuruseid X ja Y nimetatakse sõltumatuteks, kui iga $i \in I$ ja iga $j \in J$ korral kehtib võrdus*

$$p_{ij} = p_{i.} p_{.j}$$

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

Definitsioon 5 *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. Pidev juhuslik vektor

Nii pideva kui ka diskreetse juhusliku vektori jaotust saab kirjeldada jaotusfunktsioonide abil. Tuletame siinkohal meelde, et kahemõõtmelise juhusliku vektori korral oli jaotusfunktsioon defineeritud järgmiselt.

Definitsioon 6 *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. $0 \le F_{X,Y}(x,y) \le 1 \ \forall (x,y) \in \mathbb{R}^2$,
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. $P(\{a < X \le b, c < Y \le d\}) = F_{X,Y}(b,d) - F_{X,Y}(a,d) - F_{X,Y}(b,c) + F_{X,Y}(a,c)$.

Tõestus. Esimene omadus tuleneb otse definitsioonist ja sellest, et iga sündmuse tõenäosus on 0 ja 1 vahel. Omadused 2-4 nõuavad tõenäosuse pidevuse omaduse kasutamist, mida siin (ja eelnevas) kursuses ei ole käsitletud, nii et need tulemused võtame praegu lihtsalt teadmiseks. Tõestused on analoogilised juhusliku suuruse jaotusfunktsiooni omaduste tõestamisele. Viimase omaduse tõestamiseks defineerime sündmused $A = \{a < X \le b, c < Y \le d\}$, $B = \{X \le b, Y \le d\}$, $B = \{X \le b, Y \le d\}$, $C = \{X \le a, Y \le d\}$, $D = \{X \le b, Y \le c\}$, $E = \{X \le a, Y \le c\}$ ning paneme tähele, et $B = A \cup C \cup D \cup E$ (mõelge järele, miks see nii on!). Edasine aga tuleb rakendada sündmuste summa tõenäosuse valemit [1, Lemma 2 om. 6]. Tõestuse korrektne lõpetamine on harjutuseks lugejale. □
Jaotusfunktsiooni abil saame defineerida pideva juhusliku vektori. Kasutame siin lihtsuse mõttes jälle kahemõõtmelist juhtu.

Definitsioon 7 *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 8 *Juhuslikke suurusi* X *ja* Y *nimetatakse sõltumatuteks, kui iga* $x, y \in \mathbf{R}$ *korral on sündmused* $\{X \le x\}$ *ja* $\{Y \le y\}$ *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

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

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 6 *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*

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

Seega (kasutades üleminekut polaarkoordinaatidele)

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

Näide 7 *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 Teoreemi 1 põhjal

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

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*

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

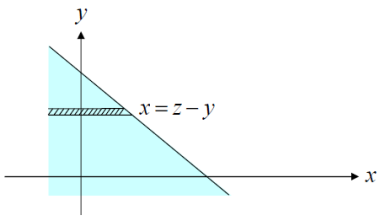
Tõestus. Meil on vaja leida juhusliku suuruse Z tihedusfunktsiooni $f_Z(z)$, mille on võimalik tuletada selle juhusliku suuruse jaotusfunktsioonist $F_Z(z)$. Leiame esmalt jaotusfunktsiooni:

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

Lemma (6) põhjal leiame

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

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

$$\begin{aligned} f_Z(z) = F_Z'(z) &= \int_{-\infty}^{\infty} \frac{\partial}{\partial z} \left(f_Y(y) F_X(z-y) \right) dy \\ &= \int_{-\infty}^{\infty} f_Y(y) f_X(z-y)\, dy \end{aligned}$$

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 8 *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:

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

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

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

Lemma (6) põhjal leiame

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

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

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

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

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

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 9 *Olgu $X \sim B(2, \frac{1}{2})$ ning olgu $Y = |X-1|$. Otseste arvutustega on lihtne näha, et $\mathrm{cov}(X,Y) = 0$ 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. $\mathrm{cov}(X,X) = DX$;
2. $\mathrm{cov}(X,Y) = E(XY) - EX \cdot EY$;
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:

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

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

Omaduse 7 näitamiseks tõestame kõigepealt nn Cauchy-Schwarz'i võrratuse:

$$[E(UV)]^2 \leq EU^2 \cdot EV^2, \tag{2.1}$$

kus U ja V on suvalised juhuslikud suurused.

Selle tõestamiseks paneme tähele, et suvalise $\theta \in \mathcal{R}$ kehtib :

$$0 \leq E(\theta U + V)^2 = \theta^2 EU^2 + 2\theta E(UV) + EV^2.$$

Saime võrratuse tüüpi $a\theta^2 + b\theta + c \geq 0$, kus $a = EU^2 > 0$, $b = 2E(UV)$ ja $c = EV^2$. Ruutfunktsioon on mittenegatiivne siis kui $b^2 - 4ac \leq 0$. (Tuletame siinkohal meelde, et vastava võrratuse reaalarvulised lahendid on $\theta_{1,2} = (-b \pm \sqrt{b^2 - 4ac})/(2a)$, kus $b^2 - 4ac > 0$. Kuid vastava võrratuse korral oleks osa ruutfunktsioonist negatiivne, mis aga on vastuolus eespool kirjeldatuga.)

Tingimusest $b^2 - 4ac \leq 0$ järeldub

$$4\left[E(UV)\right]^2 - 4 \cdot EU^2 \cdot EV^2 \leq 0,$$

ehk

$$[E(UV)]^2 \leq EU^2 \cdot UV^2.$$

Sellega on Cauchy-Schwarz'i võrratus (2.1) tõestatud. Rakendame selle omaduse 7 näitamiseks. Selleks võtame $U = X - EX$ ja $V = Y - EY$. Paneme tähele, et sel juhul kehtib:

$$cov(U,V) = E(UV) - EU \cdot EV = E(UV) = E\left((X-EX)(Y-EY)\right) = cov(X,Y),$$

kuna $EU = E(X - EX) = EX - E(EX) = EX - EX = 0$ (analoogiliselt EV). Tõstame ruutu ning rakendame Cauchy-Schwarz'i võrratuse:

$$[cov(X,Y)]^2 = [E(UV)]^2 \leq EU^2 \cdot EV^2.$$

Arvestades, et $EU^2 = E(X - EX)^2 = DX$ ja $EV^2 = DY$ on lihtne veenduda, et omadus 7 kehtib.

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 10 *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 11 *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 10 *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 11 *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 12 *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:

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

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

mis on tõepoolest suurem või võrdne nulliga. □

Näide 12 *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 13 *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 13 ***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

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_x\sigma_y} + \frac{(y-\mu_2)^2}{\sigma_y^2}\right)\right]. \tag{2.2}$$

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 Tinglik jaotus ja tinglik keskväärtus

Tingliku tõenäosuse mõistega tutvusime aines "Tõenäosusteooria ja matemaatiline statistika". Vaatame siinkohal ühte näidet vajalike teadmiste meelde tuletamiseks.

Näide 14 *(Kanad ja tibud.) Oletame, et ühe päeva jooksul munevad Rohelise talu kanad N muna. N on juhuslik, kuna päevad ei ole vennad ja mõni päev munevad kanad rohkem mune, teisel vähem. Rohelise talu talunik teab, et $N \sim Po(\lambda)$, kus λ on keskmine munade arv päevas.*

Päeva jooksul munetud munad pannakse inkubaatorisse. Tõenäosus, et munast koorub tibu on p ja see on sama kõikide munade puhul. Tibud kooruvad üksteisest sõltumatult.

Olgu X tibude arv, mis on koorunud N-st munast, kusjuures $X \mid N \sim Bin(N, p)$. Avaldis $X \mid N$ tähendab koorunud tibude arvu tingimusel, et inkubaatoris oli N muna (me korraks teeskleme, et see N on meil teada).

Analoogiliselt, olgu Y tibude arv, mis ei koorunud N-munast. Järelikult $X + Y = N$, kus kõik kolm suurust on juhusliku loomuga.

<u>Küsimus</u>: kas X ja Y on sõltumatud? Intuitiivne vastus selle – EI, kuna $X = N - Y$.

Siiski kontrollime, kas kehtib tingimus $p_{ij} = p_{i\cdot}.p_{\cdot j}$, kus antud juhul $p_{ij} = \Pr(X = i, Y = j)$ $(i, j = 1, 2, ..., N)$, $p_{i\cdot} = \Pr(X = i)$ ja $p_{\cdot j} = \Pr(Y = j)$. Antud juhul N on juhuslik suurus väärtustega $0, 1, ...,$. Sündmused $\{N = n\}$, $n = 0, 1, 2, ...$ moodustavad täissüsteemi.

Täistõenäosuse valemi järgi saame:

$$\Pr(X=i,Y=j)=\sum_{n=0}^{\infty}\Pr(X=i,Y=j\mid N=n)\Pr(N=n).$$

Avaldis sisaldab summat, mille liidetavate arv on lõpmatu. Kuid vaatame mõningaid erijuhte:

$$\Pr(X=3,Y=5\mid N=10)=0,\quad \Pr(X=3,Y=5\mid N=2)=0.$$

Tuleb välja, et ainus variant, kus tõenäosus erineb nullist on siis kui $n=i+j$:

$$\begin{aligned}\Pr(X=i,Y=j)&=\Pr(X=i,Y=j\mid N=i+j)\Pr(N=i+j)\\&=\Pr(X=i,Y=j\mid X+Y=i+j)\Pr(N=i+j)\\&=\Pr(X=i\mid N=i+j)\Pr(N=i+j)\end{aligned}$$

kus viimases avaldises $Y=j$ on ära võetud, sest see ei anna enam mingit uut infot.

Juhuslike suuruste $X|N$ ja N jaotused on teada, ka vastavad tõnäosusfunktsioonid on teada, millest saame, et

$$\begin{aligned}\Pr(X=i,Y=j)&=\left[C_{i+j}^{i}p^{i}(1-p)^{i+j-i}\right]\cdot\frac{\lambda^{i+j}e^{-\lambda}}{(i+j)!}\\&=\frac{p^{i}(1-p)^{j}}{i!j!}e^{-\lambda}\lambda^{i}\lambda^{j}\\&=\frac{(\lambda p)^{i}}{i!}\frac{(\lambda(1-p))^{j}}{j!}e^{-\lambda(p+1-p)}\\&=\frac{(\lambda p)^{i}}{i!}e^{-\lambda p}\cdot\frac{(\lambda(1-p))^{j}}{j!}e^{-\lambda(1-p)}.\end{aligned}$$

Jõudsime kahe tõenäosusfunktsiooni korrutiseni, kus $X\sim Po(\lambda p)$ ja $Y\sim Po(\lambda(1-p))$. Seega, juhuslikud suurused X ja Y on sõltumatud.

Miks see nii juhtus? Intuitiivselt tundus, et seos kahe juhusliku suuruse vahel on ju olemas. Aga me ei võtnud ju arvesse, et N on samuti juhuslik suurus, mitte konstant. Kursusest "Tõenäosus ja matemaatiline statistika" on teada, et kui $X\sim Po(\mu_1)$ ja sellest sõltumatu juhuslik suurus $Y\sim Po(\mu_2)$, siis $X+Y\sim Po(\mu_1+\mu_2)$. Sama tulemust saime ka antud näites kasutades tinglikustamise võtet. □

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

Näide 15 *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.

Olgu antud diskreetne juhuslik vektor $(X,\ Y)$ väärtustega vastavalt $(x_1, x_2, ..)$ ja $(y_1, y_2, ...)$, kus väärtuste hulk on kas lõplik või loenduv. Siis saame defineerida tingliku keskväärtust järgmiselt.

Definitsioon 14 *Keskväärtust omava diskreetse juhusliku suuruse X tinglikuks keskväärtuseks tingimusel, et sündmus $\{Y = y_j\}$ (mille korral $P(\{Y = y_j\}) > 0$) toimus, nimetatakse arvu*

$$E(X \mid \{Y = y_j\}) = \sum_{i \in I} x_i \Pr(\{X = x_i\} \mid \{Y = y_j\}),$$

kus $(x_i, \Pr(\{X = x_i\} \mid \{Y = y_j\}))$, $i \in I$ on juhusliku suuruse X tinglik jaotus tingimusel, et $\{Y = y_j\}$ toimus.

Tuletame meelde, et kehtib seos:

$$\Pr(\{X = x_i\} \mid \{Y = y_j\})) = \frac{\Pr(\{X = x_i\}, \{Y = y_j\}))}{\Pr(\{Y = y_j\}))}.$$

Näide 16 *Leiame vastused eelmises näites püstitatud küsimustele.*

1. Millega võrdub summa Y keskväärtus (teiste sõnadega kahel täringul oodatav silmade arvude summa) juhul, kui on teada, et esimesel tuli 2 silma?

$$E(Y|\{X_1 = 2\}) = \sum_j y_j P\left(\{Y = y_j\}|\{X_1 = 2\}\right) =$$

$$= 2 \cdot 0 + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} + 7 \cdot \frac{1}{6} + 8 \cdot \frac{1}{6} + 9 \cdot 0 + ... + 36 \cdot 0 = 11/2.$$

2. Millega võrdub esimese täringu keskmine saadud silmade arv juhul kui on teada, et summa tuli 5?

$$E\left(X_1|\{Y = 5\}\right) = \sum_i x_i P\left(\{X_1 = x_i\}|\{Y = 5\}\right)$$

$$= \sum_i x_i \cdot \frac{P\left(\{X_1 = x_i\} \cap \{Y = 5\}\right)}{P\{Y = 5\}} = \sum_{x=1}^{4} x \frac{1/36}{4/36} = \frac{5}{2}.$$

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

Teoreem 2 *Olgu B_j, $j \in J$ (kus J on kas lõplik või loenduv hulk) sündmuste täissüsteem. Siis kehtib võrdus*

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

Tavaliselt on sobivateks sündmusteks B_j mingi teise juhusliku suuruse Y väärtustele vastavad sündmused $\{Y = y_j\}$.

Tõestus. Alustame võrdsuse paremast poolest:

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

Näide 17 *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.)*

Tingliku keskväärtust saab defineerida ka pideva juhusliku vektori $(X,\ Y)$ jaoks, mille ühine tihedusefunktsioon on $f(x,y)$.

Definitsioon 15 *Olgu $(X,\ Y)$ kahemõõtmeline pidev juhuslik vektor. Siis juhusliku suuruse X tinglik keskväärtus tingimusel $\{Y=y\}$ on*

$$E\left(X \mid \{Y=y\}\right) = \int_{-\infty}^{+\infty} x\cdot f_X\left(x \mid \{Y=y\}\right) dx,$$

kus

$$f_X\left(x \mid \{Y=y\}\right) = \frac{f(x,y)}{f_Y(y)}.$$

Näide 18 *Olgu juhusliku suuruse $(X,\ Y)$ ühistihedus on antud järgmise valemiga:*

$$f(x,y) = \begin{cases} 8xy, & \text{kui } 0<x<y<1,\\ 0, & \text{vastasel juhul.}\end{cases}$$

Leida $E(X \mid \{Y=y\})$. Tingliku tõenäosuse definitsiooni rakendamiseks peame teadma $f_Y(y)$ ja $f_X\left(x \mid \{Y=y\}\right)$. Leiame:

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)\,dx = \int_0^y 8xy\,dx = 4yx^2\big|_0^y = 4y^3;$$

$$f_X\left(x \mid \{Y=y\}\right) = \frac{f(x,y)}{f_Y(y)} = \frac{8xy}{4y^3} = \frac{2x}{y^2},\ \text{kui } 0<y<x<1.$$

Lõpuks saame leida tingliku keskväärtust:

$$E\left(X \mid \{Y=y\}\right) = \int_{-\infty}^{+\infty} x\cdot f_X\left(x \mid \{Y=y\}\right) dx = \int_0^y x\frac{2x}{y^2}\,dx = \left.\frac{2x^3}{3y^2}\right|_0^y = \frac{2y}{3}.$$

Paneme tähele, et $E(X|\{Y=y\})$ on muutuja y funktsioon. Tähistame seda funktsiooni korraks $h(y)$. Eelmises näites $h(y)=\frac{2y}{3}$. Defineerime keskväärtust $E(X \mid Y)$ kui juhusliku

suurust $h(Y)$. Eelmises näiteks oleks $h(Y) = \frac{2Y}{3}$. Teisiti öeldes, $E(X|\{Y = y\})$ on mingi arv ühe konkreetse y väärtuse korral. Näiteks, kui eelmises näites võtta $y = 1/2$, siis $E(X|\{Y = 1/2\}) = \frac{2 \cdot 0.5}{3} = 1/3$. Samal ajal avaldis $E(X \mid Y)$ on juhuslik suurus, mis sõltub juhuslikust suurusest Y.

Kuna $E(X \mid Y)$ on juhuslik suurus, siis saame rääkida selle keskväärtusest $E\left(E(X \mid Y)\right)$. Tähtis on aru saada, et sisemine keskväärtus on võetud arvestades juhusliku suuruse X jaotust tingimusel $\{Y = y\}$. Välimine keskväärtus on aga võetud arvestades Y jaotust.

Lemma 15 *(Tingliku keskväärtuse keskväärtus.) Tingliku keskväärtuse kohta kehtib järgmine tulemus:*

$$E\left(E(X \mid Y)\right) = EX. \tag{2.3}$$

Teisiti öeldes, tingliku keskväärtuse keskväärtus on võrde "tingimata" keskväärtusega. Sageli kasutatakse antud tulemust vastupidises suunas, ehk leitakse juhusliku suuruse keskväärtus kasutades tinglikustamise võtet.

Tõestus. Tõestame pideva juhusliku vektori $(X,\ Y)$ jaoks. Diskreetne juht on harjutuseks lugejale. Definitsiooni 15 kohaselt,

$$E\left(X \mid \{Y = y\}\right) = \int_{-\infty}^{+\infty} x \cdot f_X\left(x \mid \{Y = y\}\right) dx = \int_{-\infty}^{+\infty} x \cdot \frac{f(x,y)}{f_Y(y)} dx.$$

Tähistame jälle $E\left(X \mid \{Y = y\}\right) := h(y)$. Juhusliku suurusena vaadatuna saame leida selle keskväärtust (kasutades juhusliku suuruse funktsiooni keskväärtuse tulemust, vt Lemma 24 moodle dokumendis Teoreemide ja lemmade kordamine):

$$E\left(E(X \mid Y)\right) = E\left(h(Y)\right) = \int_{-\infty}^{+\infty} h(y) \cdot f_Y(y) dy = \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} x \cdot \frac{f(x,y)}{f_Y(y)} dx\right] \cdot f_Y(y) dy.$$

Kui kõik keskväärtused eksisteerivad, siis saame muuta integraalide järjekorda:

$$E\left(E(X \mid Y)\right) = \int_{-\infty}^{+\infty} x \left[\int_{-\infty}^{+\infty} f(x,y) dy\right] dx = \int_{-\infty}^{+\infty} x f_X(x) dx = EX,$$

kus rakendasime Lemma 6 omadust 2.□

Näide 19 *Leiame eelmises näites EX kasutades nii Lemmat 15 kui ka pideva juhusliku suuruse keskväärtuse otsest definitsiooni. Lemma 15 järgi saame:*

$$EX = E\left(E(X \mid Y)\right) = \int_{-\infty}^{+\infty} E(X \mid \{Y = y\}) f_Y(y) dy = \int_0^1 \frac{2y}{3} 4y^3 dy = \left.\frac{8y^5}{15}\right|_0^1 = \frac{8}{15}.$$

Pideva juhusliku suuruse definitsiooni rakendamiseks peame teadma juhsuliku suuruse X marginaalse tihedusfunktsiooni:

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y) dy = \int_x^1 8xy\, dy = \left. 4y^2 x\right|_x^1 = 4(x - x^3),\ 0 < x < 1.$$

Seejärel leiame keskväärtuse:

$$EX = \int_{-\infty}^{+\infty} x f_X(x) dx = \int_0^1 x 4(x - x^3)\, dx = \left.\left(\frac{4y^3}{3} - \frac{4y^5}{5}\right)\right|_0^1 = \frac{8}{15},$$

mis langeb kokku eespool leituga.

Näide 20 *Ainest '"Tõenäosus ja matemaatiline statistika 'ön teada, et kui $X \sim Geom(p)$ (geomeetrilise jaotusega juhuslik suurus), siis $EX = 1/p$. Kuna $X \in \{1, 2, ...\}$ (loenduv hulk), siis keskväärtuse otsene definitsiooni rakendame keskväärtuse leidmisel võib osutuda keeruliseks. Tuleks leida järgmise summa:*

$$EX = \sum_{k=1}^{\infty} k \cdot \Pr\{X = k\} = \sum_{k=1}^{\infty} k \cdot p(1-p)^{k-1} = ...,$$

kus k on juhusliku suuruse X kõikvõimalikud väärtused.

Võime aga kasutada nn tinglikustamise võtet. Olgu $X \sim Geom(p)$, näiteks mündi viskamiste arv kuni esmakordse kulli saamiseni (k.a.) ja $p = 1/2$. Defineerime lisaks veel ühe diskreetse juhusliku suuruse:

$$Y = \begin{cases} 1, & \textit{kui huvipakkuv sündmus realiseerus 1. katsel, nt. 1. viske tulemuseks on kull} \\ 0, & \textit{vastasel juhul.} \end{cases}$$

Kasutades Lemmat 15 saame:

$$EX = E\left(E(X|Y)\right),$$

kus välimine keskväärtus on võetud arvestades Y jaotust. Kuna $Y \sim Be(p)$, siis välimise keskväärtuse leidmiseks tuleb leida summa kahe liidetavaga:

$$\begin{aligned} EX = E\left(E(X|Y)\right) &= \sum_{y} E(X|Y=y) \cdot \Pr(Y=y) \\ &= E(X|Y=0) \cdot \Pr(Y=0) + E(X|Y=1) \cdot \Pr(Y=1) \\ &= E(X|Y=0) \cdot (1-p) + E(X|Y=1) \cdot p. \end{aligned} \tag{2.4}$$

Leiame $E(X|Y=0)$ ja $E(X|Y=1)$ eraldi. Alustame viimasest. Rakendame Definitsiooni 14:

$$\begin{aligned} E(X|Y=1) &= \sum_{k=1}^{\infty} k \cdot \Pr(X = k \mid Y = 1) \\ &= \sum_{k=1}^{\infty} k \cdot \frac{\Pr(X=k, Y=1)}{\Pr(Y=1)} \end{aligned}$$

Kuna $Y \sim Be(p)$, siis $\Pr(Y=1) = p$. Sündmus $\{Y=1\}$ tähendab seda, et huvipakkuv sündmus (nt kull) realiseerus kohe 1. katsel. Järelikult

$$\Pr(X=k, Y=1) = \begin{cases} \Pr(X=1) = p, & \textit{kui k=1,} \\ 0, & \textit{vastasel juhul.} \end{cases}$$

(Tuletame siinkohal meelde, et $\Pr(X=k) = p(1-p)^{k-1}, k = 1, 2, ...$). Seega $E(X|Y = 1) = 1$.

Leiame nüüd $E(X|Y=0)$ analoogiliselt eelmisega,

$$\begin{aligned} E(X|Y=0) &= \sum_{k=1}^{\infty} k \cdot \Pr(X = k \mid Y = 0) \\ &= \sum_{k=1}^{\infty} k \cdot \frac{\Pr(X=k, Y=0)}{\Pr(Y=0)} \end{aligned}$$

Sündmus $\{X = k, Y = 0\}$ *on samaväärne sündmusega* $\{X = k\}$, $k = 2, 3, \ldots$ *Seega,*

$$\Pr(X = k, Y = 0) = \begin{cases} \Pr(X = k) = p(1-p)^{k-1}, & \textit{kui } k\textit{=2,3...}, \\ 0, & \textit{vastasel juhul.} \end{cases}$$

Teisiti kirjutades,

$$\begin{aligned} E(X|Y=0) &= \sum_{k=2}^{\infty} k \cdot \frac{p(1-p)^{k-1}}{1-p} \\ &= \sum_{k=2}^{\infty} k \cdot p(1-p)^{k-2} \\ &\overset{l:=k-1}{=} \sum_{l=1}^{\infty} (l+1)p(1-p)^{l-1} \\ &= \sum_{l=1}^{\infty} l \cdot p(1-p)^{l-1} + \sum_{l=1}^{\infty} p(1-p)^{l-1} = EX + 1. \end{aligned}$$

Asendades avaldisse 2.4 saame võrrandi

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

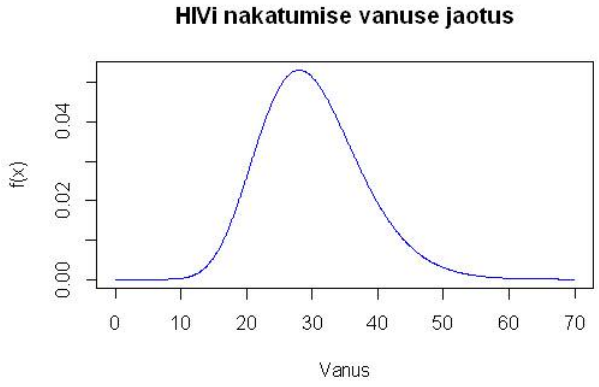
mille lahendades EX *suhtes saamegi, et* $EX = 1/p$. □

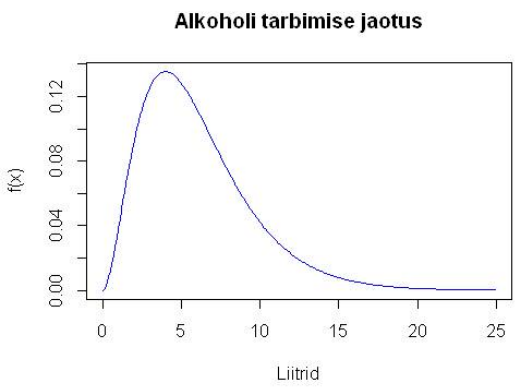
2.6 Kolm tähtsat pidevat jaotust statistikas ja seosed nende vahel

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.

2.6.1 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:



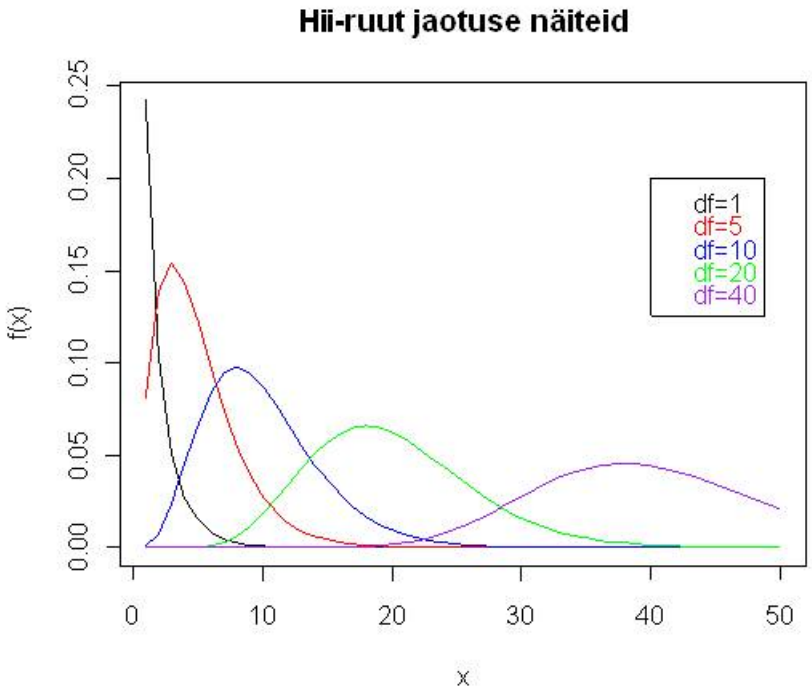


Definitsioon 16 *Juhuslik suurus X on hii-ruut jaotusega parameetriga (ka vabadusastmete arvuga) f, $X \sim \chi^2(f)$, kui tema tihedusfunktsioon avaldub seosega*

$$f(x) = kx^{\frac{f}{2}-1}e^{-\frac{x}{2}}, \quad x \geq 0, \tag{2.5}$$

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

$$f(x) = \frac{1}{\sqrt{2\pi x}}e^{-\frac{x}{2}} \tag{2.6}$$

Tõestus on harjutus lugejale.

Lemma 17 *(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*

$$M(t) = (1-2t)^{-f/2}, \quad t < \frac{1}{2}. \tag{2.7}$$

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 $DX = EX^2 - (EX)^2$,

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

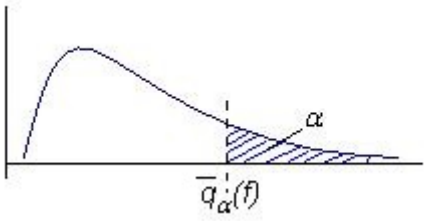
millest $DX = f^2 + 2f - f^2 = 2f$. □

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:*

$$P(X \geq \bar{q}_\alpha) = \alpha. \tag{2.8}$$

Täiendkvantiili mõistet iseloomustab ka järgmine joonis:



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

Lemma 18 *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*

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n), \tag{2.9}$$

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 \sim \chi^2(n-1), \tag{2.10}$$

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 (2.6) 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 (2.9).

Vaatame nüüd teist väidet kujul

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2. \tag{2.11}$$

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 (2.11) 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 $\mathbf{Y} = \mathbf{CX}$, kus $\mathbf{C}$ on ortogonaalne maatriks, $\mathbf{CC}^T = \mathbf{I}$. Vektori $\mathbf{Y}$ komponendid Y_i on normaaljaotusega, sest tegu on vektori $\mathbf{X}$ lineaarkombinatsiooniga. Vastavaks keskväärtusvektoriks saame:

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

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

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

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 (2.11) 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 (2.9) saame,

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

millega oleme ka seose (2.10) tõestanud. □

Lemma 19 *(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

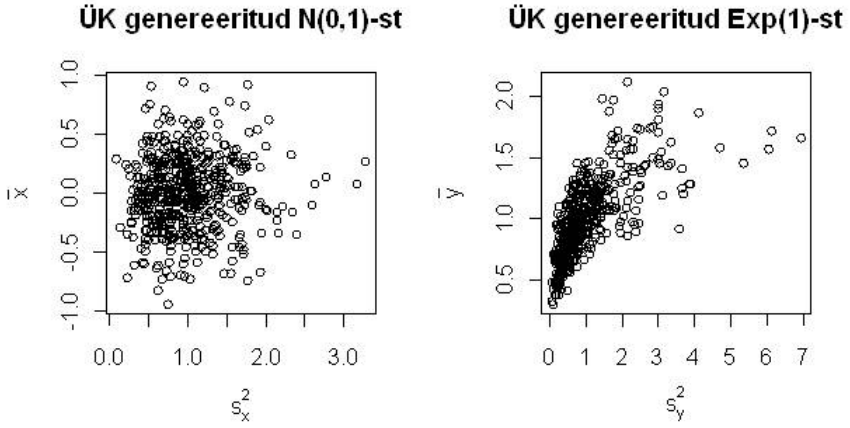
$$\sum_{i=2}^{n} Y_i^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2. \ \square$$

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 21 *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 väite (2.10) 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 väidet (2.10) saamegi järelduse 2. väidet samutigi tõestatud. □

2.6.2 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.

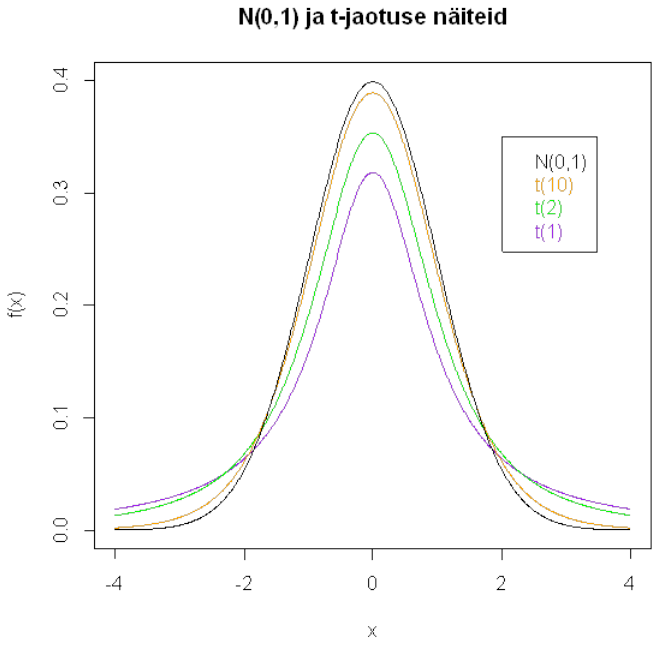
Definitsioon 18 *Juhuslik suurus X on t-jaotusega vabadusastmete arvuga f, $X \sim t(f)$, kui tema tihedusfunktsioon avaldub kujul*

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

kus

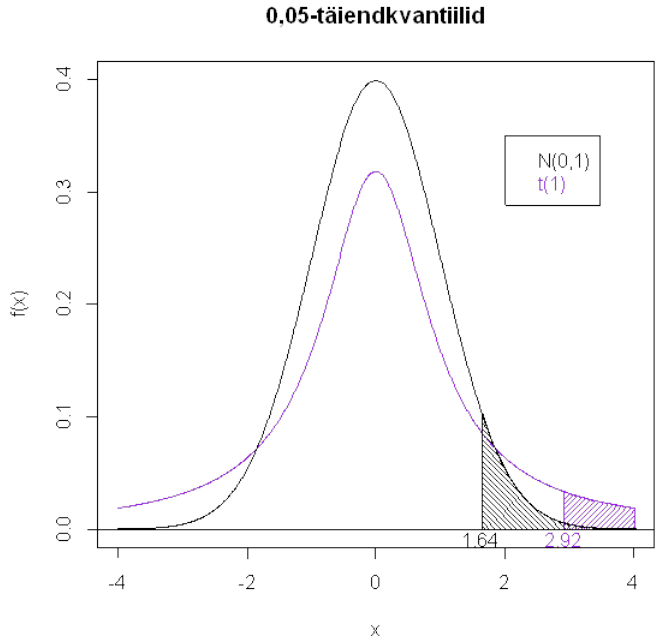
$$k = \frac{\Gamma(\frac{f+1}{2})}{\sqrt{f\pi}\Gamma(\frac{f}{2})}, \quad f \in \{1, 2, \ldots\}.$$

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

$$f_Z(x) = \int_0^\infty v f_X(vx) f_V(v) dv. \tag{2.12}$$

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 (2.5) argumendi x korrutisega fx, saame

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

kus $k_1 = k \cdot f^{f/2}$ 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 (2.12) 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:

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

Paneme tähele, et

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

ja teoreemi 4 põhjal

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

Lisaks, on Lemma 19 põhjal suurused U ja V sõltumatud (parameetrid μ, σ ning valimimaht n on ju konstandid). Siis teoreemi 5 põhjal saame

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

□

2.6.3 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*

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

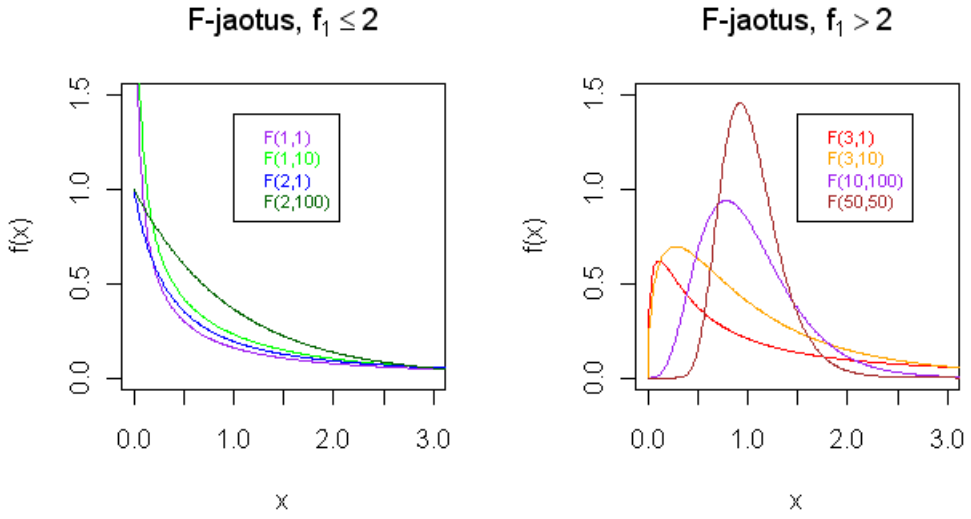
kus normeeriv konstant omab kuju

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

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.

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:

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

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.*

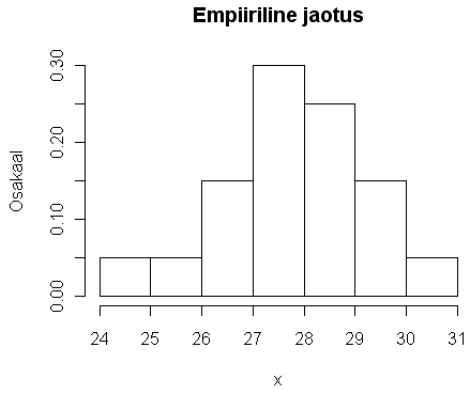
Inglise keeles on head iseloomulikud sõnad nende kahe mõiste eristamiseks: *estimate* – punkthinnang, *estimator* – hinnangufunktsioon. Üldjuhul nimetame teoreetilise valimi funktsiooni statistikuks. Hinnangufunktsioon on selline statistik, mida kasutatakse parameetri hindamise eesmärgil.

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 22 *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 23 *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,

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

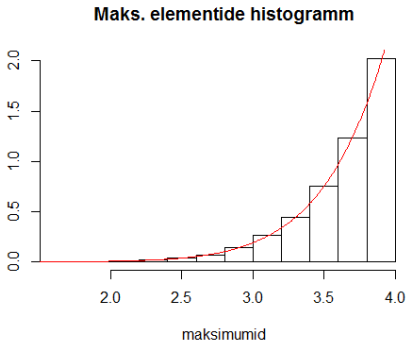
kus $f(x) = \frac{1}{\theta}$ on jaotuse $U(0, \theta)$ tihedusfunktsioon, $x \in [0, \theta]$. Jaotusfunktsioon $F(x)$ on seega

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

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 24 *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:

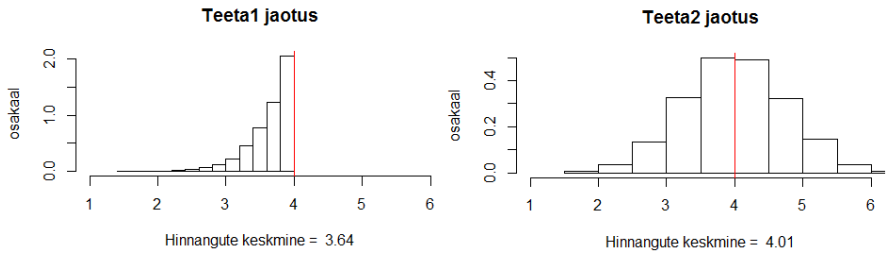
$$E\hat{\boldsymbol{\theta}}_1 = \int_{-\infty}^{\infty} x \cdot f_{\hat{\boldsymbol{\theta}}_1}(x)dx,$$

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 20 *Ruutkeskmist viga on võimalik esitada järgmisel alternatiivsel kujul:*

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

Tõestuses tuleb lähtuda MSE definitsioonist ja on harjutuseks lugejale.

Näide 25 *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 $g(x) = (x-\theta)^2$:

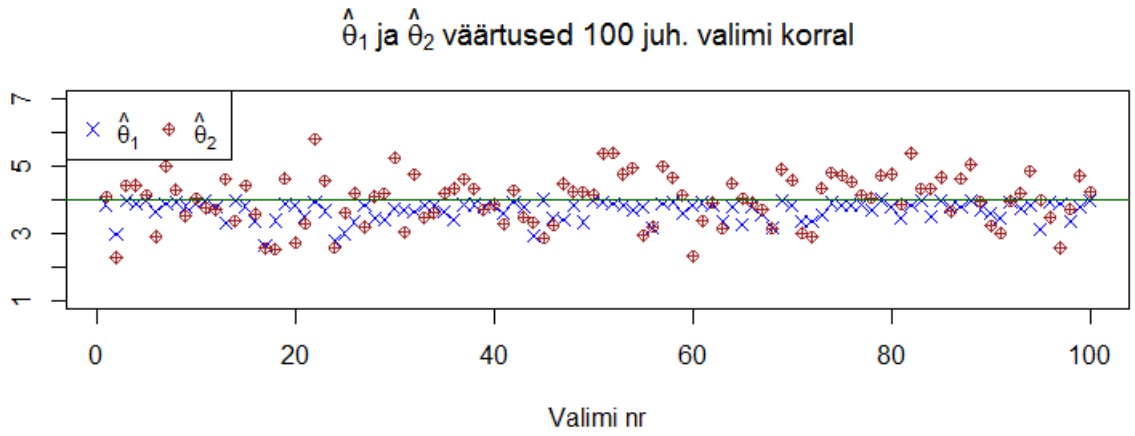
$$MSE(\hat{\boldsymbol{\theta}}_1) = E[g(x)] = \int_0^{\theta}(x-\theta)^2 f_{\hat{\boldsymbol{\theta}}_2}(x)dx = \int_0^{\theta}(x-\theta)^2 \cdot \frac{nx^{n-1}}{\theta^n}dx = \frac{2\theta^2}{(n+2)(n+1)}.$$

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 21 *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 21 tingimused 1-2 saab kokku võtta ka järgmise tingimuse abil:

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

Näide 26 *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 27 *Näites 22 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 22 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 28 *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.

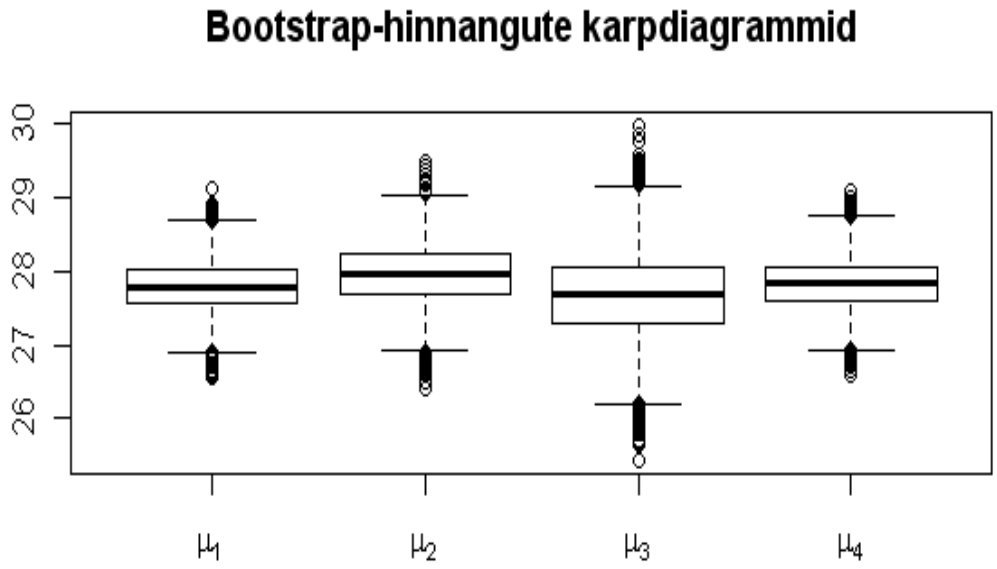
```
x=c(24.46,25.61,26.25,26.42,26.66,27.15,27.31,27.54,27.74,27.94,27.98,
28.04,28.28,28.49,28.50,28.87,29.11,29.13,29.50,30.88)
 k=10000

 # 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 $\hat{\mu}_3$ *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 $\hat{\theta}^*$.

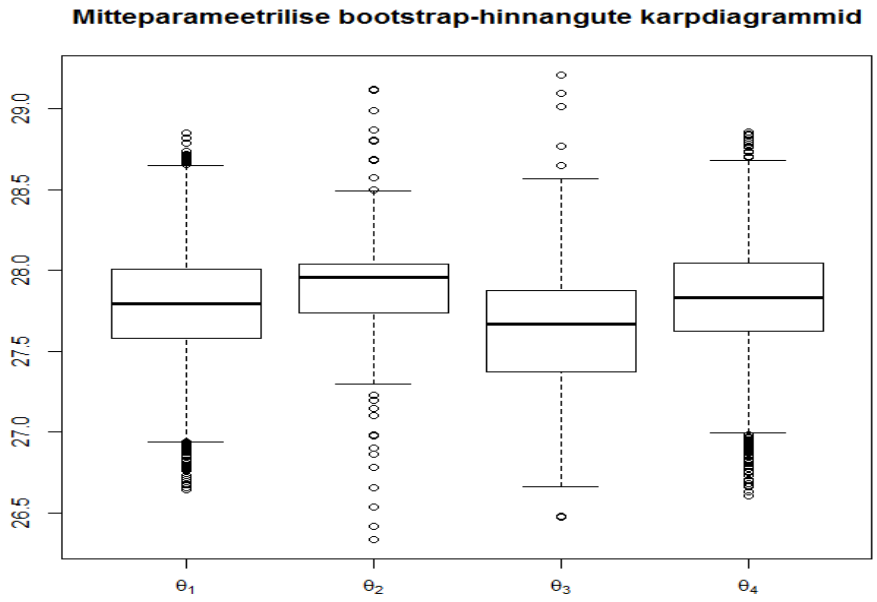
Hinnangu $\hat{\theta}$ standardviga leitakse analoogiliselt eelmise versiooniga,

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

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

Näide 29 *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 meetodi** 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 22 *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 30 *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 31 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 32 *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

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

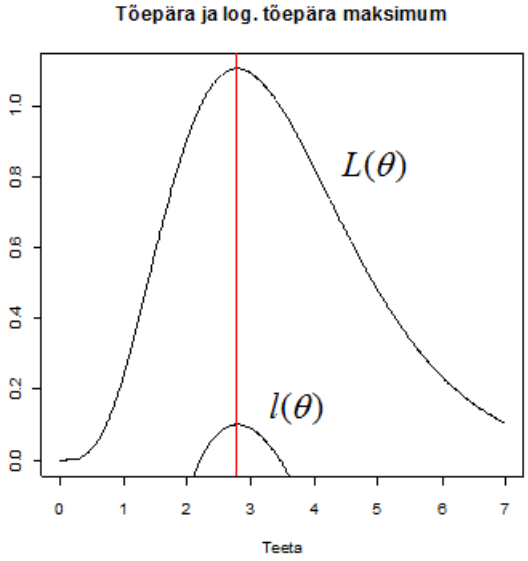
ja logaritmilise tõepärafunktsiooni

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

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 33 *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 34 *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

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

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 35 *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 36 *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 37 *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

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

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.□

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 38 *Tuletame meelde näidet 22, 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 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma/\sqrt{n})$.

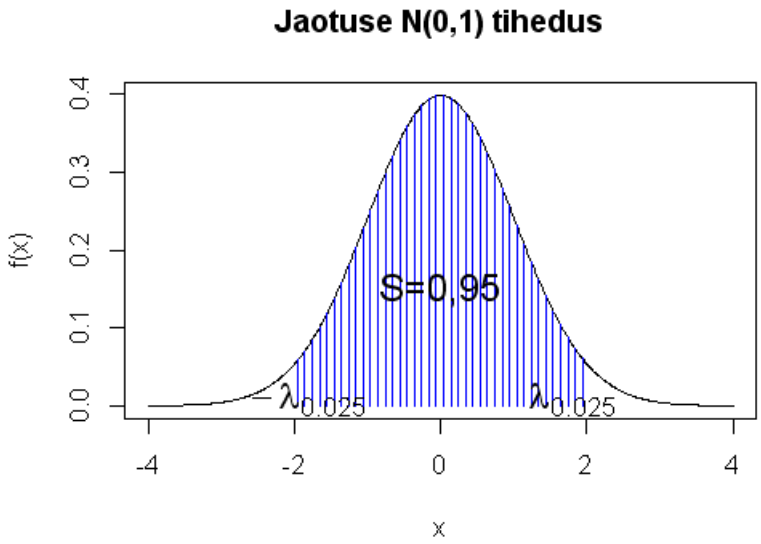
Vahemikhinnangu leidmiseks standardiseerime $\bar{X}$:

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

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

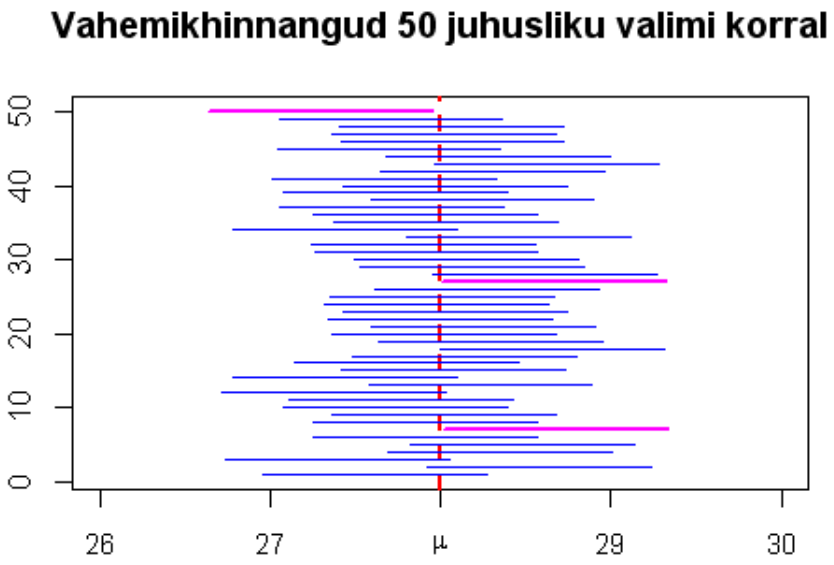
- keskpunktiks on $\bar{X}$, mis on **juhuslik**;
- vahemiku saame siis, kui liidame ja lahutame sellele $\lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, mis on **fikseeritud** üldkogumi standardhälbe- ja valimimahuga;
- vahemikhinnangu laius $c = 2 \cdot \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ on samuti 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 $1-\alpha \approx 1-3/50 = 0,94$.



Paneme tähele, et vahemikhinnangu laius (c = *ülemine piir* - *alumine piir*, näites 38 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 38 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

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

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

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 23 *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:

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

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

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

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

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

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 23 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 võrrandisse (4.24) saame dispersiooni jaoks kuju:

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

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)
  ```

4.5 Vahemikhinnang muutusele

Järgnev peatükk põhineb õpikul Traat(2006), lk. 102-103.

On ülesandeid, kus mõõdetakse samu objekte enne ja pärast mingit protseduuri ja tahetakse hinnata toimunud muutust (edaspidi Δ). Siis on kaks valimit sõltuvad (enne $x_1, x_2, ..., x_n$ ja pärast $y_1, y_2, ..., y_n$).

Sel juhul moodustatakse Δ usalduspiiride I_Δ leidmiseks uus tunnus:

$$z_i = y_i - x_i, \quad i = 1, 2, \ldots, n,$$

mis on juhusliku suuruse Z_i realisatsiooniks. Kasutatavaks mudeliks on järgmine eeldus:

$$Z_i = Y_i - X_i \sim N(\Delta,\ \sigma_z^2).$$

Parameetrid on siin tundmatud. Jõudsime tuttava ülesande juurde. Meil on andmed z_i normaaljaotusest ja on vaja anda vahemikhinnang normaaljaotuse keskväärtusele (juhul kui dispersioon σ_z^2) tundmatud). Teoreemist 9 saame sellele vastuse, mille sõnastame omaette teoreemina.

Teoreem 13 *Ülalkirjeldatud sõltuvate valimite korral on vahemikhinnang keskväärtuste erinevusele Δ järgmine:*

$$I_\Delta = \bar{z} \pm t_{\alpha/2}(f)\frac{s_z}{\sqrt{n}},$$

kus $f = n - 1$, $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$ ja

$$s_z = \Big(\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2\Big)^{1/2}$$

on hinnanguks juhusliku suuruse Z_i standardhälbele σ_z.

4.6 Rakendus binoom- ja Poissoni jaotuse parameetritele

Varasemast teame, et juhul kui hinnang on avaldatav valimikeskmise ($\bar{x}$) abil ja valimimaht on suur, siis hinnangu jaotuse lähendamiseks sobib normaaljaotus. See fakt põhineb tsentraalsel piirteoreemil (TPT):

Teoreem 14 (Tsentraalne piirteoreem) *Olgu $X_1, X_2, \ldots$ sõltumatud, sama jaotusega juhuslikud suurused, kus $EX_i = \mu$, $DX_i = \sigma^2$ (lõplikud). Olgu $Y_n = X_1 + X_2 + \ldots + X_n$, siis $n \to \infty$ kehtib, et*

$$\frac{Y_n - EY_n}{\sqrt{DY_n}} = \frac{Y_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{D} N(0,1)$$

Teiste sõnadega, suure valimi korral valimikeskmise jaotus on lähendatav normaaljaotusega,

$$Y_n \sim AsN(n\mu, \sigma\sqrt{n}) \Rightarrow \bar{X} = \frac{Y_n}{n} \sim AsN(\mu, \frac{\sigma}{\sqrt{n}}),$$

kus sümbol *As* jaotuse nime ees tähistab ligikaudset jaotust (ingl. keelsest sõnast *asymptotical*). Seega, $AsN()$ vastab ligikaudsele (asümptootilisele) normaaljaotusele. Mida suurem on valimimaht n, seda lähedasem on jaotus nomraaljaotusele.

Tsentraalse piirteoreemi abil on võimalik tuletada vahemikhinnangud nii binoomjaotuse parameetrile p (välendab osakaalu, või huvipakkuva sündmuse tõenäosust) kui ka Poissoni jaotuse parameetrile λ, mis on samal ajal ka jaotuse keskväärtus.

4.6.1 Vahemikhinnang binoomjaotuse parameetrile p ühe valimi korral

Olgu $Y \sim Bin(n,p)$, kus p on tundmatu parameeter ja n on teada ning on suur. Leiame I_p usaldusnivool $1-\alpha$ ligikaudselt tsentraalse piirteoreemi abil.

- Teame, et binoomjaotust võime vaadelda Bernoulli jaotuste summana, st

$$Y = \sum_{i=1}^{n} X_i \sim Bin(n,p), \text{ kus } X_i \sim Be(p), \text{ sõltumatud.}$$

- Teame samuti, et nihketa ja mõjus hinnang parameetrile p (Bernoulli jaotuse keskväärtus!) on

$$\hat{p} = \frac{y}{n}, \text{ ehk teisiti } \hat{p} = \frac{\sum x_i}{n} = \bar{x}.$$

- Rakendades TPT, saame

$$\hat{\boldsymbol{p}} = \frac{Y}{n} = \bar{X} \sim AsN(E\hat{\boldsymbol{p}}, \sqrt{D\hat{\boldsymbol{p}}}) = AsN\left(p, \sqrt{\frac{p(1-p)}{n}}\right).$$

  Viimane väide kehtib sest:

$$E(\hat{\boldsymbol{p}}) = E\left(\frac{Y}{n}\right) = \frac{1}{n}EY = \frac{1}{n} \cdot np = p,$$

  ja

$$D(\hat{\boldsymbol{p}}) = D\left(\frac{Y}{n}\right) = \frac{1}{n^2}DY = \frac{1}{n^2} \cdot np(1-p) = \frac{p(1-p)}{n}.$$

- Normeerime hinnangufunktsiooni,

$$\frac{\hat{\boldsymbol{p}} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim AsN(0,1).$$

- Normaaljaotuse täiendkvantiilide abil saame välja kirjutada tõenäosusavaldise (märk on $\approx$, kuna kasutame lähendamist normaaljaotusega),

$$1-\alpha \approx P\left(-\lambda_{\alpha/2} < \frac{\hat{\boldsymbol{p}} - p}{\sqrt{\frac{p(1-p)}{n}}} < \lambda_{\alpha/2}\right).$$

- Saab näidata, et teisendades tõenäosusavaldist nii, et see moodustaks vahemiku parameetri p jaoks, jõuab järgmise tulemuseni:

$$I_p \approx \tilde{p} \pm \lambda_{\alpha/2} \frac{\sqrt{\hat{p}(1-\hat{p})/n + \lambda^2_{\alpha/2}/(4n^2)}}{1 + \lambda^2_{\alpha/2}/n}, \tag{4.27}$$

  kus

$$\tilde{p} = \frac{\hat{p} + \lambda^2_{\alpha/2}/(2n)}{1 + \lambda^2_{\alpha/2}/n}. \tag{4.28}$$

- Avaldist (4.27) nimetatakse vahemikhinnanguks parameetrile p **Wilsoni skoori** meetodil (1927).

Teine levinud meetod vahemikhinnangu leidmiseks on nn **Waldi meetod**.

- Lähtume taas normeeritud kujust, kuid hinnangufunktsiooni $\hat{\boldsymbol{p}}$ dispersiooni $D(\hat{\boldsymbol{p}}) = \frac{p(1-p)}{n}$ asemel kasutame mõjusat dispersiooni hinnangut $\hat{D}(\hat{\boldsymbol{p}}) = \frac{\hat{\boldsymbol{p}}(1-\hat{\boldsymbol{p}})}{n}$:

$$\hat{\boldsymbol{p}} \sim AsN\left(p, \sqrt{\frac{\hat{\boldsymbol{p}}(1-\hat{\boldsymbol{p}})}{n}}\right).$$

- Vastav tõenäosusavaldis on järelikult

$$1-\alpha \approx P\left(-\lambda_{\alpha/2} < \frac{\hat{\boldsymbol{p}}-p}{\sqrt{\frac{\hat{\boldsymbol{p}}(1-\hat{\boldsymbol{p}})}{n}}} < \lambda_{\alpha/2}\right).$$

- Teisendame selle nii, et saaks alumise ja ülemise piiri p jaoks:

$$1-\alpha \approx P\left(\hat{\boldsymbol{p}} - \lambda_{\alpha/2}\sqrt{\frac{\hat{\boldsymbol{p}}(1-\hat{\boldsymbol{p}})}{n}} < p < \hat{\boldsymbol{p}} + \lambda_{\alpha/2}\sqrt{\frac{\hat{\boldsymbol{p}}(1-\hat{\boldsymbol{p}})}{n}}\right).$$

- Asendades hinnangufunktsiooni $\hat{\boldsymbol{p}}$ valimi põhjal leitud punkthinnanguga $\hat{p}$ saame usaldusintervalli Waldi meetodil:

$$I_p \approx \left(\hat{p} \pm \lambda_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right). \tag{4.29}$$

Märkus. Vaatamata sellele, et nii (4.27) kui ka (4.29) töötavad seda paremini, mida suurem on n (sest on tuletatud kasutades TPT), on näidatud, et (4.27) on siiski täpsem. Valem (4.29) töötab hästi siis kui kehtivad tingimused: $n\hat{p} \geq 10$ ja $n(1-\hat{p}) \geq 10$. On võimalik konstrueerida nn täpset vahemikhinnangut parameetrile p, kuid see nõuab rohkem teadmisi (näiteks Beta jaotuse kohta). Huvilised võivad uurida siin: http://arxiv.org/abs/1303.1288

Näide 39 *Ühes linnaosas otsustati rajada jalgrattatee. Üheks huvipakkuvaks näitajaks on jalgratta omanike osakaal. Osutus, et juhuslikult valitud 48 inimese seast 16 omab jalgratast. Konstrueerime vahemikhinnang rattaomanike osakaalule p usaldusnivool* 95%.

1. *Wilsoni skoori meetod. Paneme tähele, et punkthinnang p-le on $\hat{p} = \frac{16}{48} = 0,333$. Vahemikhinnang:*

$$I_p \approx \frac{0,333 + 1,96^2/(2\cdot 48)}{1+1,96^2/48} \pm 1,96 \cdot \frac{\sqrt{0,333\cdot 0,667/48 + 1,96^2/(4\cdot 48^2)}}{1+1,96^2/48}$$

$$= 0,346 \pm 0,129 = (0,217;\ 0,475).$$

2. *Waldi meetod.*

$$I_p \approx 0,333 \pm 1,96\sqrt{0,333\cdot 0,667/48} = 0,333 \pm 0,133 = (0,200;\ 0,466).$$

4.6.2 Valimimahu määramine binoomjaotuse osakaalu usalduspiiride järgi

Tähistame usaldusvahemiku laiust c. Olgu see ette antud (fikseeritud). Usaldusnivoo $1-\alpha$ ja laiuse abil saab leida ligikaudse valimimahu (*ise!*).

1. Wilsoni skoori meetod:

$$n = \frac{2\lambda^2_{\alpha/2}\hat{p}\hat{q} - \lambda^2_{\alpha/2}c^2 \pm \sqrt{4\lambda^4_{\alpha/2}\hat{p}\hat{q}(\hat{p}\hat{q} - c^2) + c^2\lambda^4_{\alpha/2}}}{c^2},$$

   kus $\hat{q} = 1 - \hat{p}$.

2. Waldi meetodi kohaselt,

$$n = \frac{4\lambda^2_{\alpha/2}\hat{p}\hat{q}}{c^2}.$$

Paneme tähele, et mõlemad nõuavad $\hat{p}$ teadmist. Juhul kui puudub info hinnangu kohta, siis asendatakse $\hat{p}\hat{q}$ funktsiooni maksimumiga (1/4). See viib nn konsermatiivse valimimahuni.

Näide 40 *Soovime jalgrattaomanike näites vahemikhinnangut laiusega $c = 0,1$. Siis*

- *Wilsoni meetodi kohaselt peaks valimisse võtma*

$$n = \frac{2 \cdot 1,96^2 \cdot 0,25 - 1,96^2 \cdot 0,01 + \sqrt{4 \cdot 1,96^4 \cdot 0,25(0,25 - 0,01) + 0,01 \cdot 1,96^4}}{0,01} = 380,3;$$

- *Waldi meetodiga*

$$n = \frac{4 \cdot 1,96^2}{4 \cdot 0,01} = 385.$$

  inimest.

Näeme, et antud juhul ei erine meetodid kuigi palju.

4.6.3 Vahemikhinnangud suvalise jaotuse parameetritele

Järgnev peatükk põhineb õpikul Traat (2006), lk. 95-98.
Olgu $\hat{\theta}$ punkthinnang (asümptootiliselt nihketa) parameetrile θ ja olgu $\hat{\boldsymbol{\theta}}$ vastav tõenäosusfunktsioon. On tõestatud, et protsessis $n \to \infty$ kehtivad enamasti järgmised jaotuse järgi koondumised:

$$\frac{\hat{\boldsymbol{\theta}} - \theta}{\sqrt{D\hat{\boldsymbol{\theta}}}} \to N(0,1),$$

$$\frac{\hat{\boldsymbol{\theta}} - \theta}{\sqrt{\hat{D}\hat{\boldsymbol{\theta}}}} \to N(0,1),$$

kus $\hat{D}\hat{\boldsymbol{\theta}}$ on dispersiooni $D\hat{\boldsymbol{\theta}}$ mõjus hinnang. See omadus laiendab ääretult statistika võimalusi, öeldes, et teatud valimimahust n alates on enamus statistikuid ligikaudu normaaljaotusega, $\hat{\boldsymbol{\theta}} \sim AsN(\theta, \sqrt{D\hat{\boldsymbol{\theta}}})$, või $\hat{\boldsymbol{\theta}} \sim AsN(\theta, \sqrt{\hat{D}\hat{\boldsymbol{\theta}}})$.

Teoreem 15 *[normaaljaotusega lähendamine ühe valimi korral] Kui statistik* $\hat{\boldsymbol{\theta}} \sim AsN(\theta, \sqrt{D\hat{\boldsymbol{\theta}}})$*, siis* θ *usaldusvahemik ligikaudsel usaldusnivool* $1-\alpha$ *on:*

$I_\theta = \hat{\theta} \pm \lambda_{\alpha/2}\sqrt{D\hat{\boldsymbol{\theta}}}$*, kui* $D\hat{\boldsymbol{\theta}}$ *ei sõltu tundmatust parameetrist* θ*,*

$I_\theta = \hat{\theta} \pm \lambda_{\alpha/2}\sqrt{\hat{D}\hat{\boldsymbol{\theta}}}$*, kui* $D\hat{\boldsymbol{\theta}}$ *sõltub tundmatust parameetrist* θ*.*

Vahemikhinnang osakaalule p Waldi meetodil ongi teoreemi teise väite rakendus.

Märkused teoreemile normaaljaotusega lähendamisest

1. Vahemikhinnangud, mis baseeruvad teoreemile normaaljaotusega lähendamisest, kehtivad ainult suure n korral. Samas pole suure n mõiste täpselt fikseeritud.
2. Usaldusnivoo on sel juhul ligikaudu $1-\alpha$. Ligikaudsus on seda suurem, mida rohkem erineb $\hat{\theta}$ normaaljaotusest antud n korral.
3. Tegelik katmistõenäosus võib olla nii väiksem kui ka suurem $(1-\alpha)$-st.
4. Statistikapakettides võetakse sageli $\lambda_{\alpha/2}$ asemel $t_{\alpha/2}(f)$. Usaldusvahemik on siis pigem laiem ja ollakse nn "kindlamal poolel"(usaldusvahemiku alahindamine on ebasoovitavam kui ülehindamine).

4.6.4 Rakendus Poissoni jaotuse parameetrile

Tuginedes teoreemile normaaljaotusega lähendamisest saab kirja panna ka vahemikhinnangu Poissoni jaotuse parameetrile.

Järeldus 5 *Suure valimi korral on üldkogumijaotuse* $Po(\mu)$ *parameetri vahemikhinnanguks*

$$I_\mu = \hat{\mu} \pm \lambda_{\alpha/2}\sqrt{\hat{D}\hat{\boldsymbol{\mu}}} = \bar{x} \pm \lambda_{\alpha/2}\sqrt{\frac{\bar{x}}{n}}.$$

Usaldusnivoo on I_μ korral ligikaudu $1-\alpha$.

Tõestus. Olgu valim $x_1, .., x_n$ realiseerunud teoreetilisest valimist $X_1, ..., X_n$, kus $X_i \sim Po(\mu)$ ning sõltumatud. Olgu valimimaht n suur. Kuna μ on jaotuse keskväärtus, siis selle hindamiseks sobib valimikeskmine, $\hat{\mu} = \bar{x}$. See on nihketa ja mõjus hinnang parameetrile μ.

Vastavalt Teoreemile 15 peame leidma $D(\hat{\boldsymbol{\mu}})$ et teada saada, kas saadud avaldis sõltub parameetrist μ või mitte. Arvestades, et Poissoni jaotuse korral $EX_i = DX_i = \mu$ ning X_i on sõltumatud, saame

$$D(\hat{\boldsymbol{\mu}}) = D(\bar{X}) = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} DX_i = \frac{1}{n^2}\sum_{i=1}^{n}\mu = \frac{\mu}{n}.$$

Näeme, et $D(\hat{\boldsymbol{\mu}})$ avaldis sõltub tundmatust parameetrist μ, seega peame rakendama Teoreemi 15 teist väidet, mille kohaselt tuleks $D(\hat{\boldsymbol{\mu}})$ hinnata valimi põhjal:

$$\hat{D}(\hat{\boldsymbol{\mu}}) = \frac{\hat{\mu}}{n} = \frac{\bar{x}}{n}.$$

Asendades saadud avaldist Teoreemi 15 teises väites, saamegi järelduse tõestatud. □

4.6.5 Vahemikhinnang kahe osakaalu vahele

Järgnev peatükk põhineb õpikul Traat (2006), lk. 104-105.

On tõestatud, et analoogiliselt normaaljaotusega lähendamisega ühe valimi korral, kehtivad koondumised ka kahe parameetri vahe korral. Vastavate vahemikhinnangute usaldusnivoo on ligikaudu $1-\alpha$, seda täpsemini, mida suurem on valimimaht. See põhineb asjaolul, et protsessis $n \to \infty$ toimuvad jaotuse järgi koondumised:

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{D(\bar{X}-\bar{Y})}} \to N(0,1),$$

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\hat{D}(\bar{X}-\bar{Y})}} \to N(0,1).$$

Olgu meil kaks valimit Bernoulli jaotustest $Be(p_1)$ ja $Be(p_2)$, kus parameetrid on p_1 ja p_2, mida võiks näiteks interpreteerida kui nähtuse A pooldajate osakaalu mingil ajahetkel Tallinnas ja Tartus.
Meid huvitab osakaalude vahe usaldusvahemik $I_{p_1-p_2}$.
Teeme nendest jaotustest vastavalt n_1 ja n_2 vaatlust.
Tähistame: y_1 – nähtuse pooldajate arv esimese valimi korral ja y_2 – pooldajate arv teise valimi korral. Siis $Y_1 \sim B(n_1, p_1)$ ja $Y_2 \sim B(n_2, p_2)$. Järgnev teoreem kehtib suurte valimite korral.

Teoreem 16 (Usaldusintervall $I_{p_1-p_2}$) *Eespool kirjeldatud eelduste korral on vahemikhinnang osakaalude vahele ligikaudsel usaldusnivool $1-\alpha$ järgmine:*

$$I_{p_1-p_2} = \left(\hat{p}_1 - \hat{p}_2 \pm \lambda_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right).$$

Tõestus. Toimime üldise algoritmi järgi, mis on kirjeldatud punktis 4.1.
1) Võtame vahe $p_1 - p_2$ punkthinnanguks $\hat{p}_1 - \hat{p}_2$, kus

$$\hat{p}_1 = \frac{y_1}{n_1}, \quad \hat{p}_2 = \frac{y_2}{n_2}.$$

2) Vaatame vastavat hinnangufunktsiooni ja selle jaotust:

$$\hat{\boldsymbol{p_1}} - \hat{\boldsymbol{p_2}} = \frac{Y_1}{n_1} - \frac{Y_2}{n_2}.$$

Leiame:

$$E(\hat{\boldsymbol{p_1}} - \hat{\boldsymbol{p_2}}) = p_1 - p_2,$$

$$D(\hat{\boldsymbol{p_1}} - \hat{\boldsymbol{p_2}}) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$$

Kui n_1 ja n_2 on suured, siis vaadeldav statistik on ligikaudu normaaljaotusega:

$$\hat{\boldsymbol{p_1}} - \hat{\boldsymbol{p_2}} \sim AsN\left(p_1 - p_2,\ \sqrt{\frac{\hat{\boldsymbol{p_1}}(1-\hat{\boldsymbol{p_1}})}{n_1} + \frac{\hat{\boldsymbol{p_2}}(1-\hat{\boldsymbol{p_2}})}{n_2}} \right),$$

mille normeerides saame

$$\frac{\hat{\boldsymbol{p_1}} - \hat{\boldsymbol{p_2}} - (p_1 - p_2)}{\sqrt{\frac{\hat{\boldsymbol{p_1}}(1-\hat{\boldsymbol{p_1}})}{n_1} + \frac{\hat{\boldsymbol{p_2}}(1-\hat{\boldsymbol{p_2}})}{n_2}}} \sim AsN(0,1).$$

3) Moodustame tõenäosusavaldise:

$$1 - \alpha \approx P\left(-\lambda_{\alpha/2} < \frac{\hat{\boldsymbol{p_1}} - \hat{\boldsymbol{p_2}} - (p_1 - p_2)}{\sqrt{\frac{\hat{\boldsymbol{p_1}}(1-\hat{\boldsymbol{p_1}})}{n_1} + \frac{\hat{\boldsymbol{p_2}}(1-\hat{\boldsymbol{p_2}})}{n_2}}} < \lambda_{\alpha/2} \right).$$

4) Teisendades viimast avaldist $(p_1 - p_2)$ suhtes saamegi vahemikhinnangu kahe osakaalu vahele:

$$I_{p_1-p_2} = \hat{p}_1 - \hat{p}_2 \pm \lambda_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

usaldusnivooga ligikaudu $1 - \alpha$. □

Näide 41 *Üks meditsiiniline ajakiri avaldas andmeid vähihaigete patsientide arvu kohta, keda raviti kahe erineva metoodika abil. Esimise korral, raviti patsiente ainult kiiritusega. Selgus välja, et 154 patsiendist 76 jäid 15 aasta jooksul ellu. Teised läbisid lisaks kemoteraapiat. Sellistest 164 patsiendist jäi ellu 98 vaadeldava perioodi jooksul.*

Olgu p_1 ellujäänute patsientide arv, keda raviti meetodika I abil ja p_2 vastavalt metoodika II abil. Punkthinnangud on:

$$\hat{p}_1 = \frac{76}{154} = 0,494 \text{ ja } \hat{p}_2 = \frac{98}{164} = 0,598.$$

Vahemikhinnang osakaalude vahele usaldusnivool 99% *on siis*

$$I_{p_1-p_2} = 0,494 - 0,598 \pm 2,58\sqrt{\frac{0,494 \cdot 0,506}{154} + \frac{0,598 \cdot 0,402}{164}} =$$

$$= -0,104 \pm 0,143 = (-0,247;\ 0,039).$$

$$I_{p_1-p_2} = (-0,247;\ 0,039)$$

- *Vahemikhinnang on küllaltki lai. Selle põhjuseks:*
  - *kõrge usaldusnivoo;*
  - *mitte piisavalt suured valimimahud.*
- *Vahemik sisaldab nulli, mis viitab sellele, et antud valimite korral erinevust meetoodikates ei selgu.*

5. Hüpoteeside kontroll

Antud peatükis uurime veel ühte võimalust jaotuse tundmatu parameetri hindamiseks juhusliku valimi põhjal. Kuid selle asemel, et leida valimi väärtuste põhjal sellele hinnang või väärtuste vahemik, pakkume kõigepealt sellele mõne hüpoteetilise väärtuse. Seejärel valimi põhjal saadud informatsiooniga püüame hüpoteesi kinnitada või kummutada. Antud peatükis tuletame meelde juba tuttavaid algoritme (kursusest "Tõenäosusteooria ja statistika I") hüpoteeside kontrollimiseks, õpime uusi võimalusi ning kirjeldame antud algoritmide omadusi teoreetiliselt.

Siinkohal tegeleme vaid parameetriliste hüpoteesidega, mis kujutavad endast väidet üldkogumi jaotus(t)e parameetrite kohta.

5.1 Parameetrilistest hüpoteesidest üldiselt

Huvitagu meid üldkogumis tunnus (ehk juhuslik suurus) $X \sim F(\theta)$ (F tähitab suvalist jaotust), kus θ on tundmatu. Oletame, et uurijal on θ väärtuse kohta teatud oletus, mida ta soovib kontrollida valimi $x_1, x_2, ..., x_n$ abil. Parameetriline statistiline hüpotees sõnastatakse väidete paarina:

nullhüpotees	$H_0 : \theta = \theta_0,$
alternatiivhüpotees	$H_1 : \theta \neq \theta_0.$

Võimalikud variandid on ka $H_0 : \ \theta \leq \theta_0$ või $H_0 : \ \theta \geq \theta_0$. Hüpotees H_1 on H_0 täiend.

Näide 42 *Tehases toodetakse ühte tüüpi detaile, mille eluiga (tundides) on normaaljaotusega juhuslik suurus, $N(100,\ 3)$. Et seda eluiga pikendada, töötati välja uus valmistamise tehnoloogia. Loodetakse, et uut tüüpi detailide eluiga on samuti normaaljaotusega juhuslik suurus, $N(\mu,\ 3)$, kus $\mu > 100$ (esialgu oletame, et standardhälve ei muutu, ehk uute detailide eluea varieeruvus jääb samaks). Seega, soovitakse tehases kontrollida järgmist hüpoteeside paari:*

$$H_0 : \ \mu = 100 \quad \textit{versus} \quad H_1 : \mu > 100.$$

Vaikimisi on siin tehtud eeldus, et uut tüüpi detailide keskmine eluiga ei saa olla kehvem kui vana tüüpi oma.

Hüpoteeside sõnastamisel võetakse arvesse järgmisi punkte:

- H_0 kirjeldab tasakaalu olukorda või seni kehtinud olukorda (uut tüüpi detilide keskmine eluiga jääb samaks).

- H_1 kirjeldab olukorda, mis uurijat huvitab, seda väidet soovib uurija tõestada (keskmine eluiga pikeneb).
- Väidet H_0 ei saa tõestada, öeldakse, et selle juurde jäädakse, H_0 kummutamisel aga loetakse H_1 tõestatuks.
- Nii kummutamisel kui ka jäämisel on olemas risk teha vea.
- Hüpoteeside kontrollieeskiri konstrueeritakse nii, et vea tegemise tõenäosus on väike (õigemini kahte tüüpi vea tõenäosused).

Mõnikord liigitatakse hüpoteesid järgmiselt

- lihthüpotees (ka kahpoolne hüpotees) – väide parameetri ühe väärtuse kohta ($H_0 : \theta = \theta_0$),
- liithüpotees (ka ühepoolne hüpotees) – väide parameetri väärtuspiirkonna kohta ($H_0 : \theta \leq \theta_0$).

Näide 43 *Jätkame eelmise näitega, kus soovisime kontrollida hüpoteeside paari*

$$H_0 : \ \mu = 100 \quad \textit{versus} \quad H_1 : \mu > 100.$$

Hüpoteeside paari kontrollimiseks võetakse juhuslik partii detailidest, mis on valmistatud uue tehnoloogia järgi, ning saadakse eluea väärtused $x_1, ..., x_n$. Nende põhjal leitakse valimikeskmine $\bar{x}$, sest varasemast teame juba, et valimikeskmine on "hea"hinnang normaaljaotuse keskväärtusele μ. Otsustatakse, et hüpoteeside paari kontrollimiseks kasutatakse järgmist eeskirja:

$$\textit{kui } \bar{x} - 100 > C', \ \textit{siis } H_1 \textit{ on tõestatud};$$

või alternatiivselt,

$$\textit{kui } \bar{x} > C, \ \textit{siis } H_1 \textit{ on tõestatud.}$$

Siinkohal nii C' kui C on mingid konstandid, mida tuleks kuidagi paika panna. Paneme kohe tähele, et otsustamise kriteerium on sama tüüpi mis H_1. Kui huvituksime kahepoolsest hüpoteeside paarist $H_0 : \ \mu = 100$ versus $H_1 : \mu \neq 100$, siis ka otsustamise kriteerium oleks: kui $|\bar{x} - 100| > C'$, siis H_1 on tõestatud.

Kriteeriumi sõnastamisel näeme, et erinevate C väärtuste korral saame erinevaid kriteeriume. Samuti juhul, kui $\bar{x}$ asemel võtame muu hinnangu (näiteks mediaani), saame moodustada erinevaid kontrollimise eeskirju. Siit võivad tekkida ka küsimused: "Kui hea on püstitatud kriteerium?", "Milliste omadustega on ta?", "Kuidas on võimalik kahte erinevat kriteeriumi omavahel võrrelda?"

Hüpoteeside kontrollimisel on oluline aru saada, et ei eksisteeri ühtegi eeskirja, mis garanteeriks täpset vastust H_1 või H_0 õigsuse kohta. Me seda ei ootagi (nii nagu ka punkthinnangust kunagi ei oota, et see langeb kokku tegeliku parameetriga). Meie poolt pakutud kontrollimise eeskiri (=test) ei pruugi anda alati õiget otsust, kuid kõige parem test on selline, mis annab seda õiget otsust "kõige rohkem".

Hüpoteeside kontrollimisel on võimalik eksida kahel suunal: kriteeriumi järgi kummutame H_0, aga tegelikult H_0 on õige (I tüüpi viga), või kriteeriumi järgi jääme H_0 juurde, kuid tegelikult H_0 on vale (II tüüpi viga).

Sageli I ja II tüüpi vigade tõenäosused on tähistatud vastavalt α ja β abil (vt. järgmist tabelt).

Otsus	Tegelik olek	
	õige on H_0	õige on H_1
Jääme H_0 juurde	+	II liiki viga, tõen. β
Kummutame H_0	I liiki viga, tõen. α	+

Definitsioon 31 *Suurimat lubatavat esimest liiki vea tõenäosust nimetatakse testi* ***olulisuse nivooks*** *ja tähistatakse* α*:*

$$\alpha = \begin{cases} P(\textit{kummutada } H_0 | \ H_0 \textit{ õige}), \textit{ kui } H_0 \textit{ on lihthüpotees,} \\ sup_{\ \theta \in H_0} \ P(\textit{kummutada } H_0 | \ H_0 \textit{ õige}), \textit{ kui } H_0 \textit{ on liithüpotees.} \end{cases}$$

Standardsed olulisuse nivood on $\alpha = 0,05$, $\alpha = 0,01$, või $\alpha = 0,001$.

5.2 Testi võimsusfunktsioon

Üsna loomulik on arvata, et hea test on see, mille I ja II liiki viga on väike ja mis suure tõenäosusega kummutab H_0, kui H_1 on õige. Testide omavaheliseks võrdlemiseks ja testi omaduste uurimiseks toome sisse võimsusfunktsiooni mõistet.

Definitsioon 32 *Testi* ***võimsusfunktsioon*** $h(\theta)$ *on nullhüpoteesi kummutamise tõenäosus vaadatuna* θ *funktsioonina:*

$$h(\theta) = P(\textit{kummutada } H_0 \mid \theta).$$

Märkus: mõnikord võimsusfunktsiooni asemel kasutatakse selle vastandtõenäosust, $P(\textit{Jääda } H_0 \textit{ juurde} | \theta) = 1 - h(\theta)$.

Võimsusfunktsiooni omadused:

- hea testi korral on $h(\theta)$ väike, kui $\theta \in H_0$ ning $h(\theta)$ suur, kui $\theta \in H_1$;
- $0 \leq h(\theta) \leq 1$,
- testi olulisuse nivoo avaldub võimsusfunktsioonist

$$sup_{\ \theta \in H_0} \ h(\theta) = \alpha,$$

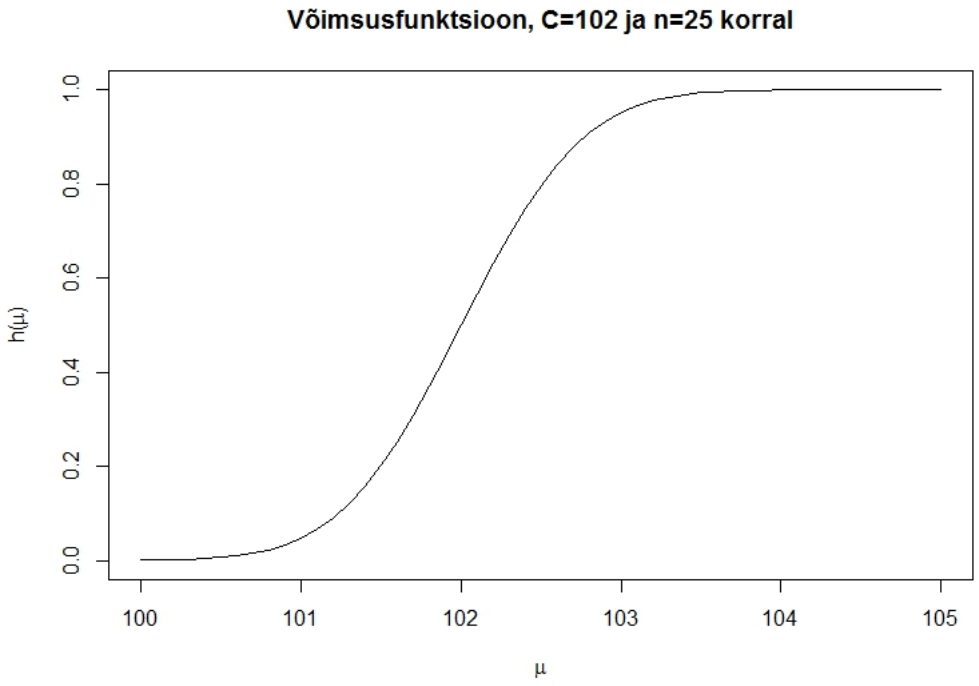
ehk kui H_0 on lihthüpotees, siis

$$h(\theta_0) = \alpha.$$

Näide 44 *Leiame eelmises näiteks võimsusfunktsioon* $h(\mu)$*. Arvestades, et uut tüüpi detailide valim on normaalajotusest,* $X_i \sim N(\mu, 3)$, $n = 1, 2, ..., n$*, siis valimikeskmise hinnangufunktsioon* $\bar{X} \sim N(\mu, \ \frac{3}{\sqrt{n}})$*. Seega,*

$$\begin{aligned} h(\mu) &= P(\textit{kummutada } H_0 \mid \mu) = P(\bar{X} > C \mid \mu) = 1 - P(\bar{X} \leq C \mid \mu) \\ &= 1 - P\left(\frac{\bar{X} - \mu}{3/\sqrt{n}} \leq \frac{C - \mu}{3/\sqrt{n}}\right) = 1 - \Phi\left(\frac{(C - \mu)\sqrt{n}}{3}\right). \end{aligned}$$

Fikseeritud C *ja* n *korral näeb* $h(\mu)$ *välja järgmiselt:*



Graafikult näeme, et kui tegelik keskmine eluiga $\mu = 101$, siis kummutab pakutud test (kui $\bar{X} > 102$, siis kummutada H_0) null-hüpoteesi üsna väikese tõenäosusega (ligikaudu 0,05):

$$h(101) = 1 - \Phi\left(\frac{(102-101)\sqrt{25}}{3}\right) \approx 1 - \Phi(1,667) = 0,0478.$$

See näitab, et kui tegelik keskmine oleks 101 (kuulub hüpoteesi H_1 hulka), siis sõltumata realiseerunud valimist teeks meie test õige otsuse vaid 4,78% jutudest. See on mõistagi väga väike tõenäosus, mis vihjab sellele, et $\mu = 101$ korral test ei töötaks kuigi hästi.

Võtame veidi suurema μ väärtuse. Oletame, et tegelik keskmine eluiga on $\mu = 103$, siis:

$$h(103) = 1 - \Phi\left(\frac{(102-103)\sqrt{25}}{3}\right) \approx 1 - \Phi(-1,667) = \Phi(1,667) = 0,952,$$

mis on väga suur kummutamise tõenäosus. Selle μ väärtuse korral töötab test väga hästi.

Näites on käsitletud ühte konstandi C väärtust, mis oli määratud tehase direktori sisetunde järgi. Kuidas aga leida selline C, mis viiks kõige parema (võimsama) testini?

Fikseeritud valimimahu n korral fikseeritakse tavaliselt I liiki vea tõenäosust (ehk olulisuse nivood) α ning saadud eeskirjast avaldatakse C.

Näide 45 *Leiame eelmises näites optimaalse konstandi C väärtuse kui $n = 25$ ja $\alpha = 0,05$. Võimsusfunktsioon ja olulisuse nivoo on omavahel seotud järgmise eeskirjaga:*

$$sup\ _{\theta \in H_0}\ h(\theta) = \alpha,$$

kusjuures supremum on saavutatud punktis $\mu = 100$. Seega, $0,05 = h(100)$, millest

$$0,05 = 1 - \Phi\left(\frac{(C-100)\cdot 5}{3}\right).$$

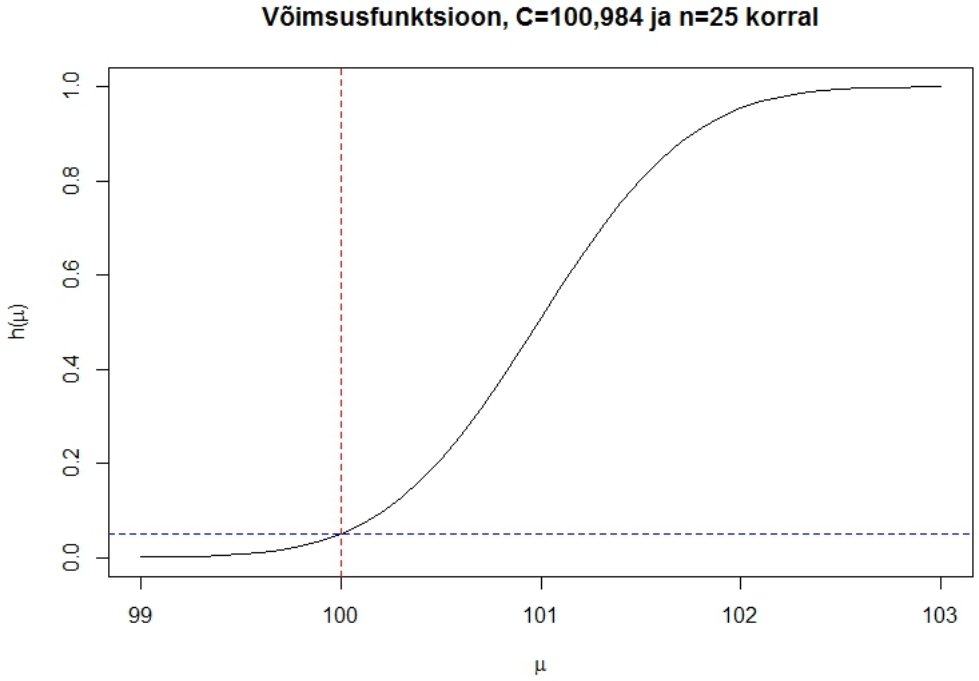
Alternatiivselt saab viimast avaldist esitada kujul

$$\Phi\left(\frac{(C-100)\cdot 5}{3}\right) = 0,95,$$

millest

$$\frac{(C-100)\cdot 5}{3} = 1,64.$$

Lahendades saadud võrratust saame $C = 100,984$. *Järgmisel joonisel on toodud võimsusfunktsioon* $C = 100,984$ *ja* $n = 25$ *korral. Graafikule on lisatud ka vertikaalne punktiirjoon, mis vastab punktile* $\mu = 100$ *ning horisontaalne punktiirjoon, mille väärtuseks y-teljel on 0,05 ehk olulisuse nivoo väärtus.*



Kui soovitakse leida optimaalsed C ja n väärtused, siis kahe muutuja leidmiseks peame kasutama kahte punkti võimsusfunktsiooni graafikul. See võimaldab hoida kontrolli all ka II liiki vea tõenäosust.

Näide 46 *Kasutame eelmist testi: 'kui* $\bar{x} > C$, *siis* H_1 *on tõestatud', kus* $X_i \sim N(\mu, 3)$, $i = 1, ..., n$. *Leiame* C *ja* n *nii, et* $h(100) = 0,05$ *ja* $h(102) = 0,99$. *Viimast tingimust võime sõnastada nii: 'kui tegelik* $\mu = 102$, *siis soovime, et test kummutaks* H_0 *tõenäosusega 0,99', ehk alternatiivselt: 'tegeliku* $\mu = 102$ *korral jätab test* H_0 *juurde tõenäosusega 0,01', ehk alternatiivselt: 'tegeliku* $\mu = 102$ *korral on II tüüpi vea tõenäosus vaid 0,01'.*

Eelnevalt oleme juba leidnud, et

$$h(\mu) = 1 - \Phi\left(\frac{(C-\mu)\sqrt{n}}{3}\right).$$

Meie lisatingimuste korral saame:

$$h(100) = 0,05 \quad ja \quad h(102) = 0,99.$$

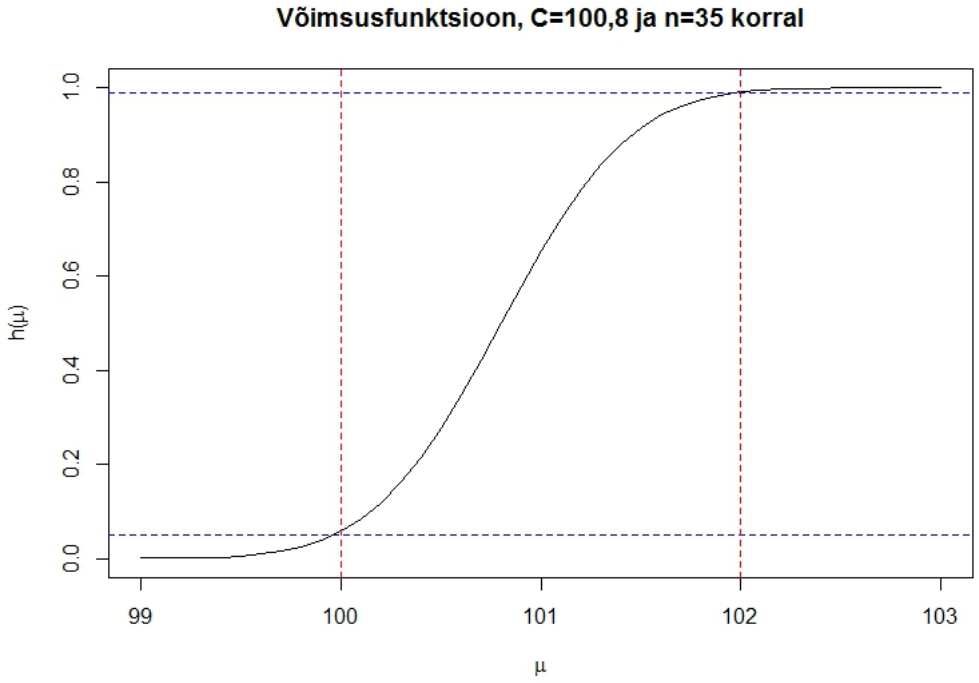
See tähendab järgmiste võrrandite lahendamist:

$$0,05 = 1 - \Phi\left(\frac{(C-100)\sqrt{n}}{3}\right) \quad ja \quad 0,99 = 1 - \Phi\left(\frac{(C-102)\sqrt{n}}{3}\right).$$

Normaaljaotuse tabelitest saame

$$1,64 = \frac{(C-100)\sqrt{n}}{3} \quad ja \quad -2,33 = \frac{(C-102)\sqrt{n}}{3}.$$

Lahendades antud süsteemi jõuab vastuseni: $C = 100,8$ *ja* $n = 34,6 \approx 35$. *Järgmine graafik näitabki funktsiooni* $h(\mu)$ *käitumist leitud väärtuste korral.*



5.3 Hüpoteeside kontroll normaaljaotuse keskväärtuse kohta

Sõnastame üldise algoritmi hüpoteeside paari kontrollimiseks normaaljaotuse keskväärtuse μ kohta. Üldjuhul võime sõnastada järgmised hüpoteeside paarid:

$$\begin{array}{lll} H_0: \ \mu = \mu_0 & H_0: \ \mu \geq \mu_0 & H_0: \ \mu \leq \mu_0 \\ H_1: \ \mu \neq \mu_0 & H_1: \ \mu < \mu_0 & H_1: \ \mu > \mu_0 \end{array}$$

Üldine algoritm hüpoteeside kontrollimiseks keskväärtuse kohta

- Hüpoteeside paari kontrollimiseks peab olema võetud juhuslik valim $x_1, x_2, ..., x_n$.
- Moodustame punkthinnangut $\hat{\mu} = \bar{x}$.
- Vastav hinnangufunktsioon on

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

- H_0 õigsuse korral (null-sesisu korral) on

$$X_i \sim N(\mu_0, \sigma),\ \bar{X} \sim N(\mu_0, \frac{\sigma}{\sqrt{n}}).$$

- Moodustame normeeritud teststatistikud U ja V, millede jaotused on H_0 kehtivuse korral teada:

$$\begin{aligned} U &= \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1), \quad \text{kui } \sigma \text{ on teada,} \\ V &= \frac{\bar{X}-\mu_0}{s/\sqrt{n}} \sim t(f),\ f = n-1, \text{ kui } \sigma \text{ on tundmatu.} \end{aligned}$$

- Kontrollreegel sõltub hüpoteeside paarist ja informatsioonist dispersiooni σ^2 kohta:

1) σ teada	2) σ ei ole teada
$H_1 : \mu \neq \mu_0$	$H_1 : \mu \neq \mu_0$
test: $\lvert u\rvert > \lambda_{\frac{\alpha}{2}} \xrightarrow{+}$ kummuta H_0	test: $\lvert v\rvert > t_{\frac{\alpha}{2}}(f) \xrightarrow{+}$ kummuta H_0
$H_1 : \mu < \mu_0$	$H_1 : \mu < \mu_0$
test: $u < -\lambda_\alpha \xrightarrow{+}$ kummuta H_0	test: $v < -t_\alpha(f) \xrightarrow{+}$ kummuta H_0
$H_1 : \mu > \mu_0$	$H_1 : \mu > \mu_0$
test: $u > \lambda_\alpha \xrightarrow{+}$ kummuta H_0	test: $v > t_\alpha(f) \xrightarrow{+}$ kummuta H_0

Näide 47 *Olgu üldkogumi dispersioon σ^2 teada ning huvitume hüpoteeside paarist $H_0 :\ \mu = \mu_0,\ H_1 :\ \mu > \mu_0$. Näitame, et fikseeritud n ja α korral on üldise algoritmi eeskiri '$u > \lambda_\alpha \xrightarrow{+}$ kummuta H_0' samaväärne näites 45 leitud eeskirjaga C jaoks:*

$$\Phi\left(\frac{(C-\mu_0)\cdot\sqrt{n}}{\sigma}\right) = 1-\alpha \tag{5.1}$$

Näites 45 kasutasime järgmist testi: 'kui $\bar{x} > C$, siis kummutada H_0', kus C on antud eeskirjaga (5.1).

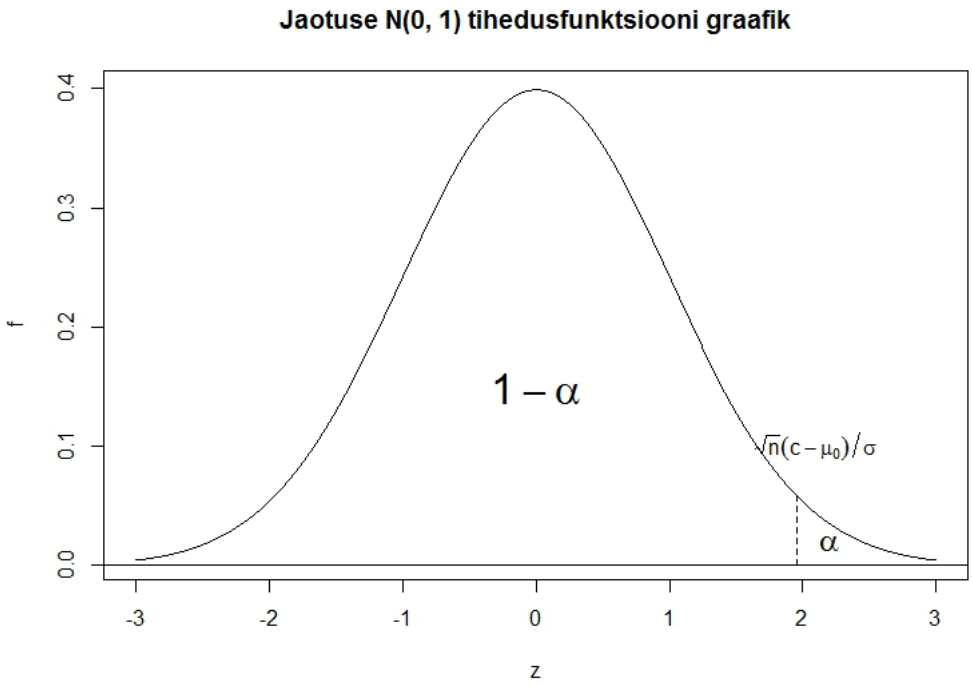
Selleks paneme kõigepealt tähele, et $\Phi\left(\frac{(C-\mu_0)\cdot\sqrt{n}}{\sigma}\right) = P\left(Z < \frac{(C-\mu_0)\cdot\sqrt{n}}{\sigma}\right)$, kus $Z \sim N(0,1)$ (vt joonist allpool). Alternatiivselt võime kirjutada:

$$\alpha = P\left(Z \geq \frac{(C-\mu_0)\cdot\sqrt{n}}{\sigma}\right),$$

mis on α-täiendkvantiili definitsioon. Seega, $\frac{(C-\mu_0)\cdot\sqrt{n}}{\sigma} = \lambda_\alpha$. Saadud avaldisest tuletame konstandi C väärtuse:

$$C = \frac{\lambda_\alpha \sigma}{\sqrt{n}} + \mu_0. \tag{5.2}$$

Näites 45 kasutatud test 'kui $\bar{x} > C$, siis kummutada H_0' on seega alternatiivne testiga 'kui $\bar{x} > \frac{\lambda_\alpha\sigma}{\sqrt{n}} + \mu_0$, siis kummutada H_0'. Viimast väidet saab kirja panna ka kujul 'kui $\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > \lambda_\alpha$, siis kummutada H_0'. See aga langeb kokku normeeritud statistikuga U ning üldise algoritmi eeskirjaga.



Näide 48 *Olgu üldkogumi dispersioon σ^2 teada ning huvitume järgmise hüpoteeside paari kontrollimisest:*

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

Kasutame testi, mis on kirjeldatud üldises algoritmis: $|u| > \lambda_{\frac{\alpha}{2}} \xrightarrow{+}$ kummuta H_0. Näitame, et selle testi I liiki vea tõenäosus on täpselt α ning võimsusfunktsioon on

$$h(\mu) = 1 - \Phi\left(\lambda_{\alpha/2} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right) + \Phi\left(-\lambda_{\alpha/2} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}\right).$$

Rakendame võimsusfunktsiooni definitsiooni:

$$\begin{aligned} h(\mu) &= P(\text{kummutada } H_0|\mu) = P(|U| > \lambda_{\alpha/2}|\mu) \\ &= 1 - P(|U| < \lambda_{\alpha/2}|\mu) = 1 - P(-\lambda_{\alpha/2} < U < \lambda_{\alpha/2}|\mu). \end{aligned}$$

Kui aga μ on üldkogumi keskväärtus, siis

$$EU = \frac{E\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\mu - \mu_0}{\sigma/\sqrt{n}},$$

$$DU = \frac{n}{\sigma^2} D\bar{X} = 1,$$

ja järelikult

$$U \sim N(\frac{\mu - \mu_0}{\sigma/\sqrt{n}}, 1).$$

Kasutades antud teadmist U kohta, saame

$$h(\mu) = 1 - \Phi(\lambda_{\alpha/2} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}) + \Phi(-\lambda_{\alpha/2} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}}),$$

kus $\Phi(\cdot)$ on standardnormaaljaotuse jaotusfunktsioon. Leiame võimsusfunktsioonist testi olulisuse nivoo, milleks asendame $\mu = \mu_0$:

$$h(\mu_0) = 1 - \Phi(\lambda_{\alpha/2}) + 1 - \Phi(\lambda_{\alpha/2}) = \alpha.$$

5.4 Kahe üldkogumi keskväärtuse võrdlemine (sõltuvad valimid)

Sõltuvate valimite korral soovitakse tavaliselt leida (ravimi) mõju olemasolu, kus ühed ja samad objektid on mõõdetud enne katset (null olukorras) ja peale katse läbiviimist (ravimi manustamist).

Punktist 4.5 teame, et valimiandmeid kirjeldab järgmine mudel:

$$(y_1 - x_1),\ \ldots,\ (y_n - x_n) \ \leftarrow N(\Delta, \sigma),$$

kus x_i on i. katseisiku mõõtetulemus enne katse läbiviimist ja y_i pärast katset, $i = 1, 2, ..., n$.
Soovitakse testida keskmist erinevust (ehk mõju Δ olemasolu):

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

Algoritm hüpoteeside paari kontrollimiseks:

- Kahe valimi baasil moodustatakse uued väärtused $z_i = y_i - x_i$, mille eeldatavaks jaotuseks on $N(\Delta, \sigma_z)$, kus σ_z on tavaliselt tundmatu parameeter.
- Seejärel moodustatakse test-statistik nii, et selle normeeritud kuju on H_0 hüpoteesi kehtivuse korral t-jaotusega:

$$V = \frac{\bar{Z} - E\bar{Z}}{\sqrt{\hat{D}(\bar{Z})}} \overset{H_0 \text{ õige}}{=} \frac{\bar{Z} - 0}{s_z/\sqrt{n}} \sim t(f),$$

  kus $f = n - 1$ ja

$$\hat{D}\bar{Z} = \frac{s_z^2}{n},$$

  milles s_z^2 on väärtuste z_i valimidispersioon.

- Testime t-jaotuse kvantiiliga:

$$|v| > t_{\alpha/2}\ (f) \xrightarrow{+} \text{kummutada } H_0.$$

- Testi olulisuse nivoo on α.

Ülejäänud olukorrad ($H_1 : \Delta > 0$ ja $H_1 : \Delta < 0$) on analoogilised punktis 5.3 toodud skeemile.

5.5 Kahe üldkogumi keskväärtuse võrdlemine (sõltumatud valimid)

Olgu nüüd uuringu all kaks üldkogumit (näiteks Eesti ja Soome tööealised inimesed). Uuritava tunnuse väärtused (näiteks töötasu) olgu normaaljaotusega juhuslikud suurused.

Tähistame: valim $x_1, x_2, \ldots, x_{n_1}$ on jaotusest $N(\mu_1, \sigma_1)$ ja sellest sõltumatu valim $y_1, y_2, \ldots, y_{n_2}$ on jaotusest $N(\mu_2, \sigma_2)$, ning hüpoteeside paar

$$H_0: \ \mu_1 = \mu_2 \Leftrightarrow \mu_1 - \mu_2 = 0,$$
$$H_1: \ \mu_1 \neq \mu_2 \Leftrightarrow \mu_1 - \mu_2 \neq 0.$$

Algoritm hüpoteeside paari kontrollimiseks:

1. Moodustame hinnangu parameetrite vahele, $\hat{\mu}_1 - \hat{\mu}_2 = \bar{x} - \bar{y}$.

2. Normeerime vastava statistiku ja leiame selle jaotuse null olukorras,

$$T = \frac{\bar{X} - \bar{Y} - E(\bar{X} - \bar{Y})}{\sqrt{D(\bar{X} - \bar{Y})}},$$

   kus $E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$.

3. Test-statistiku T jaotus sõltub informatsioonist ÜK dispersioonide $\sigma_i^2, i = 1, 2$ kohta. Teoreemide 11 ja 12 tõestamise käigus oleme näidanud järgmiste jaotuste kehtivust.

   - **ÜK dispersioonid on teada**. Siis $D(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ ja

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1).$$

   - **ÜK dispersioonid pole teada, kuid võrdsed**. Siis $D(\bar{X} - \bar{Y}) = \sigma^2(\frac{1}{n_1} + \frac{1}{n_2})$ ja

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2),$$

     kus

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n_1}(x_i - \bar{x})^2 + \sum_{i=1}^{n_2}(y_i - \bar{y})^2}{n_1 + n_2 - 2}}.$$

   - **Info ÜK dispersioonide kohta puudub**. Siis $\hat{D}(\bar{X} - \bar{Y}) = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$ ja

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(f),$$

     kus

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \bar{x})^2 \text{ ja } s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(y_i - \bar{y})^2,$$

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

4. Test-statistikut T saame välja arvutada vaid hüpoteesi $H_0 : \mu_1 - \mu_2 = 0$ eeldusel. Siis asendub vahe $(\mu_1 - \mu_2)$ test-statistiku T avaldistes nulliga.

5. Test: kui $|T| > q_{\alpha/2}$, siis kummutada H_0. Siin $q_{\alpha/2}$ on vastavalt $N(0,1)$, $t(n_1+n_2-2)$ või $t(f)$ jaotuste täiendkvantiil.

5.6 Kahe üldkogumi dispersiooni võrdlemine normaaljaotuse korral

Eespool nägime, et test-statistiku kasutamine normaaljaotuse keskväärtuste võrdlemisel oleneb sellest, kas üldkogumite dispersioonid on võrdsed või mitte. See viib vajadusele tundmatute ÜK dispersioonide korral kontrollida hüpoteeside paari nende võrdsuse kohta. Teine näide, kus kasutatakse dispersioonide võrdlust, on dispersioonanalüüs, mille kirjeldust antakse peatükis ...

Olgu meil kaks üldkogumijaotust: $X \sim N(\mu_x, \sigma_x)$, millest on valim mahuga n_x ja $Y \sim N(\mu_y, \sigma_y)$, millest on valim mahuga n_y. Valimitest on leitud s_x ja s_y.

Algoritm:

- Huvitume hüpoteesist
$$H_0 : \ \sigma_x^2 = \sigma_y^2 \text{ ehk } \sigma_x = \sigma_y.$$

- Varasemast teame, et normaaljaotusega vaatluste korral on järgmised statistikud χ^2-jaotusega:
$$U = \frac{n_x - 1}{\sigma_x^2} s_x^2 \sim \chi^2(n_x - 1), \ \ V = \frac{n_y - 1}{\sigma_y^2} s_y^2 \sim \chi^2(n_y - 1).$$

- Moodustame juhusliku suuruse Z, mis on F-jaotusega:
$$Z = \frac{U/(n_x - 1)}{V/(n_y - 1)} = \frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2} \sim F(n_x - 1, n_y - 1).$$

- Kui H_0 kehtib (null olukord), siis
$$Z = \frac{s_x^2}{s_y^2} \sim F(n_x - 1, n_y - 1).$$

- Hüpoteesi H_0 kehtivuse korral hindavad nii s_x^2 kui ka s_y^2 sama üldkogumidispersiooni, seega nende jagatis varieerub 1 ümbruses.

- Kui Z on palju suurem 1-st, on põhjust arvata, et $\sigma_x^2 > \sigma_y^2$.

- Hüpoteeside paari
$$H_0 : \ \sigma_x^2 \leq \sigma_y^2,$$
$$H_1 : \ \sigma_x^2 > \sigma_y^2,$$
kontrollimiseks teeme testi
$$z > q_\alpha(n_x - 1, n_y - 1) \xrightarrow{+} \text{kummutame } H_0,$$
kus z on statistiku Z väärtus antud valimite korral.

- Testi olulisuse nivoo ehk esimest liiki vea tõenäosus on α.
- Hüpoteeside paari

$$H_0: \ \sigma_x^2 \geq \sigma_y^2,$$
$$H_1: \ \sigma_x^2 < \sigma_y^2,$$

kontrollimiseks vahetame tunnused ära ja teeme eelmise testi või võrdleme suurust Z $(1-\alpha)$-täiendkvantiiliga:

$$Z < q_{1-\alpha}(n_x - 1, n_y - 1) \xrightarrow{+} \text{kummutame } H_0;$$

- Kahepoolse alternatiivi korral viime läbi mõlemad ülalnimetatud testid. H_0 kummutamisel on vea tegemise tõenäosus 2α.

Märkus. Otsust dispersioonide võrdumise või mittevõrdumise kohta tuleb teha keskmiste võrdlemisel, $H_0 : \mu_1 = \mu_2$, $H_1 : \mu_1 \neq \mu_2$. Kontrollimise eeskiri sõltub sellest, millist infot saame kasutada ÜK dispersioonide kohta.

Oletame, et kontrollime kõigepealt hüpoteeside paari ÜK dispersioonide kohta, kus olulisuse nivoo on 0,05. Kui dispersioonid on võrdsed (jääme H_0 juurde), siis kasutame t-testi võrdsete dispersioonidega; kui aga erinevad, siis t-testi erinevate ja tundmatute dispersioonidega. T-testi kasutamisel lubame I liiki veaks samuti 0,05. Kuid sellisel juhul me väljastame t-testi juures vale olulisuse nivoo väärtuse, kuna dispersioonide võrdlemisel juba kasutasime ühte vea tõenäosust 0,05.

5.7 Hüpoteesid, mis põhinevad normaaljaotusega lähendamisel

Olgu valim $x_1, x_2, ..., x_n$ mingist jaotusest $F(\theta)$, kus θ on tundmatu (jaotuse keskväärtus või selle funktsioon). Loobume nõudest, et F on normaaljaotusega.

Hüpoteeside paariks olgu

$$H_0: \ \theta = \theta_0,$$
$$H_1: \ \theta \neq \theta_0$$

(võivad olla ka teised alternatiivid). Vaatame parameetri θ mõjusat punkthinnangut $\hat{\theta}$. Üldistel eeldustel kehtib vastava test-statistiku jaoks, et kui H_0 on õige, siis protsessis $n \longrightarrow \infty$,

$$U = \frac{\hat{\boldsymbol{\theta}} - \theta_0}{\sqrt{D\hat{\boldsymbol{\theta}}}} \longrightarrow N(0,1),$$

$$V = \frac{\hat{\boldsymbol{\theta}} - \theta_0}{\sqrt{\hat{D}\hat{\boldsymbol{\theta}}}} \longrightarrow N(0,1).$$

Suure valimimahu korral saame nüüd kasutada testimiseks normaaljaotuse täiendkvantiile (vt teema 5.3). Antud hüpoteesidepaari korral kontrollime, kas

$$|u| > \lambda_{\alpha/2} \text{ või } |v| > \lambda_{\alpha/2}.$$

Kui võrratus on tõene, siis kummutame H_0. Peame aga meeles pidama, et testi olulisuse nivoo on nüüd *ligikaudu* α (sest teststatistik on vaid ligikaudu normaaljaotusega). Praktikas kasutatakse testimisel tavaliselt t-kvantiile, sest

- $t_\alpha(n-1) > \lambda_\alpha$ ja protsessis $n \longrightarrow \infty$, $t_\alpha(n-1) \longrightarrow \lambda_\alpha$,
- t-kvantiilide kasutamine vähendab esimest liiki vea tõenäosust, sest sel juhul on raskem kummutada hüpteesi H_0. Kui aga õnnestub kummutada, on vea tõenäosus väiksem sellest, mis oleks olnud normaaljaotuse kvantiili kasutades.

5.7.1 Rakendus binoomjaotusele (üks valim)

Huvitagu meid sündmuse A osakaal (tõenäosus) p üldkogumis. Näiteks 'vaeste osakaal Eestis' või 'tervenemise tõenäosus antud ravimi korral'.

Võtame juhusliku valimi $x_1, x_2, ..., x_n$ ja olgu y sündmuse A esinemiste arv valimis. Soovime testida hüpoteesipaari

$$\begin{aligned} H_0: \ p &= p_0, \\ H_1: \ p &\neq p_0 \text{ (või alternatiivid } p < p_0, \ p > p_0). \end{aligned}$$

Vaatame p punkthinnangut $\hat{p}$.

Vastava statistiku $\hat{\boldsymbol{p}}$ jaotus läheneb normaaljaotusele n kasvades ($np_0 \geq 10, \ n(1-p_0) \geq 10$). Seega, suure valimimahu korral

$$U = \frac{\hat{\boldsymbol{p}} - E\hat{\boldsymbol{p}}}{\sqrt{D\hat{\boldsymbol{p}}}} \overset{H_0 \text{ õige}}{=} \frac{\hat{\boldsymbol{p}} - p_0}{\sqrt{p_0(1-p_0)/n}} \dot{\sim} N(0,1).$$

Test: $|u| > \lambda_{\alpha/2}$ (teiste alternatiivide korral vastavalt $u < -\lambda_\alpha, u > \lambda_\alpha$). Vea tegemise tõenäosus hüpoteesi H_0 kummutamisel on ligikaudu α.

Näide 49 *Seoses muutunud tehnoloogiaga on vaja kontrollida, kas toote maitseomadused on ka muutunud. Kasutatakse nn kolmiktesti. Igaüks* 500 *inimesest maitseb toodet kolmes pakis, milledest kaks sisaldavad vana ja üks uue tehnoloogia järgi valmistatud toodet. Isikutel palutakse valida pakk, mis maitseb erinevalt.*

200 *isikut osutasid uue tehnoloogia pakile. Kui maitseerinevus puudub, siis valitakse uue tehnoloogia pakk tõenäosusega $p = 1/3$. Kui erinevus on, siis $p > 1/3$. Seega kontrollitav hüpoteesipaar on:*

$$\begin{aligned} H_0: \ p &= 1/3, \\ H_1: \ p &> 1/3. \end{aligned}$$

Firma on rahul olulisuse nivooga $\alpha = 0.05$, $\lambda_{0.05} = 1.64$. Arvutame

$$u = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{\frac{200}{500} - \frac{1}{3}}{\sqrt{\frac{1}{3}\frac{2}{3}\frac{1}{500}}} = 3.2.$$

Kuna $3.2 > 1.64$, siis kummutame H_0 ja väidame, et maitseerinevus on tõepoolest olemas.

5.7.2 Rakendus binoomjaotusele (kaks valimit)

Olgu meil nüüd kaks valimit vastavalt mahtudega n_1 ja n_2. Olgu y_1 – omaduse A esinemiste arv ühes valimis, st $y_1 \leftarrow Y_1 \sim B(n_1, p_1)$ ja y_2 – omaduse A esinemiste arv teises valimis, st $y_2 \leftarrow Y_2 \sim B(n_2, p_2)$. Nüüd pakub huvi kahe üldkogumi võrdlemine, st kas omaduse A tõenäosused p_1 ja p_2 on võrdsed:

$$\begin{aligned} H_0: \ p_1 &= p_2 \ \Leftrightarrow p_1 - p_2 = 0, \\ H_1: \ p_1 &\neq p_2 \ (p_1 < p_2, \ p_1 > p_2). \end{aligned}$$

Vaatame statistikut $\hat{\boldsymbol{p}}_1 - \hat{\boldsymbol{p}}_2 = \frac{Y_1}{n_1} - \frac{Y_2}{n_2}$. Leiame

$$E(\hat{\boldsymbol{p}}_1 - \hat{\boldsymbol{p}}_2) = p_1 - p_2,$$

$$D(\hat{\boldsymbol{p}}_1 - \hat{\boldsymbol{p}}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$$

Moodustame normeeritud suuruse (keskväärtusega 0 ja standardhälbega 1):

$$U = \frac{\hat{\boldsymbol{p}}_1 - \hat{\boldsymbol{p}}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \overset{H_0 \text{ õige}}{=} \frac{\hat{\boldsymbol{p}}_1 - \hat{\boldsymbol{p}}_2}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}},$$

kus p nimetajas tähistab ühist osakaalu ($p = p_1 = p_2$). Kuna p pole teada, kasutame tema suurima tõepära hinnangut (*kontrollida*!),

$$\hat{p} = \frac{y_1 + y_2}{n_1 + n_2}.$$

Vastav statistik on H_0 õigsuse ja suure valimimahu korral ligikaudu standardse normaaljaotusega:

$$U = \frac{\hat{\boldsymbol{p}}_1 - \hat{\boldsymbol{p}}_2}{\sqrt{\hat{\boldsymbol{p}}(1-\hat{\boldsymbol{p}})(\frac{1}{n_1} + \frac{1}{n_2})}} \dot{\sim} N(0,1).$$

Sellele rajame testi: kui $|u| > \lambda_{\alpha/2}$, siis kummutame H_0 (vastavalt $u < -\lambda_\alpha$ või $u > \lambda_\alpha$ teiste alternatiivide korral). Testi olulisuse nivooks on ligikaudu α.

Näide 50 *Juhuslikult valitud 563 mehele ja 684 naisele esitati küsimus: "Kas Te usute elu peale surma?". 425 meest (75,49 %) ja 550 naist (80,40 %) vastasid jaatavalt.*

Olgu p_1 meeste osakaal, kes vastasid jaatavalt ja p_2 - vastav naiste osakaal. Hüpoteeside paar, mida soovitakse kontrollida olulisuse nivool 0,05 on

$$H_0 : \ p_1 = p_2, \ H_1 : \ p_1 \neq p_2$$

Leiame:

$$\hat{p} = \frac{425 + 550}{563 + 684} = 0,7819.$$

Seega,

$$u = \frac{0,7549 - 0,8040}{\sqrt{0,7819(1-0,7819)\left(\frac{1}{563} + \frac{1}{684}\right)}} \approx -2,0936.$$

Kuna $|u| = 2,0936 > \lambda_{0,05/2} = 1,96$, siis kummutame H_0 ja loeme erinevuse tõestatuks.

5.8 P-meetod

Otsene test, ehk p-meetod on alternatiivne meetod hüpoteeside kontrollimisel. Oleme seda juba varasemalt õppinud, kuid siin tuletame meelde. Alustame järgmisest näitest.

Näide 51 *Üks rehvifirma reklaamib, et nende uute talverehvidega libisemine jääl on kõige lühem.*

Üks teine firma, mis omab suurt autoparki plaanib neid rehve kasutusele võtta ainult peale põhjaliku kontrolli.

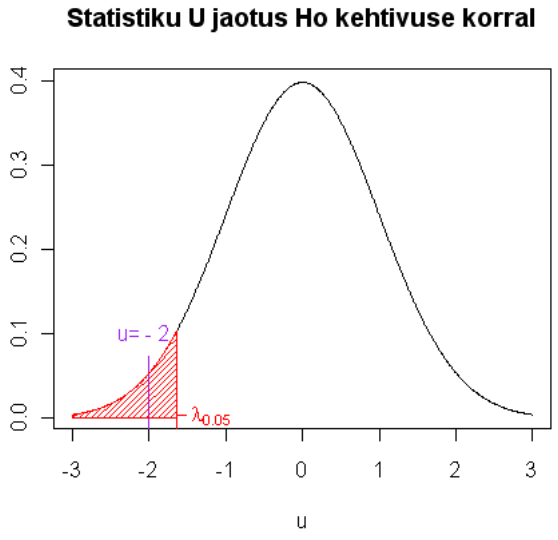
Sõltumatu laboratoorium testib neid rehve 36 auto peal ja väljastab, et keskmise libisemise kaugus on 44,4 m. Samuti annab laboratoorium ka infot traditsiooniliste rehvide kohta: libisemise kaugused on normaaljaotusest keskväärtusega 45,6 m ning standardhälbega 3,6 m.

Kontrollitav hüpoteeside paar on: $H_0: \ \mu \geq 45,6; \ \ H_1: \ \mu < 45,6.$

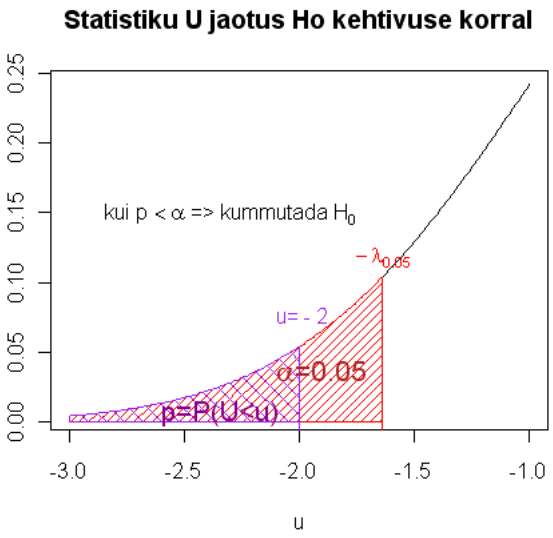
Kontrollime kõigepealt hüpoteside paari traditsioonilise meetodiga:

1. *Punkthinnang keskväärtusele,* $\hat{\mu} = \bar{x}$.
2. *Hinnangu jaotus* H_0 *kehtivuse korral:* $\bar{X} \overset{H_0}{\sim} N(45,6; \frac{3,6}{\sqrt{36}})$
3. *Normeeritud kujul:*
$$U = \frac{\bar{X} - 45,6}{\frac{3,6}{\sqrt{36}}} \overset{H_0}{\sim} N(0,1)$$
4. *Arvutame statistiku väärtuse valimi korral ja võrdleme kriitilise väärtusega* λ_α,
$$u = \frac{44,4 - 45,6}{\frac{3,6}{\sqrt{36}}} = -2; \ \ \lambda_{0,05} = 1,64.$$
5. *Kuna* $u < -\lambda_{0,05}$, *kummutame* H_0, *ehk uued rehvid on tõepoolest paremad (valekummutamise tõenäosus on 0,05).*

Järgmine joonis demonstreerib statistiku u realiseerunud väärtust ning kriitilist piirkonda, moodustatud täiendkvantiili $-\lambda_{0,05}$ *abil. Näeme, et tõepoolest langeb statistiku väärtus kriitilisse piirkonda, mis omakorda tähendab, et eeldatud* H_0 *väide pole väga tõenäoline, ehk tuleks* H_0 *kummutada.*



Traditsiooniline meetod võrdleb omavahel väärtuseid, mis asuvad joonise x-teljel. P-meetod aga võrdleb pindalaid, mis on moodustatud test-statistiku- ja täiendkvantiiliga (ning vastavalt test-statistiku jaotuse tõenäosusfunktsiooniga). Järgmisel joonisel on suurendatud eelmise grafiku vasakpoolne saba:



Näeme, et väide $u < -\lambda_{0,05}$ on samaäärne väitega $p < \alpha$, kus p on nn olulisuse tõenäosus, mis vastab piirkonnale $(-\infty,\ -\lambda_\alpha)$.

Üldiselt võib **p-meetodi algoritmi** sõnastada järgmiselt:

1. Olgu teststatistiku $T(\mathbf{X})$ väärtus antud valimi korral t (meie näites U ja $u = -2$).
2. Püüame otsustada, kas väärtus t on ebatavaline H_0 kehtivuse jaoks, kas ta pigem viitab vajadusele H_0 kummutada.
3. Leiame t esinemistõenäosuse H_0 kehtivuse korral. Et üldjuhul on $T(\mathbf{X})$ jaotus pidev ja üksikväärtuse tõenäosus on 0, siis leitakse väärtusega t määratud piirkonna C_t tõenäosus. (Meie näites $C_u = (-\infty, -2)$.)
4. Leiame $p = P(T(\mathbf{X}) \in C_t \ |H_0)$. Seda tõenäosust nimetatakse olulisuse tõenäosuseks (inglise keeles p-value).
5. Otsene test: kui p on väike, siis kummutame H_0.

Näide 52 *Eelmise näite jätk. Leiame olulisuse tõenäosuse rehvinäites,*

$$p = P(T(\mathbf{X}) \in C_t \ |H_0) = P(U < -2 | U \sim N(0,1)) = \Phi(-2) = 0,0228.$$

Kui olulisuse nivoo $\alpha = 0,05$, siis saame, et

$$p < \alpha => \ \textit{kummutame } H_0.$$

Järedus peab olema sama, mis traditsioonilisel meetodil.

Järgmisena vaatleme veel ühte p-meetodi näidet.

5.8.1 Märgitest

Märgitest on nn mitteparameetriline meetod, mis võrdleb mitte jaotuse parameetreid, vaid kahte jaotust omavahel (mediaane) sõltuvate valimite korral. Näiteks, kas inimeste kaalud enne ja pärast 3-kuuset dieedi on sama jaotusega või mitte. Kuna test on mitteparametriline, siis ei nõua see mingit jaotuse eeldust vaatlusandmetele.

Märgitesti saab rakendada nii arvulise kui ka järjestustunnuse korral.

Valimiks on n paari väärtuseid/mõõtmistulemusi $(x_i, y_i), i = 1, 2, ..., n$ ja n on valimimaht. Kontrollitavaks hüpoteeside paariks on

$$H_0: \ X \text{ ja } Y \text{ jaotused on võrdsed;}$$

$$H_1 : Y \text{ jaotus on nihutatud } X \text{ suhtes.}$$

Kontrollimise algoritm:

1. Välja jätta need paarid, kus $x_i = y_i, \ i = 1, ..., n$.
2. Ülejäänute jaoks moodustada uus tunnus V, mille väärtusteks on

$$v_i = \begin{cases} 1, \text{ kui } x_i < y_i \\ 0, \text{ muidu.} \end{cases}$$

3. Kui X ja Y jaotused on võrdsed, siis $V \overset{H_0}{\sim} Be(1/2)$.
4. Olgu $z = \sum v_i$. Siis vastav juhuslik suurus on Z, mille jaotuseks on

$$Z \overset{H_0}{\sim} Bin(m, 1/2),$$

   kus m on erinevate paaride arv valimis.

5. Edasi rakendada p-meetodit.

Näide 53 *17-le inimesele näidati kahte erinevat reklaami ja paluti neil mõlemat reklaami hinnata 5 pallises skaalas. Tabelis on toodud küsitluse tulemused:*

Reklaam 1	*4*	*2*	*4*	*4*	*4*	*2*	*3*	*4*	*3*	*5*	*3*	*3*	*4*	*5*	*4*	*5*	*2*
Reklaam 2	*2*	*4*	*5*	*4*	*4*	*5*	*3*	*5*	*5*	*4*	*4*	*5*	*5*	*5*	*5*	*4*	*3*
v_i	*0*	*1*	*1*	.	.	*1*	.	*1*	*1*	*0*	*1*	*1*	*1*	.	*1*	*0*	*1*

Seega, $z = \sum v_i = 10$. *Kuivõrd on selline väärtus tõenäone kui* $p = 1/2$*? Olulisuse tõenäosuse leidmiseks leiame nii 10 kui ka teiste ekstreemsete väärtuste tõenäosuse:*

$$p = P\left(Z \geq 10 | Z \sim Bin(13, \frac{1}{2})\right) = \sum_{k=10}^{13} C_{13}^k (0,5)^k (1-0,5)^{13-k}.$$

$$= 0,03491 + 0,00952 + 0,00159 + 0,00012 = 0,04614$$

Näeme, et saadud p-väärtus on väiksem kui olulisuse nivoo $\alpha = 0,05$. *Seega, z väärtusega määratud piirkond on küllaltki ebatavaline* H_0 *kehtivuse korral ja seega kummutame* H_0. *Hinnangud kahele reklaamile on erinevad.*

6. Lihtne lineaarne regressioon

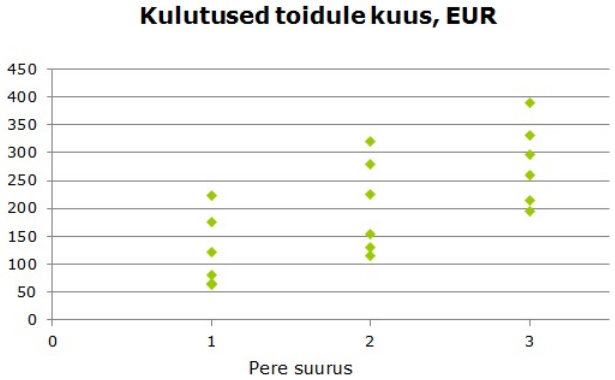
Antud peatükis võtame hindamisteooria kokku ühe konkreetse analüüsi näitel - regressioonanalüüs lihtsa lineaarse mudeli korral. Antud peatükk ei hõlma kogu regressioonanalüüsi teooriat, vaid demonstreerib õpitu teooria rakendust. Paetükk põhineb suures osas õpikul Traat (2007).

6.1 Regressioonimudel

Seni oli meil tegemist ühe tunnuse vaatlusandmetega. Meid huvitas selle tunnuse üldkogumijaotus, õieti tema jaotusparameetrid. Statistilised otsustused üldkogumi jaotuse kohta rajasime sellele, et vaatlusväärtuste asemel vaatlesime juhuslikke suurusi, mis produtseerisid need väärtused. Seega seni oli meil tegu vaatluste mudeliga $y_i \sim F(\theta),\ i = 1, 2, \ldots, n$, y_i sõltumatud.

Edasi vaatleme olukorda, kus nähtust iseloomustatakse kahe tunnusega (y, x). Seejuures x-tunnuse väärtusi vaadeldakse fikseeritutena (mittejuhuslikena), aga y-tunnust vaadeldakse juhuslikuna iga fikseeritud x-väärtuse korral.

Näide 54 *Olgu x – pere suurus ja y – pere kulutus toidule. Vaadeldes näiteks 3-liikmelisi peresid ($x = 3$), märkame et nende perede kulutus toidule varieerub (on teatava juhusliku suuruse realisatsioonid). Samuti on $x = 1$ korral y-tunnus juhuslik, küll aga ilmselt teise keskväärtuse ja ehk ka teise dispersiooniga. Selles näites x kasvamisel y keskmiselt kasvab.*



Kahe tunnuse korral huvitabki meid tunnustevaheline seos, st kuidas mõjub x-tunnuse muutumine y-tunnusele. Kui seos on olemas, siis saab x-väärtuste abil prognoosida y-väärtusi. Siinjuures nimetatakse y-tunnust uuritavaks (ka sõltuvaks ehk funktsioontunnuseks) ja x-tunnust seletavaks (ka sõltumatuks ehk argumenttunnuseks).

Kahe tunnuse vaatlusandmeteks on paarid:

$$\begin{pmatrix} y_i \\ x_i \end{pmatrix},\ i = 1, 2, \ldots, n. \tag{6.1}$$

Neid paare saab sageli vaadelda kui realisatsioone järgmisest mudelist.

Definitsioon 33 *Vaatlusandmete* (6.1) *lihtsaks lineaarseks regressioonimudeliks on:*

$$\begin{cases} y_i \sim N(\mu_i, \sigma),\ \textit{sõltumatud}\ \forall i, \\ \mu_i = \alpha + \beta x_i, \\ x_i\ \textit{mittejuhuslikud.} \end{cases} \tag{6.2}$$

Mudelit nimetatakse lihtsaks, sest tegemist on üheainsa argumenttunnusega (üldjuhul võib neid olla rohkem), ja lineaarseks seetõttu, et funktsioontunnuse keskväärtus on lineaarselt seotud argumenttunnusega (üldjuhul on võimalikud ka teistsugused seosed). Ka funktsioontunnuse dispersioon võib üldjuhul sõltuda vaatlusest i, siin on ta aga konstantselt σ^2 iga i korral.

Mudeli (6.2) saab alternatiivselt esitada seostega

$$\begin{cases} y_i = \alpha + \beta x_i + \varepsilon_i, \\ \varepsilon_i \sim N(0, \sigma),\ \text{sõltumatud}\ \forall i, \\ x_i\ \text{mittejuhuslikud.} \end{cases} \tag{6.3}$$

Definitsioon 34 *Sirget $y = \alpha + \beta x$ nimetatakse tunnuste x ja y vaheliseks regressioonisirgeks (y regressioon x järgi).*

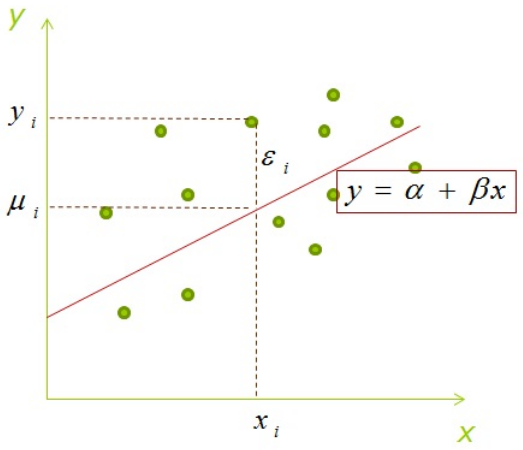
Regressioonisirge väljendab y-tunnuse keskmise lineaarset sõltuvust x-tunnusest.

Definitsioon 35 *Suurust $\mu_i = \alpha + \beta x_i$ nimetatakse y-tunnuse prognoosiks punktis x_i. Prognoosid paiknevad regressioonisirgel. Vaatlusväärtuse ja prognoosi vahet $y_i - \mu_i = \varepsilon_i$ nimetatakse prognoosijäägiks (ka jäägiks või mudeli veaks).*

Vaatlusväärtust (y_i, x_i), prognoosi μ_i, ja prognoosijääki ε_i iseloomustab allpool toodud joonis. Märgime, et regressioonisirge abil saame y-väärtust prognoosida suvalises punktis x, kuigi, nagu hiljem näeme, mitte ühesuguse täpsusega kõikjal.

Regressioonimudeli parameetriteks on α, β ja σ^2 ning kõik nad iseloomustavad tunnuste (y, x) vahelist seost.

Mudeli vabaliige α näitab y prognoosiväärtust punktis $x = 0$. Sisulistest kaalutlustest lähtuvalt peab mõnikord $x = 0$ korral olema ka $y = 0$. Siis on kasulik fikseerida $\alpha = 0$, mis tähendab, et tegemist on vabaliikmeta ehk koordinaatide alguspunkti läbiva regressioonisirgega.



Tähtis sõltuvuse näitaja on regressioonikordaja ehk sirge tõus β. Kui $\beta = 0$, siis sõltuvust tunnusest x ei ole ($y_i = \alpha + \varepsilon_i$), st kõigi x-väärtuste korral on y-tunnus keskmiselt võrdne konstandiga α. Kui $\beta > 0$ ($\beta < 0$), on samasuunaline (vastassuunaline) sõltuvus. Mida suurem on β absoluutväärtus, seda kiiremini muutub y prognoos tunnuse x muutumisel. Tunnuse x muutumisel ühiku võrra muutub y prognoos β võrra (y ühikutes).

Kolmas mudeli parameeter, vaatlusandmete dispersioon σ^2 regressioonisirge (st oma prognoosi ümber) on ühtlasi jääkide dispersioon. Kui $\sigma^2 = 0$ on $\varepsilon_i \equiv 0$ (konstantselt 0), ning siis on y ja x vahel funktsionaalne lineaarne seos. Mida väiksem on σ^2, seda paremini lähendab regressioonisirge y-tunnuse väärtusi.

6.2 Mudeli parameetrite hindamine

Nägime, et regressioonimudeli parameetritel on oluline tähtsus tunnustevahelise sõltuvuse kirjeldamisel. Mudeli parameetrid pole aga arvuliselt teada.

Uurija käsutuses on juhuslik valim üldkogumist, st sõltumatute vaatluste paarid $\begin{pmatrix} y_i \\ x_i \end{pmatrix}$, $i = 1, \ldots, n$, mille kohta eeldatakse, et nad on realisatsioonid mudelist, ja seetõttu rahuldavad tingimusi (6.2) või (6.3). Osutub, et seni läbivaadatud meetoditega oleme võimelised mudeli parameetreid hindama nii punkthinnangute kui vahemikhinnangute mõttes, kontrollima hüpoteese parameetrite kohta ja andma ka oma väidete õigsuse mõõdud – usaldus- ja olulisusnivood.

6.2.1 Regressiooniparameetrite punkthinnangud

Tuletame parameetrite α ja β punkthinnangud vähimruutude meetodil. Selleks vaatame vaatluste hälbeid keskväärtusest

$$Q(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \mu_i)^2 = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2.$$

ja minimiseerime avaldise α ja β suhtes.

Osutub, et arvutusi on lihtsam teha, kui parametriseerime regressioonisirge teisiti:

$$\mu_i = \alpha + \beta x_i + \beta\bar{x} - \beta\bar{x} = \alpha + \beta\bar{x} + \beta(x_i - \bar{x}) = \alpha' + \beta(x_i - \bar{x}),$$

kus

$$\alpha' = \alpha + \beta\bar{x}. \tag{6.4}$$

Parameeter α' on y prognoos $x = \bar{x}$ korral. Tal on sageli sisukam tähendus kui parameetril α, näiteks on α' keskmise suurusega pere prognoositav toidukulu ülaltoodud näites. Vajadusel saab esialgse α hinnangu leida seosest (6.4), kui α' ja β on hinnatud.

Vaatame hälvete ruutude summat:

$$Q(\alpha', \beta) = \sum_{i=1}^{n} (y_i - (\alpha' + \beta(x_i - \bar{x})))^2.$$

Definitsioon 36 *Parameetrite α' ja β vähimruutude hinnanguks nimetatakse $\hat{\alpha}'$, $\hat{\beta}$, mis minimiseerivad $Q(\alpha', \beta)$, st*

$$\min_{\alpha',\ \beta} Q(\alpha', \beta) = Q(\hat{\alpha}', \hat{\beta}).$$

Leides funktsiooni Q osatuletised α' ja β järgi, saame

$$\begin{cases} \partial \frac{Q(\alpha',\beta)}{\partial \alpha'} = -2 \sum_{i=1}^{n} (y_i - \alpha' - \beta(x_i - \bar{x})), \\ \partial \frac{Q(\alpha',\beta)}{\partial \beta} = -2 \sum_{i=1}^{n} [(y_i - \alpha' - \beta(x_i - \bar{x}))(x_i - \bar{x})]. \end{cases}$$

Võrdsustades osatuletised nulliga ja arvestades, et $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, saame esimesest võrandist $\hat{\alpha}' = \bar{y}$. Asendades tulemuse teise võrrandisse, saame ka β hinnangu. Lõppkokkuvõttes on regressiooniparameetrite vähimruutude hinnanguteks:

$$\hat{\alpha}' = \bar{y}, \tag{6.5}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{6.6}$$

Märkus. Lihtsad esialgsed hinnangud regressiooniparameetritele saab anda ka käsitsi. Selleks tuleb vaatlused (y_i, x_i) kanda graafikule ja silma järgi tõmmata sirge nii, et punktide (y_i, x_i) vertikaalkaugused sirgest oleksid minimaalsed. Graafikult loetud sirge tõus ja vabaliige ongi regressiooniparameetrite α ja β hinnangud.

Uurime leitud hinnangute omadusi – nihketust ja mõjusust. Kasutades eeldusi (6.2) vaatlustele (y_i, x_i), saame:

$$\begin{aligned} E\hat{\alpha}' &= E\bar{y} = \frac{1}{n}\sum_{i=1}^{n} Ey_i = \frac{1}{n}\sum_{i=1}^{n} \mu_i \\ &= \frac{1}{n}\sum_{i=1}^{n} (\alpha' + \beta(x_i - \bar{x})) = \frac{1}{n}\sum_{i=1}^{n} \alpha' = \alpha', \end{aligned}$$

$$D\hat{\alpha}' = D\bar{y} = \frac{1}{n^2}\sum_{i=1}^{n} Dy_i = \frac{\sigma^2}{n} \to 0.$$

Seega $\hat{\alpha}'$ on nihketa ja mõjus hinnang parameetrile α'. Teise parameetri hinnangu korral saame

$$E\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})E(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta,$$

sest $E(y_i - \bar{y}) = \mu_i - E\bar{y} = \mu_i - \alpha' = \alpha' + \beta(x_i - \bar{x}) - \alpha' = \beta(x_i - \bar{x})$. Protsessis $n \to \infty$

$$D\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \; Dy_i}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \to 0.$$

Seega $\hat{\beta}$ on nihketa ja mõjus hinnang parameetrile β.

Hinnatud sirge abil saame anda hinnangud ka y-tunnuse prognoosidele. Punktis x_0 on prognoosi $\mu_0 = \alpha' + \beta(x_0 - \bar{x})$ hinnanguks $\hat{\mu}_0 = \hat{\alpha}' + \hat{\beta}(x_0 - \bar{x})$. Uurides selle hinnangu omadusi, saame

$$E\hat{\mu}_0 = E\hat{\alpha}' + E\hat{\beta}(x_0 - \bar{x}) = \mu_0,$$

$$D\hat{\mu}_0 = D\hat{\alpha}' + (x_0 - \bar{x})^2 D\hat{\beta} = \frac{\sigma^2}{n} + (x_0 - \bar{x})^2 \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Dispersiooni leidmisel kasutasime ka hinnangute $\hat{\alpha}'$ ja $\hat{\beta}$ sõltumatust, mis on õigustatud üksnes normaaljaotusega vaatluste y_i korral. Näeme, et prognoos punktis x_0 on seda parem, mida lähemal asub punkt x_0 keskmisele $\bar{x}$.

Vaatame valimist arvutatud prognoosijääke $y_i - \hat{\mu}_i$, kus $\hat{\mu}_i = \hat{\alpha}' + \hat{\beta}(x_i - \bar{x})$. Jääkide ruutude summa,

$$Q_0 = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2,$$

iseloomustab regressioonisirge võimet andmeid lähendada (antud valimi korral).

Lõpuks tuletame kolmandale mudeli parameetrile σ^2 hinnangu suurima tõepära meetodil. Kuna $y_1, y_2, \ldots, y_n \leftarrow N(\mu_i, \sigma^2)$, siis tõepärafunktsioon esitub kujul

$$L(\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \mu_i)^2}.$$

Logaritmides saame

$$l(\sigma^2) = \sum_{i=1}^{n}[-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_i - \mu_i)^2],$$

millest leides $\frac{\partial l(\sigma^2)}{\partial \sigma^2}$ ja võrdsustades nulliga saame võrrandi

$$\sum_{i=1}^{n}\left[-\frac{1}{2}\frac{2\pi}{2\pi\sigma^2} + \frac{1}{2\sigma^4}(y_i - \mu_i)^2\right] = 0.$$

Summeerides ja nimetajast vabanedes saame $-n\sigma^2 + \sum_{i=1}^{n}(y_i - \mu_i)^2 = 0$. Avaldades siit σ^2 ja asendades tundmatu μ_i tema hinnanguga, saame:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 = \frac{Q_0}{n}.$$

Osutub, et saadud hinnang on nihkega. Nihketa hinnanguks dispersioonile σ^2 on

$$\hat{\sigma}^2 = \frac{Q_0}{n-2}. \tag{6.7}$$

Saadud hinnang iseloomustab jääkide hajuvust regressioonisirge ümber. Mida väiksem hajuvus, seda paremini kirjeldab sirge y-tunnuse väärtusi. Suuruse Q_0 jaotus on teada normaaljaotusega y_i korral, nimelt $Q_0 \sim \chi^2(n-2)$.

6.2.2 Vahemikhinnangud ja hüpoteeside kontroll regressiooniparameetrite korral

Paneme tähele, et regressioonimudeli eeldustel on parameetrite hinnangud normaaljaotusega:

$$\hat{\alpha}' \sim N(\alpha', \sigma/\sqrt{n}) \text{ ja } \hat{\beta} \sim N(\beta, \frac{\sigma}{\sqrt{S_{xx}}}),$$

kus

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

Vahemikhindamine taandub seega normaaljaotuse keskväärtuse vahemikhinnangu leidmisele. Varasemast teame, et sellisel juhul on kahepoolseteks usaldusvahemikeks

$$I_{\alpha'} = \hat{\alpha}' \pm \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

kui σ on teada ning

$$I_{\alpha'} = \hat{\alpha}' \pm t_{\alpha/2}(f)\frac{\hat{\sigma}}{\sqrt{n}},$$

kui σ on tundmatu. Samuti

$$I_{\beta} = \hat{\beta} \pm \lambda_{\alpha/2}\frac{\sigma}{\sqrt{S_{xx}}},$$

kui σ on teada ning

$$I_{\beta} = \hat{\beta} \pm t_{\alpha/2}(f)\frac{\hat{\sigma}}{\sqrt{S_{xx}}},$$

kui σ on tundmatu, kus $\hat{\sigma}^2 = \frac{Q_0}{n-2}$ ja $f = n - 2$.

Lõpuks vaatame hüpoteeside paari regressioonimudeli tähtsaima parameetri kohta:

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

Hüpotees H_0 väidab, et y-tunnus ei sõltu x-st. Meid huvitab sõltuvuse tuvastamine. Kuna $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{S_{xx}})$, on tegemist hüpoteesidega normaaljaotuse keskväärtuse kohta, mis on varasemas osas põhjalikult käsitletud. Kui σ on teada, moodustame statistiku

$$U = \frac{\hat{\beta}}{\sigma/\sqrt{S_{xx}}},$$

mis on $N(0,1)$ jaotusega H_0 kehtivuse korral. Testiks on:

$$|U| > \lambda_{\alpha/2} \xrightarrow{+} \text{kummutame } H_0.$$

Kui σ ei ole teada, vaatame statistikut

$$V = \frac{\hat{\beta}}{\hat{\sigma}/\sqrt{S_{xx}}},$$

mis on $t(f)$-jaotusega H_0 õigsuse korral, $f = n - 2$. Testiks on:

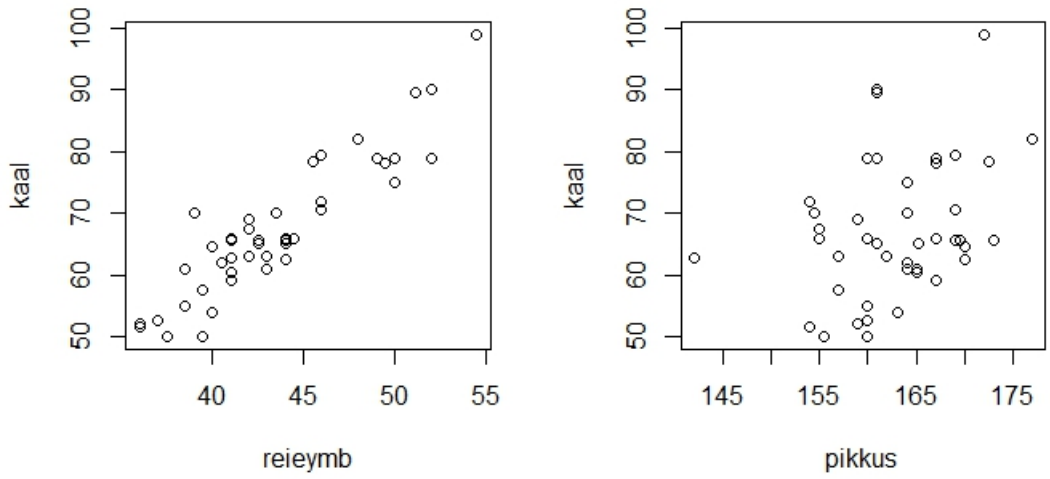
$$|V| > t_{\alpha/2}(f) \xrightarrow{+} \text{kummutame } H_0.$$

Teiste alternatiivide korral toimime analoogiliselt nagu kirjeldatud normaaljaotuse keskväärtuse jaoks.

Veendugem iseseisvalt, et vaadeldava regressioonimudeli korral α ja β suurima tõepära hinnangud langevad kokku nende vähimruutude hinnangutega.

Näide 55 *Nagu märgitud eespool, on regressioonimudeli koostamisel 2 eesmärki: sõltuvuse tuvastamine ja/või prognoosimine. Kordaja β erinevus 0-st väljendab y-tunnuse sõltuvust x-st. Suur β väärtus (kasutatud mõõtmisskaalal) väljendab tugevat sõltuvust. Tugev sõltuvus ei väljenda veel prognoosivõimet selles mõttes, et argumenttunnuse x korral võiksime väärtusega sirgel lähendada y-tunnuse väärtust. Viimaseks peab jääkide dispersioon olema väike. Tuleb meeles pidada, et valimist hinnatud mudeli parameetrid ei lange kokku tegelikega. Otsustusteks mudeli tegelike parameetrite kohta tuleb kasutada usalduspiire ja statistilisi teste.*

Soovime uurida, kas tunnus 'kaal' sõltub pikkusest ja ümbermõõtude tunnustest. Vaatame naiste andmestikku. Esmased otsustused saame teha hajuvusdiagramme vaadates (joonis allpool). Selget lineaarset trendi näeme kaalu ja reieümbermõõdu vahel. Kujutletava sirge tõus on suur, viidates tugevale sõltuvusele ja punktide hajuvus sirge suhtes on küllalt väike, viidates kaalu üsna täpsele prognoositavusele reieümbermõõdu kaudu.



Statistikapaketiga R arvutades saame hinnatud parameetritega mudeli:

$$kaal = -29.93 + 2.24\ reieymb + \varepsilon,$$

$$\hat{D}\varepsilon = \hat{\sigma}^2 = 4.83.$$

Hinnangu $\hat{\beta} = 2.24$ standardviga on 0.17, mille abil leitud 95%-usaldusvahemik $I_\beta = (1.91, 2.57)$ ütleb, et kordaja β erineb tõepoolest nullist, veelgi enam, on positiivne. Hüpoteesi
$H_0 : \beta = 0$ kontrollimiseks on arvutatud t-statistik, $t = 13.51$. Sellise väärtuse saamine on H_0 õigsuse korral vähetõenäoline, $p < 2 \cdot 10^{-16}$. Seega oleme päris kindlad, et $\beta \neq 0$. Usaldusvahemik on siinjuures märksa informatiivsem, öeldes et suure tõenäosusega on β võimalik väärtus 1.91 kuni 2.57.

Lõpuks, mida huvitavat on hinnatud mudelil sisuliselt öelda? Osutub, et reieümbermõõdu kasvades 1 cm võrra kasvab naise kaal keskmiselt 2.24 kg. Naine reieümbermõõduga 43.4 cm kaalub keskmiselt 67.3 kg. Kaalu standardhälve iga fikseeritud reieümbermõõdu korral on üsna väike, $\hat{\sigma} = 2.2$ kg, mistõttu saame regressioonisirge abil prognoosida naise kaalu küllaltki täpselt.

7. Ühefaktoriline dispersioonanalüüs

Järgnev peatükk põhineb samuti õpikul Traat (2006) ja demonstreerib õpitu teooria rakendust konkreetsel näitel – dispersioonanalüüsis.

7.1 Dispersioonanalüüsi mudel

Ka dispersioonanalüüsis on tegu uuritava tunnusega, mida vaadeldakse juhuslikuna, nimelt normaaljaotusega, ja seletavate tunnustega, mida vaadeldakse mittejuhuslikena. Eesmärgiks on seletavate tunnuste mõju avastamine uuritavale tunnusele. Seletavaid tunnuseid nimetatakse dispersioonanalüüsis *faktoriteks*. Faktor on diskreetne väheste väärtustega tunnus, mis võib olla nii arvuline kui ka mittearvuline (regressioonanalüüsis on seletav tunnus kindlasti arvuline). Meie vaatame 1 faktoriga dispersioonanalüüsi mudelit.

Illustreerime dispersioonanalüüsi andmestikku näite varal. Olgu y rottide kaalu juurdekasv ja x toit, mida nad saavad. Faktoril 'toit' on järgmised väärtused/tasemed: kaer, rukis,..., nisu. Igal faktori tasemel on katsealuseks m rotti, kelle kaalu juurdekasvu teatud ajavahemiku möödudes mõõdetakse. Vaatlusandmed saab esitada järgmise tabelina.

	Roti nr.				
x	1	2	...	m	
1. Kaer	y_{11}	y_{12}	$\cdots$	y_{1m}	$\bar{y}_{1\cdot}$
2. Rukis	y_{21}	y_{22}	$\cdots$	y_{2m}	$\bar{y}_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
k. Nisu	y_{k1}	y_{k2}	$\cdots$	y_{km}	$\bar{y}_{1\cdot}$

Tabelis on y_{ij} j-nda roti kaalu juurdekasv faktori tasemel i ja seda saab vaadelda kui juhusliku suuruse realisatsiooni, $\bar{y}_{i\cdot}$ on rea aritmeetiline keskmine. Tabelis toodud andmestiku kirjeldamiseks sobib dispersioonanalüüsi mudel.

Definitsioon 37 *Ühe faktoriga dispersioonanalüüsi mudeliks nimetatakse järgmist andmete kirjeldust:*

$$y_{ij} \sim N(\mu_i, \sigma),\ j = 1, 2, \ldots, m, \tag{7.1}$$

$$\mu_i = \mu + \alpha_i,\ i = 1, 2, \ldots, k. \tag{7.2}$$

St et vaatlused ($j = 1, 2, \ldots, m$) faktori fikseeritud tasemel i on sama normaaljaotusega ja taseme keskmine μ_i avaldub üldkeskmise μ ja faktori mõju α_i abil. Vaatlused y_{ij} on sõltumatud nii faktori samal tasemel kui ka erinevatel tasemetel.

Mudeli parameetriteks on μ, α_i ja σ^2. Tegemist on nn tasakaalustatud dispersioonanalüüsi mudeliga, kus vaatluste arv faktori tasemetel on sama (m), üldjuhul see nii ei pea olema.

Seosed (7.2) esitavad k võrrandit $k+1$ parameetri α_i ja μ määramiseks, kui μ_i on fikseeritud. Ühest lahendit ei leidu. Tõepoolest, kui μ', α_i' on lahendid, siis on seda ka $\mu'' = \mu' - \delta,\ \alpha_i'' = \alpha_i' + \delta,\ i = 1, 2, \ldots, k$. Et mudeli parameetrid oleksid üheselt hinnatavad, selleks tuleb kasutada kitsendusi. Klassikaliselt vaadeldakse kahte tüüpi kitsendusi faktori mõjule:

$$\sum_{i=1}^{n} \alpha_i = 0 \tag{7.3}$$

või

$$\alpha_1 = 0. \tag{7.4}$$

Erilist huvi pakub faktori mõju kindlakstegemine, st kas mõni α_i erineb nullist. Jaatava vastuse korral on faktoril mõju olemas. Mõju interpreteerimine rajaneb seosele (7.2) ja sõltub kitsenduste tüübist. Kitsenduste (7.3) korral näitab $\alpha_i > 0$, faktori taseme i suurendavat mõju uuritava tunnuse keskmisele (näiteks $\alpha_2 > 0$ ütleb, et toit 'rukis' suurendab keskmiselt α_2 võrra kaalu juurdekasvu). Kitsenduste (7.4) korral on $\mu_1 = \mu$ ja sel juhul mõõdab α_i erinevust esimese taseme keskmisest (näiteks $\alpha_2 > 0$ ütleb, et võrreldes kaera sööjatega on rukki sööjate keskmine juurdekasv α_2 võrra suurem.

Faktori mõju kindlakstegemine rajaneb y-tunnuse hajuvuse uurimisele valimis. Valimidispersioon lahutatakse sobivalt komponentideks. Sellest ka nimi – dispersioonanalüüs. Vaatame üldkeskmist ja reakeskmisi valimis:

$$\bar{y} = \frac{1}{km} \sum_{i=1}^{k} \sum_{j=1}^{m} y_{ij}, \quad \bar{y}_{i\cdot} = \frac{1}{m} \sum_{j=1}^{m} y_{ij}\ . \tag{7.5}$$

Nüüd lahutame koguvarieeruvuse (Sum of Squares Total) tasemete vaheliseks (Sum of Squares Between) ja tasemesiseseks (Sum of Squares Within) varieeruvuseks:

$$\begin{aligned} \underbrace{\sum_{i=1}^{k} \sum_{j=1}^{m} (y_{ij} - \bar{y})^2}_{SST} &= \sum_{i=1}^{k} \sum_{j=1}^{m} \left[(\bar{y}_{i\cdot} - \bar{y}) + (y_{ij} - \bar{y}_{i\cdot}) \right]^2 \\ &= \underbrace{\sum_{i=1}^{k} \sum_{j=1}^{m} (\bar{y}_{i\cdot} - \bar{y})^2}_{SSB} + \underbrace{\sum_{i=1}^{k} \sum_{j=1}^{m} (y_{ij} - \bar{y}_{i\cdot})^2}_{SSW}, \end{aligned}$$

sest

$$\sum_{i=1}^{k} \sum_{j=1}^{m} (\bar{y}_{i\cdot} - \bar{y})(y_{ij} - \bar{y}_{i\cdot}) = \sum_{i=1}^{k} (\bar{y}_{i\cdot} - \bar{y}) \sum_{j=1}^{m} (y_{ij} - \bar{y}_{i\cdot}) \overset{(7.5)}{=} 0.$$

Kui SSB on väike, siis $\bar{y}_{i\cdot} \approx \bar{y}$ ehk grupikeskmised ja üldkeskmine valimis on ligikaudu võrdsed, mis viitab sellele, et faktoril mõju puudub. Et otsus faktori mõju kohta oleks statistiliselt korrektne, vaatame hüpoteeside kontrolli faktori mõju kohta.

7.2 Hüpoteesid faktori mõju kohta

Soovime kontrollida hüpoteesi, et faktoril mõju puudub (y ei sõltu tunnusest x)

$$H_0 :\ \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0.$$

Meie mudeli eeldustel

$$\bar{y}_{i\cdot} = \frac{1}{m}\sum_{j=1}^{m} y_{ij} \sim N(\mu_i, \frac{\sigma}{\sqrt{m}}), \tag{7.6}$$

kus

$$E\bar{y}_{i\cdot} = \mu_i = \mu + \alpha_i. \tag{7.7}$$

Lihtne on näha, et reakeskmiste keskmine on üldkeskmine:

$$\frac{1}{k}\sum_{i=1}^{k} \bar{y}_{i\cdot} = \frac{1}{km}\sum_{i=1}^{k}\sum_{j=1}^{m} y_{ij} = \bar{y}.$$

Olgu H_0 õige, siis $E\bar{y}_{i\cdot} = \mu$ ja kehtib

$$\bar{y} \sim N(\mu, \frac{\sigma}{\sqrt{km}}). \tag{7.8}$$

Kasutades teoreeme 4.2.1 ja 4.2.2 (hii-ruut jaotuse kohta), saame

$$\frac{1}{\sigma^2/m}\sum_{i=1}^{k}(\bar{y}_{i\cdot} - \bar{y})^2 \sim \chi^2(k-1),$$

mis esitab tasemetevahelist varieeruvust, ehk

$$U = \frac{SSB}{\sigma^2} \sim \chi^2(k-1). \tag{7.9}$$

Saime, et $U \sim \chi^2(k-1)$ kehtib H_0 õigsuse korral.

Edasi vaatame i-nda rea hälvete ruutude summat. Saame

$$z_i = \frac{1}{\sigma^2}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{i\cdot})^2 \sim \chi^2(m-1).$$

Aditiivsuse teoreemist

$$\frac{SSW}{\sigma^2} = \sum_{i=1}^{k} z_i \sim \chi^2(k(m-1)) = \chi^2(n-k),$$

kus $n = km$ on kogu valimimaht. Järelikult

$$V = \frac{SSW}{\sigma^2} \sim \chi^2(n-k)$$

ja seda olenemata H_0 õigsusest. H_0 õigsuse korral kehtib

$$Z = \frac{U/(k-1)}{V/(n-k)} \sim F(k-1, n-k),$$

ehk

$$Z = \frac{(n-k)SSB}{(k-1)SSW} \sim F(k-1, n-k),$$

Näeme, et suur SSB toob kaasa Z suurenemise. Piisavalt suure Z korral kummutame H_0. Toetudes statistiku Z jaotuse omadustele, kummutame H_0 olulisuse nivool α ja ütleme, et faktoril on mõju uuritavale tunnusele, kui

$$Z > q_\alpha(k-1, n-k).$$

Leitud suurused asuvad programmipakettide väljundis ANOVA-tabelis (ANalysis Of VAriance), millel on harilikult järgmine kuju:

Source	DF	Sum of Squares	Mean Square	F-Stat	$Pr > F$
Model	$k-1$	SSB	$SSB/(k-1)$	z	$P(Z > z)$
Error	$n-k$	SSW	$SSW/(n-k)$		
C Total	$n-1$	SST			

Siin SSW iseloomustab andmete summaarset varieeruvust tasemete sees, keskmistatuna aga hindab mudeli parameetrit σ^2:

$$\hat{\sigma}^2 = \frac{SSW}{n-k}.$$

Valemist (7.1) ilmneb, et σ^2 on ühtlasi jääkide $\varepsilon_i = y_{ij} - \mu_i$ dispersiooniks. Jääki nimetatakse ka mudeli veaks, siit rea nimetus 'Error'. Mõned paketid kasutavad siin sõna 'Residual'. Tasemete vaheline varieeruvus SSB näitab, kui suure osa kirjeldab valitud faktortunnus koguvarieeruvusest, mida suurem see on, seda parem on mudel (siit rea nimetus 'Model'). Maksimaalselt hea mudel on selline, et $SSW = 0$ ja $SSB = SST$. Siis oleks iga y-väärtus x-tasemel konstant (mittejuhuslik) ja faktori mõju kohe näha. Tabeli põhjal teeme otsese olulisustesti ja ütleme, et faktoril on mõju uuritavale tunnusele, kui $P(Z > z)$ on väike. See aga ei ütle meile, missugusel faktori tasemel on funtsioontunnusele suurendav, missugusel vähendav mõju. Nende otsustuste tegemiseks on vaja uurida parameetrite hinnanguid.

7.3 Parameetrite hinnangud

Seosest (7.8) näeme, et $\bar{y}$ on nihketa hinnang üldkeskmisele μ, dispersiooniga $D\bar{y} = \sigma^2/(km)$. Seosest (7.6) järeldub, et $\bar{y}_{i\cdot}$ on nihketa hinnang taseme keskmisele μ_i. Seose (7.7) tõttu on faktori mõju nihketa hinnanguks

$$\hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y},$$

dispersiooniga

$$D\hat{\alpha}_i = D(\bar{y}_{i\cdot} - \bar{y}) = \frac{\sigma^2}{m}\frac{k-1}{k}.$$

Dispersiooni leidmisel on silmas peetud, et $\bar{y}_{i\cdot}$ ja $\bar{y}$ pole sõltumatud.

Nagu juba eespool öeldud, on dispersiooni σ^2(ehk mudeli jäägi) hinnanguks keskmine tasemesisene hajuvus

$$\hat{\sigma}^2 = \frac{SSW}{n-k}.$$

Nimetajas on n asemel $n-k$, selleks, et hinnang oleks nihketa. Kõike seda kasutades, saame välja kirjutada parameetrite usaldusvahemikud ja testida hüpoteese parameetrite kohta. Usaldusvahemikud α_i jaoks aitavad välja selgitada, missugused faktoritasemed suurendavad uuritava tunnuse keskmist, missugused vähendavad ja missuguste kohta me ei suuda konstateerida ei suurenemist ega vähenemist. Sama eesmärki täidab hüpoteesi

$H_0: \ \alpha_i = 0$ kontroll iga i korral olulisustõenäosuse abil. Usaldusvahemiku parameetrile σ^2 saame konstrueerida teadmise abil, et

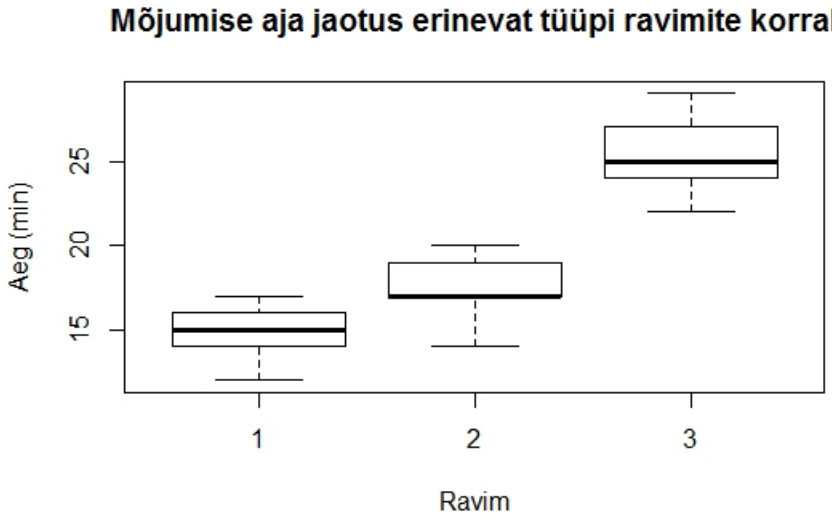
$$\frac{\hat{\sigma}^2(n-k)}{\sigma^2} = \frac{SSW}{\sigma^2} \sim \chi^2(n-k).$$

Näide 56 *Seljavalu ravimiseks on turu kolme tüüpi ravimeid valu leevendamiseks. Selleks, et välja selgitada nende efektiivsust valitakse juhuslikult 15 patsienti, kellele antakse üks ravimitest (iga ravimitüübi jaoks 5 juhuslikult valitud patsienti). Seejärel fikseeritakse aeg kuni mõju saabumiseni. Andmed on toodud järgmises tabelis:*

Ravimi tüüp	*1*	*1*	*1*	*1*	*1*	*2*	*2*	*2*	*2*	*2*	*3*	*3*	*3*	*3*	*3*
Aeg (min)	*12*	*15*	*16*	*17*	*14*	*14*	*17*	*19*	*20*	*17*	*25*	*27*	*29*	*24*	*22*

Soovime selgitada, kas aeg sõltub ravimi tüübist. Kuna seletav tunnus (ravimi tüüp) on mittearvuline (sest väljendab kõigest ravimi tüüpi), siis ei sobi siin uuritava tunnuse (Aeg) modelleerimiseks kasutada regressioonanalüüsi, vaid tuleb rakendada dispersioonanalüüsi.

Enne parameetrite hindamist on alati kasulik andmeid graafiliselt uurida. Järgmisel joonisel on toodud tunnuse 'Aeg (min)' jaotus ravimigruppides karpdiagrammide abil. Karbis asub 50% tunnuse 'Aeg (min)' väärtustest valimis, kusjuures joon karbi sees näitab valimi mediaani.



On näha, et 3. grupi mediaan ning terve karp asub kahe esimesega võrreldes ülalpool, mis viitab selle grupi erinevusele teistest. Ka esimest kaks paistavad mõnevõrra erinevad. Kas see erinevus on ka statistiliselt oluline? Selleks viime läbi dispersioonanalüüsi.

Järgmisel pildil on toodud vastavad R-käsud ja väljundid (tunnus 'Ravim' on fator tunnus). Arvutused on tehtud kitsenduse $\alpha_1 = 0$ korral. Sel juhul näitab vabaliige (intercept) 1. ravimitüübi keskmist aega (14,8 min). ANOVA tabelist saame F-statistiku väärtuse 28,086 ja olulisuse tõenäosuse (Pr(>F)) $2{,}975e-05 = 2{,}975 \cdot 10^{-5}$, mis ütleb, et tuleks vastu võtta sisukas hüpotees (aeg sõltub ravimi tüübist).

Kordajate tabelist näeme, et ravimtüübi 2 korral on keskmine aeg 2,6 minuti võrra pikem kui 1. tüübi korral. Kolmanda ravimi keskmine aeg on 10,6 minuti võrra esimesest pikem.

```
> mudel=lm(Min~Ravim, data=andmed)
> anova(mudel)
Analysis of Variance Table

Response: Min
          Df Sum Sq Mean Sq F value    Pr(>F)
Ravim      2  305.2 152.600  28.086 2.975e-05 ***
Residuals 12   65.2   5.433
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> mudel

Call:
lm(formula = Min ~ Ravim, data = andmed)

Coefficients:
(Intercept)       Ravim2       Ravim3
       14.8          2.6         10.6
```

8. 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

[5] E. B. Wilson (1927) Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 22: 209-212.

9. Lisa A. χ^2-jaotuse täiendkvantiilid

df\α	**0.1**	**0.05**	**0.01**	**df\α**	**0.1**	**0.05**	**0.01**
1	2.71	3.84	6.63	36	47.21	51.00	58.62
2	4.61	5.99	9.21	37	48.36	52.19	59.89
3	6.25	7.81	11.34	38	49.51	53.38	61.16
4	7.78	9.49	13.28	39	50.66	54.57	62.43
5	9.24	11.07	15.09	40	51.81	55.76	63.69
6	10.64	12.59	16.81	41	52.95	56.94	64.95
7	12.02	14.07	18.48	42	54.09	58.12	66.21
8	13.36	15.51	20.09	43	55.23	59.30	67.46
9	14.68	16.92	21.67	44	56.37	60.48	68.71
10	15.99	18.31	23.21	45	57.51	61.66	69.96
11	17.28	19.68	24.72	46	58.64	62.83	71.20
12	18.55	21.03	26.22	47	59.77	64.00	72.44
13	19.81	22.36	27.69	48	60.91	65.17	73.68
14	21.06	23.68	29.14	49	62.04	66.34	74.92
15	22.31	25.00	30.58	50	63.17	67.50	76.15
16	23.54	26.30	32.00	51	64.30	68.67	77.39
17	24.77	27.59	33.41	52	65.42	69.83	78.62
18	25.99	28.87	34.81	53	66.55	70.99	79.84
19	27.20	30.14	36.19	54	67.67	72.15	81.07
20	28.41	31.41	37.57	55	68.80	73.31	82.29
21	29.62	32.67	38.93	56	69.92	74.47	83.51
22	30.81	33.92	40.29	57	71.04	75.62	84.73
23	32.01	35.17	41.64	58	72.16	76.78	85.95
24	33.20	36.42	42.98	59	73.28	77.93	87.17
25	34.38	37.65	44.31	60	74.40	79.08	88.38
26	35.56	38.89	45.64	61	75.51	80.23	89.59
27	36.74	40.11	46.96	62	76.63	81.38	90.80
28	37.92	41.34	48.28	63	77.75	82.53	92.01
29	39.09	42.56	49.59	64	78.86	83.68	93.22
30	40.26	43.77	50.89	65	79.97	84.82	94.42
31	41.42	44.99	52.19	66	81.09	85.96	95.63
32	42.58	46.19	53.49	67	82.20	87.11	96.83
33	43.75	47.40	54.78	68	83.31	88.25	98.03
34	44.90	48.60	56.06	69	84.42	89.39	99.23
35	46.06	49.80	57.34	70	85.53	90.53	100.43

10. Lisa B. t-jaotuse täiendkvantiilid

df\α	**0.2**	**0.15**	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**
1	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	0.978	1.250	1.638	2.353	3.182	4.541	5.841
4	0.941	1.190	1.533	2.132	2.776	3.747	4.604
5	0.920	1.156	1.476	2.015	2.571	3.365	4.032
6	0.906	1.134	1.440	1.943	2.447	3.143	3.707
7	0.896	1.119	1.415	1.895	2.365	2.998	3.499
8	0.889	1.108	1.397	1.860	2.306	2.896	3.355
9	0.883	1.100	1.383	1.833	2.262	2.821	3.250
10	0.879	1.093	1.372	1.812	2.228	2.764	3.169
11	0.876	1.088	1.363	1.796	2.201	2.718	3.106
12	0.873	1.083	1.356	1.782	2.179	2.681	3.055
13	0.870	1.079	1.350	1.771	2.160	2.650	3.012
14	0.868	1.076	1.345	1.761	2.145	2.624	2.977
15	0.866	1.074	1.341	1.753	2.131	2.602	2.947
16	0.865	1.071	1.337	1.746	2.120	2.583	2.921
17	0.863	1.069	1.333	1.740	2.110	2.567	2.898
18	0.862	1.067	1.330	1.734	2.101	2.552	2.878
19	0.861	1.066	1.328	1.729	2.093	2.539	2.861
20	0.860	1.064	1.325	1.725	2.086	2.528	2.845
21	0.859	1.063	1.323	1.721	2.080	2.518	2.831
22	0.858	1.061	1.321	1.717	2.074	2.508	2.819
23	0.858	1.060	1.319	1.714	2.069	2.500	2.807
24	0.857	1.059	1.318	1.711	2.064	2.492	2.797
25	0.856	1.058	1.316	1.708	2.060	2.485	2.787
26	0.856	1.058	1.315	1.706	2.056	2.479	2.779
27	0.855	1.057	1.314	1.703	2.052	2.473	2.771
28	0.855	1.056	1.313	1.701	2.048	2.467	2.763
29	0.854	1.055	1.311	1.699	2.045	2.462	2.756
30	0.854	1.055	1.310	1.697	2.042	2.457	2.750
10000	0.842	1.036	1.282	1.645	1.960	2.327	2.576