Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Matemaatika ja statistika instituut

Liis Simmul

# Pidevate jaotuste parameetrite hindamisest suurima tõepära ja suurimate vahemike meetodil

Matemaatilise statistika eriala

Bakalaureusetöö (9 EAP)

Juhendaja: vanemteadur Kristi Kuljus

Tartu 2019

**Pidevate jaotuste parameetrite hindamisest suurima tõepära ja suurimate vahemike meetodil**

Bakalaureusetöö

Liis Simmul

**Lühikokkuvõte.** Iga statistilise probleemi üks osa on valimi põhjal kogu populatsiooni kohta kehtivate üldistuste tegemine. Mõnikord on probleemi lahendamiseks vaja hinnata jaotuse parameetreid. Käesolevas bakalaureusetöös on vaadeldud Kullback-Leibleri informatsioonimõõdu lähendamisel põhinevaid kahte pidevate jaotuste parameetrite hindamise meetodit – suurima tõepära ja suurimate vahemike meetodit. Muu hulgas uuritakse juhte, kus suurima tõepära meetod ei tööta ja vajab seetõttu alternatiivi. Näiteid tuuakse meetodite käitumisest, kus uuritavaks jaotuseks on normaaljaotus, ühtlane jaotus või Weibulli jaotus.

**CERCS teaduseriala:** P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kintdlustusmatemaatika

**Märksõnad:** tõenäosusjaotused, statistilised mudelid, parameetrilised meetodid, optimeerimine, simulatsioon

**Estimating parameters of continuous distributions using maximum likelihood and maximum spacing method**

Bachelor's thesis

Liis Simmul

**Abstract.** Drawing inferences from a sample that apply on the whole population is a part of every statistical problem. To solve the problem it is sometimes necessary to estimate the parameters of the distribution. This bachelor's thesis studies two parameter estimation methods, which are both based on approximating the Kullback-Leibler information measure – maximum likelihood and maximum spacing method. We consider cases where maximum likelihood does not work and therefore an alternative method is needed. Examples are provided on the behaviour of the methods, where the studied distribution is normal distribution, univariate or Weibull distribution.

**CERCS research specialisation:** P160 Statistics, operations research, programming, financial and actuarial mathematics

**Keywords:** probability distributions, statistical models, parametric methods, optimization, simulation

# Sisukord

# Sissejuhatus

Iga statistiliste probleemi korral on antud mingi hulk vaatlusi. Nende vaatluste näol on tegemist mingi juhusliku suuruse realisatsioonidega, mis pärinevad mingist jaotusest. Vastamaks vaatlustega seonduvatele küsimustele, on mõnikord tarvis leida hinnangud selle jaotuse parameetritele. Kui oleme loonud eelduse, millisest jaotusest valim võiks pärineda, on parameetritele hinnangute leidmiseks vaja rakendada mõnda parameetrite hindamise meetodit. Soovime valimi abil anda võimalikult täpse hinnangu ehk leida parameetrid nii, et tegeliku jaotuse ja valimi abil hinnatud jaotuse (meie mudel) vaheline keskmine erinevus oleks võimalikult väike.

Üheks jaotustevahelise keskmise erinevuse mõõduks on Kullback-Leibleri informatsioonimõõt. Käesoleva bakalaureusetöö eesmärk on Kullback-Leibleri informatsioonimõõdust tuletada kaks pidevate jaotuste parameetrite hindamise meetodit ning nende meetodite abil leitud hinnangute omadusi uurida. Üheks meetoditest on statistikas enim kasutust leidev, kuid siiski mõningate puudustega meetod – suurima tõepära (*maximum likelihood*) meetod. Teiseks on suurimate vahemike (*maximum spacing*) meetod, mis on küll keerukam, kuid suurima tõepära meetodi mittetöötamise juhtudel vajalik alternatiiv. Vaatluse all on ühemõõtmelised pidevad jaotused.

Töö esimeses osas tuuakse välja Kullback-Leibleri kaugusmõõdu omadused ja parameetrite hindamise meetodite intuitiivne tuletuskäik. Seal tuuakse näiteks juht, mil suurima tõepära meetod ei tööta. Vaadeldakse ka olukorda, mil mõlemad meetodid töötavad, andes seejuures asümptootiliselt võrdseid hinnanguid. Töö teises osas uuritakse simulatsiooninäidete abil juhte, kus suurima tõepära meetodiga hinnangute leidmine on ühel juhul raskendatud ja teisel juhul võimatu, ning näidatakse, et suurimate vahemike meetod töötab mõlemal juhul hästi. Lisaks eeltoodule uuritakse meetodite abil saadud hinnangute omadusi.

# 1 Pidevate jaotuste parameetrite hindamine

Olgu meil antud valim, mis esindab mingit jaotuste klassi. Soovime valimi abil hinnata selle jaotuse parameetreid. Üheks võimaluseks on minimeerida eeldatava ja tegeliku jaotuse vaheline keskmine erinevus kasutades Kullback-Leibleri informatsioonimõõtu. Nii leiame parameetrid selliselt, et eeldatava ja tegeliku jaotuse vaheline keskmine erinevus on vähim. Esimeses peatükis uurime, millised omadused on Kullback-Leibleri informatsioonimõõdul ning kuidas selle lähendamisel tuletada kaks parameetrite hindamise meetodit.

## 1.1 Kullback-Leibleri informatsioonimõõt

Kullback-Leibleri informatsioonimõõdu definitsioon põhineb allikatel Lember (2018) ja Ranneby (1984).

**Definitsioon 1.** *Vaatleme juhuslikku suurust $X$, mis võtab väärtusi hulgal $\chi = \{x_1, x_2, ...\}$. Olgu $P$ ja $Q$ kaks diskreetset jaotust tõenäosustega $p_i = P(x_i)$ ja $q_i = Q(x_i)$ $\forall x_i \in \chi$ korral, kusjuures $P$ on juhusliku suuruse $X$ tegelik jaotus. Kullback-Leibleri informatsioonimõõt jaotuste $P$ ja $Q$ vahel on defineeritud järgmiselt*

$$KL(P||Q) := \sum_i p_i \ln \frac{p_i}{q_i}.$$

*Seejuures $q_i \geq 0$ korral defineeritakse $0 \cdot \ln(\frac{0}{q_i}) = 0$ ja $p_i > 0$ korral $p_i \cdot \ln(\frac{p_i}{0}) = \infty$.*
*Kui funktsioonid $p(x)$ ja $q(x)$ on pidevate jaotuste $P$ ja $Q$ tihedusfunktsioonid, mis on määratud hulgal $I = (a, b)$, $-\infty \leq a < b \leq \infty$, siis Kullback-Leibleri kaugus avaldub kujul*

$$KL(p||q) = \int p(x) \ln \frac{p(x)}{q(x)} dx = E\Big( \ln \frac{p(X)}{q(X)} \Big), \tag{1}$$

*eeldusel, et juhusliku suuruse $X$ tegelik jaotus on $P$.*

Kullback-Leibleri informatsioonimõõt on tuntud ka nime all Kullback-Leibleri kaugus, kuid tegemist ei ole klassikalise kaugusega. Seda näitab selle suuruse esimene omadus, milleks on mittesümmeetrilisus. Selle selgitamiseks uurime kahte näidet kahe diskreetse jaotuse KL-kauguse kohta.

Esiteks defineerime kaks tõenäosusjaotust $P$ ja $Q$ hulgal $\{x_1, x_2, x_3\}$, mille väärtused on toodud tabelis 1.

Tabel 1. Juhusliku suuruse $X$ jaotused

| | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| $P(x_i)$ | 0.2 | 0.5 | 0.3 |
| $Q(x_i)$ | 0.1 | 0.7 | 0.2 |

Kullback-Leibleri informatsioonimõõt jaotuste $P$ ja $Q$ vahel on

$$KL(P||Q) = 0.2 \ln \frac{0.2}{0.1} + 0.5 \ln \frac{0.5}{0.7} + 0.3 \ln \frac{0.3}{0.2} \approx 0.092$$

ning jaotuste $Q$ ja $P$ vahel on

$$KL(Q||P) = 0.1 \ln \frac{0.1}{0.2} + 0.7 \ln \frac{0.7}{0.5} + 0.2 \ln \frac{0.2}{0.3} \approx 0.085.$$

Näeme, et $KL(P||Q) \neq KL(Q||P)$.

Teiseks uurime näidet, mille idee pärineb informatsiooniteooria loengukonspektist (Lember, 2018). Tähistagu $P$ juhusliku suuruse $X$ tegelikku jaotust ning eeldame, et seda jaotust on sobiv lähendada jaotusega $Q$. Olgu teatud sündmuse esinemise tõenäosus positiivne ehk $p_i > 0$ mingi $i$ korral. Meie aga arvame, et seda sündmust esineda ei saa ehk $q_i = 0$. Liikme $i$ panus Kullback-Leibleri kaugusesse on

$$p_i \ln \frac{p_i}{0} = \infty.$$

See tähendab, et positiivse tõenäosusega sündmuse võimatuks tunnistamine suurendab jaotustevahelist keskmist erinevust lõpmata suurel määral. Vastupidisel juhul, kui vahetame eeldatava ja tegeliku jaotuse rollid ehk eeldame, et $Q$ on juhusliku suuruse tegelik jaotus ning $P$ meie mudel, on liikme $i$ panus KL-kaugusesse

$$0 \ln \frac{0}{p_i} = 0.$$

Siit järeldub, et tegelikult mittetoimuva sündmuse võimalikuks arvamine jaotustevahelist keskmist erinevust ei suurenda. Vaadeldud sündmuste panus KL-kaugusesse on erinev. Seega illustreerib viimane näide samuti KL-kauguse mittesümmeetrilisust.

Kullback-Leibleri kauguse teist omadust, et selle väärtus on alati mittenegatiivne, selgitame järgnevalt. See arutelu on mõistmaks, et KL-kauguse väärtust saab interpreteerida kui kahe funktsiooni $p(x)$ ja $q(x)$ keskmise erinevuse mõõtu ning põhineb allikatel Lehmann ja Casella (1998) ning Bishop (2006).

**Definitsioon 2.** *Funktsiooni $\phi(x)$ kutsutakse kumeraks, kui mistahes $0 < \lambda < 1$ ja mistahes väärtuste $x < y$ korral funktsiooni määramispiirkonnast $(a, b)$, kus $-\infty \leq a < b \leq \infty$, kehtib*

$$\phi(\lambda x + (1 - \lambda)y) \leq \lambda\phi(x) + (1 - \lambda)\phi(y).$$

*Funktsiooni nimetatakse rangelt kumeraks, kui kehtib range võrratus.*

**Lemma 1.** ***Jenseni võrratus.*** *Olgu $X$ lõpliku keskväärtusega juhuslik suurus, mis võtab väärtusi hulgal $I = (a, b)$, $-\infty \leq a < b \leq \infty$. Kumera funktsiooni $\phi(x)$, mis on samuti defineeritud hulgal $I$, korral kehtib võrratus*

$$\phi[E(X)] \leq E[\phi(X)].$$

*Kui $\phi$ on rangelt kumer, kehtib range võrratus, välja arvatud juhul, kui $X$ on konstantne tõenäosusega 1.*

***Märkus.*** *Pidevate jaotuste juhul on Jenseni võrratus kujul*

$$\phi\Big(\int xp(x)dx\Big) \leq \int \phi(x)p(x)dx,$$

*kus $p(x)$ on juhusliku suuruse $X$ tihedusfunktsioon.*

**Järeldus.** *Kullback-Leibleri informatsioonimõõt on mittenegatiivne, $KL(p||q) \geq 0$.*

*Tõestus.* Kasutades teadmist, et funktsioon $-\ln(x)$ on kumer funktsioon ja et $\int q(x)dx = 1$, saame näidata, et KL-kauguse väärtus on alati mittenegatiivne. Jenseni võrratuse abil järeldub, et

$$\begin{aligned} KL(p||q) = \int p(x) \ln \frac{p(x)}{q(x)} dx = -\int p(x) \ln \frac{q(x)}{p(x)} dx \geq \\ \geq -\ln \int p(x) \frac{q(x)}{p(x)} dx = -\ln \int q(x)dx = 0. \end{aligned}$$

Funktsioon $-\ln(x)$ ei ole mitte ainult kumer, vaid rangelt kumer funktsioon. Seega kehtib võrdus ainult juhul, kui $\frac{q(x)}{p(x)} = 1$ (vt lemma 1) ehk kui funktsioonid $q(x)$ ja $p(x)$ on võrdsed. See tähendab, et funktsioonide $q(x)$ ja $p(x)$ võrdsuse korral on KL-kauguse väärtuseks 0, muudel juhtudel on selle väärtus nullist suurem.

■

Kui funktsioonide $p(x)$ ja $q(x)$ näol on tegemist juhuslike suuruste tihedusfunktsioonidega, iseloomustab KL-kaugus kahe jaotuse vahelist keskmist erinevust ehk seda, kui hästi suudab ühest jaotusest pärineva juhusliku suuruse jaotust teine jaotus kirjeldada. Eksitav on mõista seda suurust kui jaotustevahelist „kaugust". Nii võib tunduda, et mõõdame jaotustevahelist nihet, kuid nii see ei ole. Otsime sobivat jaotust (mudel), mis kirjeldaks tundmatut jaotust, millest pärineb valim, kõige paremini.

**Näide 1.** Uurime Kullback-Leibleri kauguse käitumist kahe normaaljaotuse $\mathcal{N}(\mu_1, \sigma_1^2)$ ja $\mathcal{N}(\mu_2, \sigma_2^2)$ korral. Jaotuste tihedused on vastavalt $p(x)$ ja $q(x)$, kusjuures juhusliku suuruse $X$ tegelikuks jaotuseks on $\mathcal{N}(\mu_1, \sigma_1^2)$. Kullback-Leibleri kaugus avaldub järgmiselt:

$$
\begin{aligned}
KL(p||q) = E\Big(\ln\frac{p(X)}{q(X)}\Big) = E\Big(\ln\Big(\frac{1}{\sqrt{2\pi}\sigma_1}e^{\frac{-(X-\mu_1)^2}{2\sigma_1^2}}\Big) - \ln\Big(\frac{1}{\sqrt{2\pi}\sigma_2}e^{\frac{-(X-\mu_2)^2}{2\sigma_2^2}}\Big)\Big) = \\
E\Big(\ln\frac{1}{\sqrt{2\pi}\sigma_1} - \frac{(X-\mu_1)^2}{2\sigma_1^2} - \ln\frac{1}{\sqrt{2\pi}\sigma_2} + \frac{(X-\mu_2)^2}{2\sigma_2^2}\Big) = \\
\ln\frac{1}{\sqrt{2\pi}\sigma_1} - E\Big(\frac{(X-\mu_1)^2}{2\sigma_1^2}\Big) - \ln\frac{1}{\sqrt{2\pi}\sigma_2} + E\Big(\frac{(X-\mu_2)^2}{2\sigma_2^2}\Big) \overset{E(X-\mu_1)=\sigma_1^2}{=} \\
\ln\sigma_2 - \ln\sigma_1 - \frac{1}{2} + E\Big(\frac{(X-\mu_2)^2}{2\sigma_2^2}\Big) = \\
\ln\sigma_2 - \ln\sigma_1 - \frac{1}{2} + \frac{1}{2\sigma_2^2}E(X^2 - 2X\mu_2 + \mu_2^2) \overset{E(X^2)=\sigma_1^2+\mu_1^2}{=} \\
\ln\frac{\sigma_2}{\sigma_1} - \frac{1}{2} + \frac{1}{2\sigma_2^2}\Big(\sigma_1^2 + \mu_1^2 - 2\mu_2\mu_1 + \mu_2^2\Big) = \\
\ln\frac{\sigma_2}{\sigma_1} + \frac{1}{2}\Big(-1 + \frac{\sigma_1^2}{\sigma_2^2} + \frac{(\mu_1-\mu_2)^2}{\sigma_2^2}\Big).
\end{aligned}
$$

Illustreerimaks keskväärtuse ja dispersiooni erinevat mõju KL-kauguse väärtusele, on tabelites 2 ja 3 toodud mõned näited. Oletame esmalt, et jaotuste keskväärtused on võrdsed, $\mu_1 = \mu_2$. Siis $KL(p||q) = \ln\frac{\sigma_2}{\sigma_1} - \frac{1}{2} + \frac{\sigma_1^2}{2\sigma_2^2}$.

Tabel 2 kirjeldab KL-kauguse muutumist, kui sellisel juhul väärtustame $\sigma_2$ võrdeliselt $\sigma_1$-ga.

Tabel 2. KL-kaugus kahe normaaljaotuse korral, kus $\mu_1 = \mu_2$

| $\sigma_2$ | $2\sigma_1$ | $4\sigma_1$ | $8\sigma_1$ | $1000\sigma_1$ |
|---|---|---|---|---|
| $KL(p\|\|q)$ | 0.318 | 0.918 | 1.587 | 6.408 |

Oletame nüüd, et $\sigma_1 = \sigma_2 = 1$. Siis $KL(p||q) = \frac{(\mu_1-\mu_2)^2}{2}$. Vaatleme, kuidas muutub Kullback-Leibleri kauguse väärtus suurendades jaotuste keskväärtuste vahet (tabel 3).

Tabel 3. KL-kaugus kahe normaaljaotuse korral, kus $\sigma_1 = \sigma_2 = 1$

| $\mu_1 - \mu_2$ | 2 | 4 | 8 | 1000 |
|---|---|---|---|---|
| $KL(p\|\|q)$ | 2 | 8 | 32 | 500000 |

Oleme kahe jaotuse vahelise keskmise kauguse mõõtmiseks defineerinud Kullback-Leibleri informatsioonimõõdu (1) ning uurinud selle omadusi. Kui minimeerime seda suurust, saame leida hinnangud mudeli jaotuse parameetritele nii, et erinevus tegeliku ja oletatava jaotuse vahel on vähim. Pärast hinnangu mõiste ja omaduste defineerimist näitame, et KL-kaugust kahel erineval viisil minimeerides on võimalik jõuda kahe erineva parameetrite hindamise meetodini.

Hinnangu mõiste ja omaduste defineerimiseks on kasutatud allikaid Lepik (2017) ning Dudewicz ja Mishra (1988). Olgu meil antud jaotus $F$, mis sõltub parameetrist $\theta$. Parameetrile $\theta$ antud hinnangu $\hat{\theta}_n \in \Theta$ all mõtleme hinnangufunktsiooni $\hat{\theta}_n = \hat{\theta}_n(X_1, ..., X_n)$, kus $X_i \sim F$.

Hinnangute uurimisel huvitab meid näiteks see, milline on hinnangu nihe ehk keskmine kõrvalekalle parameetri tegelikust väärtusest. Vaadeldakse veel ka hinnangu keskmist ruutviga.

**Definitsioon 3.** *Jaotuse $F$ parameetrile $\theta$ leitud hinnangu $\hat{\theta}_n$ keskmiseks ruutveaks ja nihkeks nimetatakse vastavalt keskväärtusi*

$$E(\hat{\theta}_n - \theta)^2, \quad E(\hat{\theta}_n - \theta). \tag{2}$$

Lisaks sellele huvitavad meid hinnangute asümptootilised omadused: kas hinnangute varieeruvus valimimahu kasvades muutub lõpmata väikeseks ning kas suurte valimimahtmahtude korral parameetrile antud hinnangute keskväärtus koondub parameetri tegelikuks väärtuseks. Teisisõnu uurime, kas $D(\hat{\theta}_n) \overset{n\to\infty}{\longrightarrow} 0$ ja $E(\hat{\theta}_n - \theta) \overset{n\to\infty}{\longrightarrow} 0$ (hinnang on nihketa). Need tingimused on hinnangu mõjususe alternatiivsed esitused.

**Definitsioon 4.** *Öeldakse, et tegelikule parameetrile $\theta$ leitud hinnang $\hat{\theta}_n$ on mõjus, kui $\forall \epsilon > 0$ korral*

$$P\Big(|\hat{\theta}_n - \theta| > \epsilon\Big) \overset{n\to\infty}{\longrightarrow} 0, \quad \forall \theta \in \Theta.$$

Kui hinnangu $\hat{\theta}_n$ keskmine ruutviga $E(\hat{\theta}_n - \theta)^2$ läheneb valimimahu kasvades nullile, järeldub sellest hinnangu mõjusus,

$$E(\hat{\theta}_n - \theta)^2 \overset{n\to\infty}{\longrightarrow} 0 \Longrightarrow P\Big(|\hat{\theta}_n - \theta| > \epsilon\Big) \overset{n\to\infty}{\longrightarrow} 0. \tag{3}$$

## 1.2 Suurima tõepära meetod

Käesoleva alapeatüki tuletuskäik põhineb artiklil Ranneby (1984), kui ei ole märgitud teisiti.

Olgu $x_1, ..., x_n$ valim juhuslikust suurusest $X$, mille tihedus on $g(x)$. Olgu tihedusfunktsioonid $\{f_\theta(x); \theta \in \Theta\}$, kus $\Theta \subset R^k$ sellised, mis meie arvates võiksid kirjeldada juhusliku suuruse $X$ jaotust.

**Definitsioon 5.** *Pideva juhusliku suuruse $X$, mille jaotust kirjeldab tihedusfunktsioon $f_\theta(x)$, tõepärafunktsiooniks nimetatakse avaldist*

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i)$$

*ning logaritmiliseks tõepärafunktsiooniks avaldist*

$$l(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln f_\theta(x_i).$$

*Parameetri väärtust $\hat{\theta}_n$, mille korral $l(\theta)$ saavutab maksimumi, nimetatakse* ***suurima tõepära (maximum likelihood) hinnanguks****.*

KL-kaugus (1) avaldub järgmiselt

$$KL(g||f_\theta) = \int g(x) \ln g(x) dx - \int g(x) \ln f_\theta(x) dx = E\Big( \ln g(X) \Big) - E\Big( \ln f_\theta(X) \Big).$$

Kuna keskväärtust saab hinnata valimi keskmise abil, saame kirjutada, et

$$KL(g||f_\theta) \approx \frac{1}{n} \sum_{i=1}^{n} \ln g(x_i) - \frac{1}{n} \sum_{i=1}^{n} \ln f_\theta(x_i). \tag{4}$$

Nagu varasemalt selgitasime, on mudeli parameetrite leidmiseks loomulik minimeerida KL-kaugus. See on aga samaväärne avaldises (4) oleva liikme $\frac{1}{n} \sum_{i=1}^{n} \ln f_\theta(x_i)$ maksimeerimisega, mis omakorda on ekvivalentne logaritmilise tõepärafunktsiooni maksimeerimisega. Seega, et

$$\frac{1}{n} \sum_{i=1}^{n} \ln f_\theta(x_i) = \frac{1}{n} \ln L(\theta),$$

olemegi jõudnud KL-kauguse lähendamisel suurima tõepära meetodini.

**Näide 2.** Suurima tõepära meetod töötab hästi, kui tõepärafunktsiooni (või logaritmilise tõepärafunktsiooni) iga liige on ülalt tõkestatud. Näiteks normaaljaotuste segu korral võib parameetrite hindamine olla problemaatiline, sest tõepärafunktsioon on ülalt tõkestamata. Järgnevas näites, milles on toetutud allikale Bishop (2006), näitame kahekomponendilise normaaljaotuste segu korral, et tõepärafunktsioon on tõkestamata. Kahe normaaljaotuse segu tihedus avaldub kujul

$$f_\theta(x) = \lambda \, f(x; \mu_1, \sigma_1^2) + (1 - \lambda) \, f(x; \mu_2, \sigma_2^2),$$

kus $f$ on normaaljaotuse tihedus, $\theta = (\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2)$ ja $0 < \lambda < 1$ määrab komponentide kaalu.

Tõepärafunktsioon on seega kujul

$$L(\theta) = \prod_{i=1}^{n} \left[ \lambda \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x_i - \mu_1)^2}{2\sigma_1^2}} + (1 - \lambda) \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x_i - \mu_2)^2}{2\sigma_2^2}} \right].$$

Uurime, mis juhtub, kui kahekomponendilise normaaljaotuste segu korral üks valimi väärtustest langeb kokku esimese komponendi keskväärtusega. Oletame üldisust kitsendamata, et $\mu_1 = x_1$.

Siis tõepärafunktsioon, kus juhime tähelepanu esimesele liikmele, on kujul

$$L(\theta) = \left[\lambda \frac{1}{\sqrt{2\pi}\sigma_1} + (1-\lambda)\frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x_1-\mu_2)^2}{2\sigma_2^2}}\right] \cdot$$
$$\cdot \prod_{i=2}^{n} \left[\lambda \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}} + (1-\lambda)\frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}}\right].$$

Vaatleme olukorda, kus $\sigma_1 \to 0$. Näeme, et siis $\lambda \frac{1}{\sqrt{2\pi}\sigma_1} \to \infty$, mis põhjustab kogu tõepärafunktsiooni tõkestamatu kasvamise. Teised parameetrit $\sigma_1$ sisaldavad liikmed on lõplikud, sest kui $\sigma_1 \to 0$, siis $\frac{1}{\sigma_1} / e^{\frac{(x_i-\mu_1)^2}{2\sigma_1^2}} \to 0$, $i = 2, ..., n$ korral.

Olgu mainitud, et normaaljaotuse $\mathcal{N}(\mu, \sigma^2)$ korral sellist probleemi ei teki. Siis avaldub normaaljaotuse tõepärafunktsioon kujul

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

Kui oletame, et $\mu = x_1$, siis saame tõepärafunktsiooni kujule

$$L(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_2-\mu)^2}{2\sigma^2}} \cdot \prod_{i=3}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{2\pi\sigma^2} e^{-\frac{(x_2-\mu)^2}{2\sigma^2}} \cdot \prod_{i=3}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}.$$

Näeme, et vaadeldavas protsessis ei kasva ükski liige tõkestamatult: kui $\sigma \to 0$, siis $\frac{1}{\sigma^2} / e^{\frac{(x_2-\mu)^2}{2\sigma^2}} \to 0$ ja $\frac{1}{\sigma} / e^{\frac{(x_i-\mu)^2}{2\sigma^2}} \to 0$, $i = 3, ..., n$ korral. Sellepärast ei esine suurima tõepära meetodiga normaaljaotuse $\mathcal{N}(\mu, \sigma^2)$ parameetrite hindamisel probleeme.

Kui tõepärafunktsioon on tõkestamata, ei leidu funktsioonil globaalset maksimumi. Seega ei leidu ka globaalse maksimumi kaudu defineeritud hinnanguid ja meetod ei tööta. Mõnikord defineeritakse suurima tõepära hinnangud tõepärafunktsiooni tuletise kaudu, mis võrdsustatakse nulliga (Cheng ja Amin, 1983). See võimaldab lahenditena vaadelda ka lokaalseid maksimume. Teine võimalus on hinnangud leida mõne muu meetodi abil.

Nagu öeldud, on Kullback-Leibleri kaugust võimalik lähendada veel teiselgi viisil. Vaatleme järgmiseks lähendust, mis annab tulemuseks suurimate vahemike meetodi.

## 1.3 Suurimate vahemike meetod

Järgnevas alapeatükis on toetutud artiklitele Ranneby (1984) ning Cheng ja Amin (1983).

Vaatleme pidevat juhuslikku suurust $X$, mis võtab väärtusi hulgal $I = (a, b)$, $-\infty \leq a < b \leq \infty$. Olgu selle juhusliku suuruse tihedus $g(x)$ ja jaotusfunktsioon $G(x)$. Vaatleme lisaks ka tihedusfunktsioone $\{f_\theta(x); \theta \in \Theta\}$ vastavate jaotusfunktsioonidega $F_\theta(x)$. Sarnaselt eelmise meetodi tuletuskäigus kasutatud tähistustele, olgu $F_\theta(x)$ jaotuste klass, mille arvame sobivat kirjeldama juhusliku suuruse $X$ tegelikku jaotusfunktsiooni, milleks, nagu öeldud, on $G(x)$.

Järjestades juhusliku suuruse $X$ realisatsioonid $x_1, ..., x_n$ ning lisades valimi otspunktidesse väärtused $a$ ja $b$, saame

$$a = x_{(0)} < x_{(1)} < ... < x_{(n)} < x_{(n+1)} = b.$$

**Definitsioon 6.** *Olgu vaatluse all eeltoodud järjestatud valim lisatud otspunktidega. Järjestikustele valimi väärtustele vastavate jaotusfunktsiooni väärtuste vahesid nimetatakse vahemikeks,*

$$D_j(x_j, x_{j-1}) = F_\theta(x_{(j)}) - F_\theta(x_{(j-1)}) = \int_{x_{j-1}}^{x_j} f_\theta(x)dx, \quad j = 1, ..., n+1.$$

*Vahemike funktsioon $S(\theta)$ on defineeritud järgmiselt,*

$$S(\theta) = \frac{1}{n+1} \sum_{j=1}^{n+1} \ln D_j = \frac{1}{n+1} \sum_{j=1}^{n+1} \ln \Big( F_\theta(x_{(j)}) - F_\theta(x_{(j-1)}) \Big).$$

***Märkus.*** Väärtuste $x_{(0)} = a$ ja $x_{(n+1)} = b$ korral

$$F(x_{(0)}) = 0 \text{ ja } F(x_{(n+1)}) = 1.$$

Uurime, kuidas nimetatud jaotuste vahelise KL-kauguse minimeerimise kaudu jõuda suurimate vahemike hinnanguteni. Selleks tuletame meelde ühe abitulemuse.

**Lemma 1.** ***Lagrange'i keskväärtusteoreem.*** *Pideva funktsiooni $f : [a, b] \to \mathbb{R}$ korral, mis vahemikus $(a, b)$ on diferentseeruv, leidub selline $c \in (a, b)$, et*

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Järgnevalt, kasutades Lagrange'i kesväärtusteoreemi ning teadmist, et tihedus on jaotusfunktsiooni tuletis, saame leida väärtused $\overline{x}_{(j)}, \overline{\overline{x}}_{(j)} \in (x_{(j-1)}, x_{(j)})$, $j = 1, ..., n+1$, nii et

$$F_\theta(x_{(j)}) - F_\theta(x_{(j-1)}) = f_\theta(\overline{x}_{(j)}) \cdot (x_{(j)} - x_{(j-1)}),$$

$$G(x_{(j)}) - G(x_{(j-1)}) = g(\overline{\overline{x}}_{(j)}) \cdot (x_{(j)} - x_{(j-1)}). \tag{5}$$

Teame, et suuruse (4) saab välja kirjutada järgmiselt:

$$\frac{1}{n}\sum_{i=1}^{n} \ln g(x_i) - \frac{1}{n}\sum_{i=1}^{n} \ln f_\theta(x_i) = \frac{1}{n}\sum_{i=1}^{n} \ln \frac{g(x_i)}{f_\theta(x_i)}$$

ning et see koondub suuruseks $KL(g||f_\theta)$.
Kasutades Lagrange'i keskväärtusteoreemi abil saadud avaldisi (5), on intuitiivselt selge, et ka avaldis

$$\frac{1}{n+1}\sum_{j=1}^{n+1} \ln \frac{G(x_{(j)}) - G(x_{(j-1)})}{F_\theta(x_{(j)}) - F_\theta(x_{(j-1)})} \tag{6}$$

koondub suuruseks $KL(g||f_\theta)$.

Nüüd, nagu ka eespool, muutmaks tegeliku ja eeldatava jaotuse vahelist erinevust vähimaks, minimeerime KL-kauguse. Nii leiame sellised hinnangud tundmatu jaotuse parameetritele, et vahe tegeliku jaotusega on minimaalne. Avaldise (6) minimeerimine on ekvivalentne avaldise

$$\frac{1}{n+1}\sum_{j=1}^{n+1} \ln \Big( F_\theta(x_{(j)}) - F_\theta(x_{(j-1)}) \Big)$$

maksimeerimisega. See ongi aga vahemike funktsioon $S(\theta)$.

**Definitsioon 7.** *Parameetri väärtust $\hat{\theta}_n$, mille korral $S(\theta)$ saavutab maksimumi, nimetatakse* ***suurimate vahemike (maximum spacing) hinnanguks.***

On oluline tähele panna, et funktsioon $S(\theta)$ on ülevalt tõkestatud. Näitame, et selleks tõkkeks on suurus $-\ln(n+1)$.

Esiteks on vahemike summa alati 1,

$$\sum_{j=1}^{n+1} D_j = \sum_{j=1}^{n+1} \int_{x_{(j-1)}}^{x_{(j)}} f_\theta(x)dx = \int_{x_{(0)}}^{x_{(1)}} f_\theta(x)dx + \int_{x_{(1)}}^{x_{(2)}} f_\theta(x)dx + ...$$
$$... + \int_{x_{(n)}}^{x_{(n+1)}} f_\theta(x)dx \overset{\text{Määratud integraali aditiivsus}}{=}$$
$$= \int_{x_{(0)}}^{x_{(n+1)}} f_\theta(x)dx = F(x_{(n+1)}) - F(x_{(0)}) = 1.$$

Teiseks teame, et geomeetriline keskmine on alati väiksem või võrdne aritmeetilisest keskmisest, ehk

$$\Big(\prod_{j=1}^{n+1} D_j\Big)^{\frac{1}{n+1}} \leq \frac{1}{n+1}\sum_{j=1}^{n+1} D_j.$$

Logaritmides eelneva avaldise mõlemaid pooli, saamegi tõkke vahemiku funktsioonile,

$$\frac{1}{n+1}\sum_{j=1}^{n+1} \ln D_j \leq \ln\Big(\frac{1}{n+1}\sum_{j=1}^{n+1} D_j\Big) = \ln\frac{1}{n+1}.$$

**Näide 3.** Uurime suurima tõepära ja suurimate vahemike hinnangute erinevust ühtlase jaotuse korral. Olgu $X$ ühtlase jaotusega juhuslik suurus, $X \sim \mathcal{U}(a,b)$ ning olgu $x_1, x_2, ..., x_n$ selle juhusliku suuruse juba järjestatud realisatsioonid. Ühtlase jaotusega juhusliku suuruse tihedusfunktsioon on

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b] \\ 0, & \text{mujal.} \end{cases}$$

Viimasest saame ka logaritmilise tõepärafunktsiooni,

$$l(a,b) = \sum_{i=1}^{n} \ln f(x_i) = n \cdot \ln\Big(\frac{1}{b-a}\Big) = -n \cdot ln(b-a).$$

Selleks, et leida suurima tõepära hinnangud otsitavatele parameetritele $a$ ja $b$, on vaja maksimeerida tõepärafunktsioon. Funktsioon $-ln(x)$ on rangelt monotoonselt kahanev funktsioon, millest järeldub see, et funktsiooni väärtus saavutab maksimumi minimaalse argumendi väärtuse korral. Järelikult peab argument $b-a$ olema minimaalne, et tõepärafunktsiooni väärtus saaks olla maksimaalne, arvestades ka seda, et kõik valimi väärtused peavad mahtuma otspunktide vahelise vahemiku sisse. Kokkuvõtlikult järeldub

viimasest arutelust, et suurima tõepära hinnanguteks on $\hat{a} = x_1$ ja $\hat{b} = x_n$. Järgmiseks leiame suurimate vahemike hinnangud ning defineerime selleks suurimate vahemike funktsiooni ühtlase jaotuse jaoks,

$$S(a,b) = \frac{1}{n+1}\sum_{j=1}^{n+1} \ln\Big(F_{a,b}(x_j) - F_{a,b}(x_{j-1})\Big) =$$
$$\frac{1}{n+1}\sum_{j=1}^{n+1} \ln\int_{x_{j-1}}^{x_j} f_{a,b}(x)dx = \frac{1}{n+1}\sum_{j=1}^{n+1} \ln\frac{x_j - x_{j-1}}{b-a}.$$

Lisades valimi järjestatud väärtuste võrratusahela otsadesse otsitavad parameetrid, saame $a < x_1 < x_2 ... < x_n < b$. Nüüd saame välja kirjutada suurimate vahemike funktsiooni, mis sisaldab ka otsitavaid parameetreid. Paneme selle kirja tuues eraldi välja parameetreid sisaldavad liikmed,

$$S(a,b) = \frac{1}{n+1}\sum_{j=1}^{n+1} \ln(x_j - x_{j-1}) - \ln(b-a) =$$
$$\frac{1}{n+1}\Big[\ln(x_1 - a) + \sum_{j=2}^{n} \ln(x_j - x_{j-1}) + \ln(b - x_n) - (n+1)\ln(b-a)\Big].$$

Järgmiseks saame võtta suurimate vahemike funktsioonist osatuletised otsitavate parameetrite järgi, $\frac{\partial S(a,b)}{\partial a}$ ja $\frac{\partial S(a,b)}{\partial b}$ ning võrdsustada nulliga. Tulemuseks on võrrandisüsteem:

$$\begin{cases} \frac{-1}{(n+1)(x_1-a)} + \frac{1}{(b-a)} = 0 \\ \frac{1}{(n+1)(b-x_n)} - \frac{1}{(b-a)} = 0, \end{cases}$$

mille lahendamisel avalduvad hinnangud $\tilde{a}$ ja $\tilde{b}$ järmiselt

$$\tilde{a} = \frac{nx_1 - x_n}{(n-1)}, \qquad \tilde{b} = \frac{nx_n - x_1}{(n-1)}.$$

Suurima tõepära ja suurimate vahemike hinnangud muutuvad valimi kasvades üha lähedasemaks, sest

$$\lim_{n\to\infty} \tilde{a} = \lim_{n\to\infty}\left(\frac{nx_1}{n-1} - \frac{x_n}{n-1}\right) = x_1 = \hat{a},$$
$$\lim_{n\to\infty} \tilde{b} = \lim_{n\to\infty}\left(\frac{nx_n}{n-1} - \frac{x_1}{n-1}\right) = x_n = \hat{b}.$$

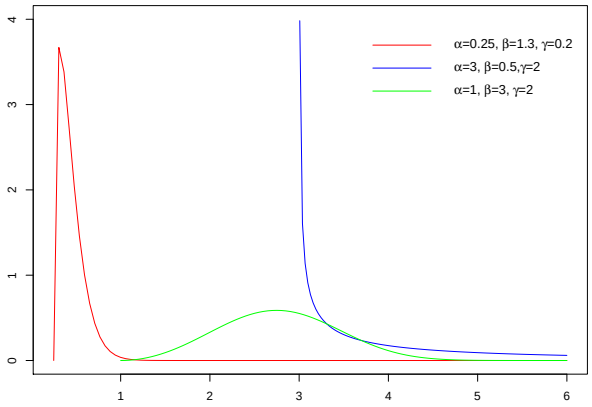
# 2 Simulatsiooninäited kahe meetodi võrdlemiseks

Järgnevas illustreerime töö esimeses osas vaadeldud pidevate jaotuste parameetrite hindamise meetodite erinevusi simulatsiooninäidete abil. Näitejaotuseks on kolmeparameetriline Weibulli jaotus. Weibulli jaotuse tihedusfunktsioon ja jaotusfunktsioon on vastavalt kujul

$$f_{\mathcal{W}}(x) = \frac{\beta}{\gamma}\left(\frac{x-\alpha}{\gamma}\right)^{\beta-1} e^{-\left(\frac{x-\alpha}{\gamma}\right)^{\beta}}, \quad F_{\mathcal{W}}(x) = 1 - e^{-\left(\frac{x-\alpha}{\gamma}\right)^{\beta}}, \ x > \alpha,$$

kus $\alpha \in \mathbb{R}$ on asukohaparameeter *(location parameter)*, $\beta > 0$ on kujuparameeter *(shape parameter)* ja $\gamma > 0$ on skaalaparameeter *(scale parameter)*.

Joonisel 1 on kujutatud Weibulli jaotuse tihedusfunktsioonid erinevate parameetri kombinatsioonide korral. Kujutatud on need tihedused, mida kasutatakse allolevates simulatsiooninäidetes.




Joonis 1. Weibulli jaotuse kuju parameetrite erinevate väärtuste korral

**Näide 4.** Eksponentjaotus on Weibulli jaotuse erijuht, kus $\alpha = 0$, $\beta = 1$ ja $\gamma = \frac{1}{\lambda}$. Sel juhul on tihedus kujul

$$f_{\mathcal{W}}(x) = \lambda e^{-\lambda x}.$$

## 2.1 Tõepärafunktsiooni ja vahemike funktsiooni käitumine

Järgnevas alapeatükis on näidete aines võetud artiklist Cheng ja Amin (1983).

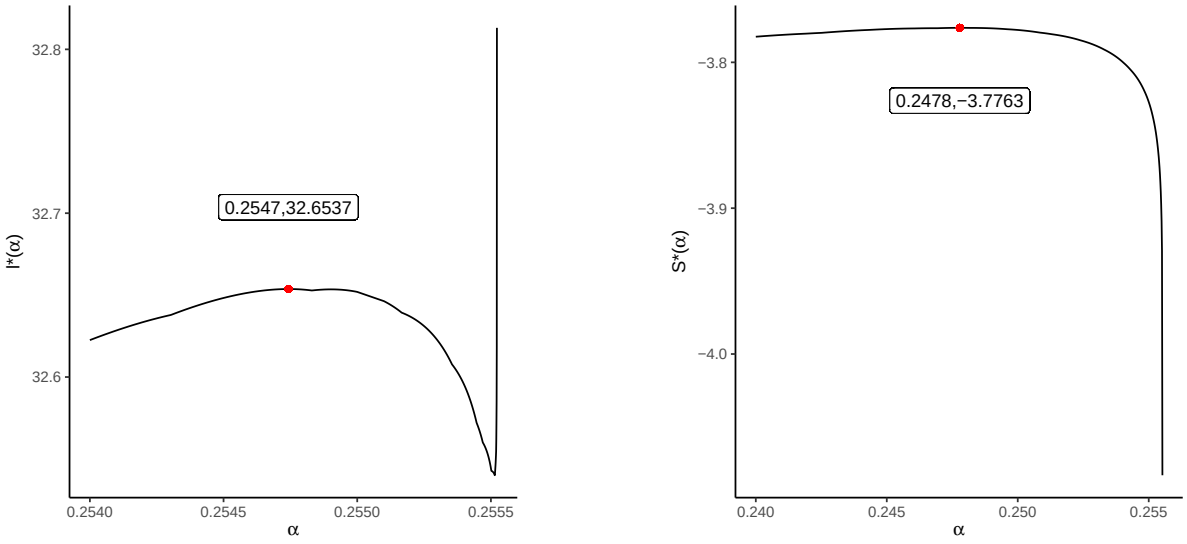
Teooriast on teada, et lisaks normaaljaotuste segu juhule, kus suurima tõepära hinnangud ei ole mõjusad, ei tööta suurima tõepära meetod alati ka kolmeparameetrilise Weibulli jaotuse korral. Tõepärafunktsioon on kujuparameetri $\beta < 1$ korral ülalt tõkestamata ning suurima tõepära meetodil saadud hinnangud ei ole mõjusad (Cheng ja Amin, 1983). Siis on jaotus J-kujuline ning tõenäosusmass koondub asukohaparameetri $\alpha$ ehk jaotuse algpunkti lähedusse (joonis 1). Samal ajal aga suurimate vahemike meetodil saadud hinnangud on mõjusad. Näidetena vaadeldakse olukordi, kus suurima tõepära meetodil hinnangute leidmine on ühel juhul raskendatud (näide 5) ja teisel juhul võimatu (näide 6).

Olgu meil antud valim, mis pärineb kolmeparameetrilisest Weibulli jaotusest. Soovime hinnata selle jaotuse parameetreid kasutades suurima tõepära ja suurimate vahemike meetodit. Kuna soovime lisaks hinnangute leidmisele uurida ka tõepära ja vahemike funktsiooni käitumist, leiame parameetrite hinnangud kahel viisil – statistikatarkvara R paketiga „lmomco“ ja võremeetodil (*grid search*).

Võremeetod tähendab seda, et valime esmalt sobiva vahemiku parameetri $\alpha$ jaoks. Fikseerime järjest väärtusi selles vahemikus ja leiame iga kord teiste parameetrite $\beta$ ja $\gamma$ väärtuste paari, mis maksimeerib vaatluse all oleva funktsiooni. Osaliselt maksimeeritud funktsioone tähistame vastavalt $l^*(\alpha) = l\Big(\alpha, \hat{\beta}(\alpha), \hat{\gamma}(\alpha)\Big)$ ja $S^*(\alpha) = S\Big(\alpha, \tilde{\beta}(\alpha), \tilde{\gamma}(\alpha)\Big)$. Nii saame lisaks hinnangute leidmisele joonisel kujutada maksimeeritud funktsiooni väärtused iga vaadeldud $\alpha$ korral ning seega näha, kas funktsioonidel leidub globaalne (või lokaalne) maksimum.

Nagu öeldud, on hinnangute leidmiseks kasutatud ka statistikatarkvara R paketti „lmomco“, mis on üks vähestest, mis võimaldab arvutada suurima tõepära ja suurimate vahemike hinnanguid 3-parameetrilise Weibulli jaotuse parameetritele. Alglähenditena on kasutatud paketi poolt vaikimisi seadistatud meetodil leitud väärtusi. Täispikk kood, mille abil on loodud ka ka allolevad joonised, on vastavalt ära toodud lisas 1 ja 2.

**Näide 5.** Vaatleme juhtu, kus $\beta > 1$ (joonis 2). Andmed on simuleeritud Weibulli jaotusest parameetritega $\alpha = 0.25$, $\beta = 1.3$ ja $\gamma = 0.2$. Valimi suuruseks on 30.




Joonis 2. Maksimeeritud tõepära ja vahemike hinnangufunktsioonid näites 5

Jooniselt 2 on näha, et funktsioon $l^*(\alpha)$ saavutab lokaalse maksimumi $\hat{\alpha} = 0.2547$ korral. Kui parameetri $\alpha$ väärtus hakkab lähenema valimi vähimale väärtusele (0.25552), hakkab tõepärafunktsioon tõkestamatult kasvama. Vahemike funktsioon $S^*(\alpha)$ on tõkestatud ja saavutab $\tilde{\alpha} = 0.2478$ korral globaalse maksimumi. Asukohaparameetri hinnangud koos vastavate funktsiooni väärtustega on joonisel märgitud punase punktiga (lisaks toodud punkti koordinaadid).

Seda juhtu nimetame suurima tõepära meetodil hinnangute raskendatud leidmiseks, sest selles töös on suurima tõepära hinnang defineeritud tõepära funktsiooni globaalse maksimumi kaudu (peatükk 1.2). Nagu selles peatükis mainitud, võib mõnel juhul hinnangut defineerida ka funktsiooni lokaalse maksimumi kaudu. Näeme, et $\hat{\alpha} = 0.2547$ ja $\tilde{\alpha} = 0.2478$, seega võime seda lokaalse maksimumi kaudu saadud hinnangut praegusel juhul arvesse võtta.

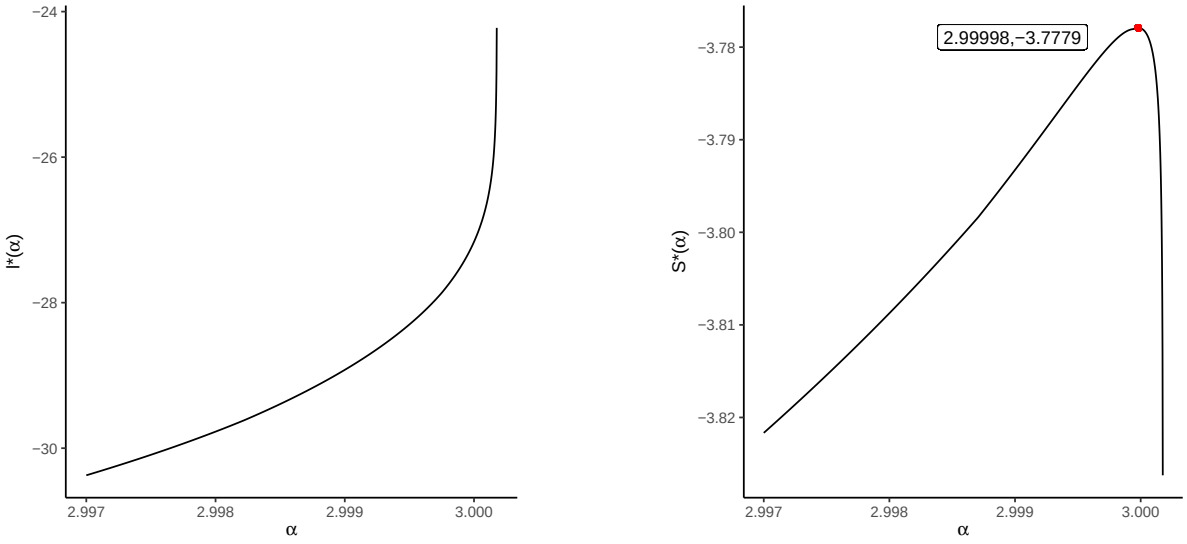
Pakett „lmomco“ andis parameetritele järgnevad hinnangud:

1. suurima tõepära meedodil $\hat{\alpha} = 0.2548, \hat{\beta} = 1.1192$ ja $\hat{\gamma} = 0.1299$,
2. suurimate vahemike meetodil $\tilde{\alpha} = 0.2475, \tilde{\beta} = 1.1720$ ja $\tilde{\gamma} = 0.1453$.

Võremeetodiga saadud hinnangud olid:

1. suurima tõepära meetodil: $\hat{\alpha} = 0.2547, \hat{\beta} = 1.1286$ ja $\hat{\gamma} = 0.1298$,
2. suurimate vahemike meetodil: $\tilde{\alpha} = 0.2478, \tilde{\beta} = 1.1694$ ja $\tilde{\gamma} = 0.1431$.

**Näide 6.** Teiseks vaatleme juhtu, kus $\beta < 1$ (joonis 3). Valimi suuruseks on jälle 30 ning andmed on simuleeritud Weibulli jaotusest parameetritega $\alpha = 3$, $\beta = 0.5$ ja $\gamma = 2$. Valimi vähimaks väärtuseks on 3.00017.




Joonis 3. Maksimeeritud tõepära ja vahemike hinnangufunktsioonid näites 6

Jooniselt 3 on näha maksimeeritud tõepärafunktsiooni $l^*(\alpha)$ tõkestamatu kasvamine, samal ajal kui suurimate vahemike funktsioon $S^*(\alpha)$ saavutab $\tilde{\alpha} = 2.99998$ korral globaalse maksimumi.

Võremeetodiga leitud suurimate vahemike hinnangud olid: $\tilde{\alpha} = 2.99998$, $\tilde{\beta} = 0.4265$ ja $\tilde{\gamma} = 0.7857$. Asukohaparameetri hinnangu $\tilde{\alpha}$ väärtus ja sellele vastav suurimate vahemike funktsiooni väärtus on joonisel märgitud punase punktiga (lisaks toodud punkti koordinaadid). Suurima tõepära meetod ei tööta ja seega ei saa selle meetodiga parameetrite hin-

nanguid leida. Pakett „lmomco“ väljastas mittekoondumisele viitava veateate. Suurimate vahemike meetodil olid „lmomco“ hinnanguteks: $\tilde{\alpha} = 2.99997, \tilde{\beta} = 0.4357$ ja $\tilde{\gamma} = 0.8131$.

## 2.2 Suurima tõepära ja suurimate vahemike meetodil saadud parameetrite hinnangute võrdlus

Järgnevas alapeatükis illustreerime suurima tõepära ja suurimate vahemike hinnangute omadusi juhul, kui mõlemad meetodid töötavad ja on teada, et hinnangute asümptootiline jaotus on sama. See kehtib juhul, kui $\beta > 2$ (Teoreem 1 ja 2 allikast Cheng ja Amin, 1983).

**Näide 7.** Genereeriti erinevate suurustega valimeid Weibulli jaotusest parameetritega $\alpha = 1$, $\beta = 3$ ja $\gamma = 2$. Vaadeldud valimimahtudeks olid $n = 50$, $n = 100$ ja $n = 1000$. Iga kord simuleeriti 1000 vastava suurusega valimit ning iga valimi põhjal leiti hinnangud kolmele parameetrile paketiga „lmomco“. Simuleerimiseks kasutatud kood on ära toodud lisas 3.

**Definitsioon 8.** *Olgu n valimimaht, m vaadeldud valimite arv ja $\theta$ hinnatav parameeter. Hinnangu $\hat{\theta}_n$ ruutkeskmist viga ja nihet (2) saab hinnata vastavalt*

$$MSE(\hat{\theta}_n) = \frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_n^{(i)} - \theta)^2,$$

$$B(\hat{\theta}_n) = \frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_n^{(i)} - \theta),$$

*kus $\hat{\theta}_n^{(i)}$ on i-nda valimi põhjal leitud hinnang.*

Näeme (tabel 4), et mõlema meetodi korral hinnangute ruutkeskmine viga ja nihe muutuvad valimi kasvades järjest väiksemaks. See ühtib teadmisega, et kui vaatluse all on Weibulli kolmeparameetriline jaotus, kus $\beta > 2$, ja valimimaht on suur, koonduvad suurima tõepära ja suurimate vahemike hinnangute ruutkeskmine viga ja nihe nulliks, ehk vastavalt

$$MSE(\hat{\theta}_n) \xrightarrow{n\to\infty} 0, \qquad B(\hat{\theta}_n) \xrightarrow{n\to\infty} 0.$$

See tähendab seda, et tegemist on mõjusate hinnangutega (3).

Tabel 4. Kokkuvõte hinnangute omadustest erinevate valimimahtude korral

| | Suurima tõepära meetod | | Suurimate vahemike meetod | |
|---|---|---|---|---|
| $n$ | $MSE(\hat{\alpha}_n)$ | $B(\hat{\alpha}_n)$ | $MSE(\tilde{\alpha}_n)$ | $B(\tilde{\alpha}_n)$ |
| 50 | 1.2115 | $-0.0404$ | 1.9103 | $-0.3191$ |
| 100 | 0.1033 | 0.0183 | 0.1672 | $-0.1331$ |
| 1000 | 0.0058 | 0.0080 | 0.0067 | $-0.0247$ |
| $n$ | $MSE(\hat{\beta}_n)$ | $B(\hat{\beta}_n)$ | $MSE(\tilde{\beta}_n)$ | $B(\tilde{\beta}_n)$ |
| 50 | 4.7608 | 0.1893 | 6.4253 | 0.4657 |
| 100 | 0.4536 | 0.0271 | 0.5768 | 0.1686 |
| 1000 | 0.0254 | $-0.0063$ | 0.0276 | 0.0334 |
| $n$ | $MSE(\hat{\gamma}_n)$ | $B(\hat{\gamma}_n)$ | $MSE(\tilde{\gamma}_n)$ | $B(\tilde{\gamma}_n)$ |
| 50 | 1.2645 | 0.0293 | 1.9860 | 0.3309 |
| 100 | 0.1231 | $-0.0249$ | 0.1767 | 0.1364 |
| 1000 | 0.0074 | $-0.0100$ | 0.0083 | 0.0255 |

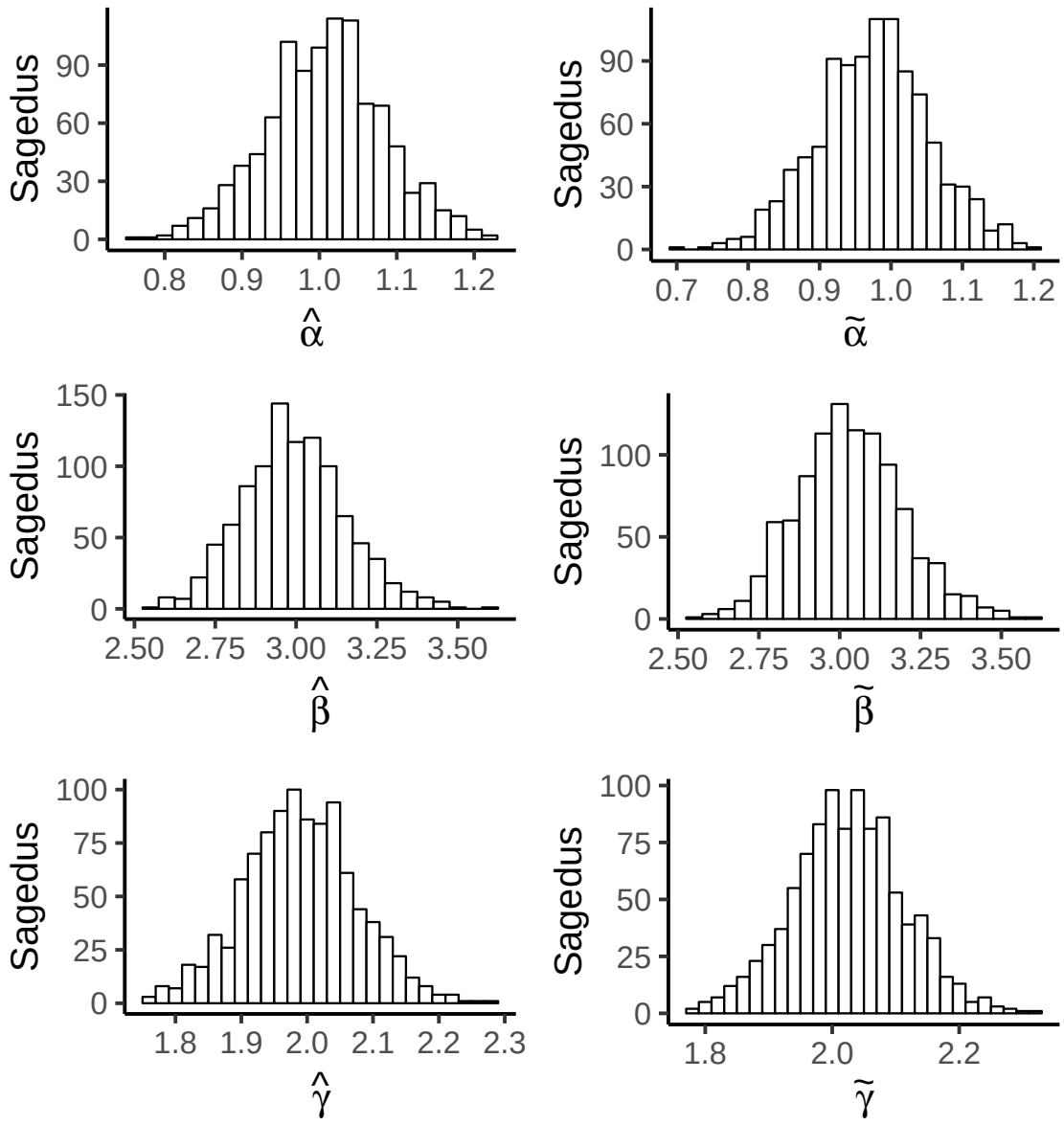
Lisaks sellele saame tähele panna, et asukohaparameetri $\alpha$ hinnangu nihke hinnang on suurima vahemike meetodil alati negatiivne, $B(\tilde{\alpha}_n) < 0$. Seda nägime ka joonistelt 2 ja 3, kus suurimate vahemike meetod hindas parameetri väärtust tegelikust väiksemaks.

Tabelist 4 jäävad silma suured ruutkeskmiste vigade hinnangud valimimahu $n = 50$ korral. Uurides neid juhte lähemalt, selgub, et teatud valimite korral on saadud hinnangud väga kaugel parameetri tegelikust väärtusest. Näiteks ühe konkreetse valimi korral olid „lmomco“ hinnanguteks

1. suurima tõepära meedodil $\hat{\alpha} = 6.46, \hat{\beta} = 18.29$ ja $\hat{\gamma} = 9.65$,
2. suurimate vahemike meetodil $\tilde{\alpha} = 19.47, \tilde{\beta} = 41.18$ ja $\tilde{\gamma} = 22.68$.

Vaadeldes sama valimi histogrammi, on näha, et see ei järgi uuritava jaotuse tihedust. Väikese valimimahu probleemiks ongi mõnikord väga ebatõenäolise valimi realiseerumine ning seega, kui valimi jaotus ei sarnane tegelikule jaotusele, on ka parameetrite hinnangud tegelikest parameetritest väga erinevad.

Suurte valimimahtude korral, kus $\beta > 2$, on suurima tõepära ja suurimate vahemike hinnangud Weibulli jaotuse parameetritele normaaljaotusega ning hinnangute keskväärtus koondub parameetri õigeks väärtuseks. Saadud hinnangute histogrammid valimimahu $n = 1000$ korral on toodud joonisel 4.




Joonis 4. Suurima tõepära ja suurimate vahemike meetodil leitud parameetrite hinnangute jaotus $n$=1000 korral näites 7

# Kokkuvõte

Käesoleva bakalaureusetöö eesmärk oli kahe pidevate jaotuste parameetrite hindamise meetodi uurimine. Vaadeldavateks meetoditeks olid suurima tõepära ja suurimate vahemike meetod. Mõlemad meetodid tuletati Kullback-Leibleri informatsioonimõõdu abil, mis seisneb kahe funktsiooni – juhusliku suuruse tegeliku jaotusfunktsiooni ja valimi abil hinnatud jaotuse – keskmise kauguse minimeerimises.

Lisaks meetodite tuletuskäigule uuriti juhte, kus suurima tõepära meetod ei tööta. Selgus, et see juhtub siis, kui tõepärafunktsioon on ülalt tõkestamata. See probleem esineb näiteks normaaljaotuste segu ja kolmeparameetrilise Weibulli jaotuse korral, kus kujuparameeter $\beta < 1$. Weibulli jaotuse korral, kus kujuparameeter $\beta > 1$, ei pruugi globaalse maksimumi kaudu defineeritud suurima tõepära hinnangud leiduda, sest tõepärafunktsioon võib saavutada vaid lokaalseid maksimume. Suurimate vahemike funktsioon aga saavutab alati globaalse maksimumi, sest on ülevalt tõkestatud.

Meetodite eripärade kõrval vaadeldi ka nende ühiseid omadusi. Näiteks ühtlase jaotuse korral on suurima tõepära ja suurimate vahemike hinnangud suurte valimite korral sarnased. Ka Weibulli jaotuse korral, kus $\beta > 2$, käituvad hinnangud sarnaselt: suurte valimimahtude korral on mõlema meetodi hinnangud normaaljaotusega, mille keskväärtus koondub parameetri tegelikuks väärtuseks.

# Kasutatud kirjandus


Bishop, C.M. (2006). *Pattern Recognition and Machine Learning.* Springer, New York.

Cheng, R.C.H., Amin, N.A.K. (1983). Estimating parameters in continuous univariate distributions with a shifted origin. *Journal of the Royal Statistical Society, B.* 45 (3), 394-403.

Dudewicz, E., Mishra, S.N. (1988). *Modern Mathematical Statistics.* Wiley, New York.

Lehmann, E.L., Casella, G. (1998). *Theory of Point Estimation.* Springer, New York.

Lember, J. (2018). *Informatsiooniteooria. Loengukonspekt ja ülesanded.* Tartu Ülikool.

Lepik, N. (2017). *Tõenäosusteooria ja statistika II.* Loengukonspekt. Tartu Ülikool.

Ranneby, B. (1984). The maximum spacing method. An estimation method related to the maximum likelihood method. *Scandinavian Journal of Statistics,* 11 (2), 93-112.

# Lisad

## Lisa 1. Kood simulatsiooninäite 5 jaoks

```
library("lmomco")
library("ggplot2")

# Jaotus- ja tihedusfunktsioon:
weibull_distribution<-function(x,alfa,beeta,gamma){
  1-exp(-1*(((x-alfa)/gamma)**beeta))
}

weibull_density<-function(x,alfa,beeta,gamma){
  beeta*(gamma**(-beeta))*((x-alfa)**(beeta-1))*exp(-1*((x-alfa)/(gamma))**beeta)

}
options(scipen=999999)
set.seed(1230)
x<-rweibull(30,1.3,0.2)+0.25
x<-sort(x)
x

# Hinnangud paketiga "lmomco"
mle2par(x,type="wei",silent=FALSE)$optim$value
mle2par(x,type="wei")$para
mps2par(x,type="wei",silent=FALSE)$optim$value
mps2par(x,type="wei")$para

# Parameetrite muutumisvahemikud
# asukohaparameeter
alfa<-seq(0.254,0.255,length.out=300)
alfa<-append(alfa,seq(0.2551,0.2555225,length.out=700))
# kujuparameeter
beeta<-seq(0.7,1.7,length.out=50)
beeta
# sklaalaparameeter
gamma<-seq(0.01,0.336,length.out=50)
gamma

##### SUURIMA TÕEPÄRA HINNANGUD VÕREMEETODIL

f_max_lik<-function(x,a,b,c){
  sum(log(weibull_density(x,a,b,c)))
}

M_MAX_LIK <- matrix(data = c(1,2,3,4), nrow = 1, ncol = 4)

for (k in 1:length(alfa)){
  m_max_lik <- matrix(data = c(1,2,3,4), nrow = 1, ncol = 4)
  valim<-x
  for (i in 1:length(beeta)){
    for (j in 1:length(gamma)){
      max_lik<-f_max_lik(valim,alfa[k],beeta[i],gamma[j])
```

```
    m_max_lik <- rbind(m_max_lik, c(alfa[k],beeta[i],gamma[j],max_lik))
   }
  }
  m_max_lik<-m_max_lik[-1,]
  l<-which(m_max_lik[,4]==max(m_max_lik[,4],na.rm=T))
  M_MAX_LIK <-rbind(M_MAX_LIK, c(m_max_lik[l,1],m_max_lik[l,2],m_max_lik[l,3],m_max_lik[l,4]))

}
M_MAX_LIK<-M_MAX_LIK[-1,]

# maksimumi saavutamise iteratsioonisamm
which(M_MAX_LIK[,4]==max(M_MAX_LIK[,4])) # kõige viimane

# lokaalse maksimumi iteratsioonisamm
which((M_MAX_LIK[,4]==max(M_MAX_LIK[,4][1:926])))
# sellise alfa korral
a<-M_MAX_LIK[,1][which((M_MAX_LIK[,4]==max(M_MAX_LIK[,4][1:926])))]
a
# beeta
b<-M_MAX_LIK[,2][which((M_MAX_LIK[,4]==max(M_MAX_LIK[,4][1:926])))]
b
# gamma
g<-M_MAX_LIK[,3][which((M_MAX_LIK[,4]==max(M_MAX_LIK[,4][1:926])))]
g
# tõepära väärtus
t<-M_MAX_LIK[,4][which((M_MAX_LIK[,4]==max(M_MAX_LIK[,4][1:926])))]
t

data<-data.frame(M_MAX_LIK[,1],M_MAX_LIK[,4])

ggplot(data=data, aes(x=M_MAX_LIK[,1], y=M_MAX_LIK[,4])) +
  geom_line() +
  xlab(expression(alpha)) +ylab((expression(paste("l*(",alpha,")"))))+
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black")) +
  geom_point(aes(x=a,y=t), color="red")  +
  geom_label(aes(a,t+0.05),label=paste(round(a,4),round(t,4),sep=","))

##### SUURIMATE VAHEMIKE HINNANGUD VÕREMEETODIL

alfa<-seq(0.24,0.2555225,length.out=1000)
n<-length(x)
x<-c(NA,x,NA)

f_max_sp<-function(x,a,b,c){
  s<-c(NA)
  for(i in 1:(length(x)-1)){
    if(i==1){
      p<-0
      o<-weibull_distribution(x[i+1],a,b,c)
      q<-(o-p)
      r<-log(q)
      s<-append(s,r)
    }
    else if (i==(length(x)-1)){
      p<-weibull_distribution(x[i],a,b,c)
```

```
    o<-1
    q<-(o-p)
    r<-log(q)
    s<-append(s,r)
   }
   else{
    o<-weibull_distribution(x[i+1],a,b,c)
    p<-weibull_distribution(x[i],a,b,c)
    q<-(o-p)
    r<-log(q)
    s<-append(s,r)
   }
  }
  sum(s[-1])/(n+1)}

M_MAX_SP <- matrix(data = c(1,2,3,4), nrow = 1, ncol = 4)

for (k in 1:length(alfa)){
  m_max_sp<-matrix(data=c(1,2,3,4), nrow=1, ncol=4)
  valim<-x
  for (i in 1:length(beeta)){
    for (j in 1:length(gamma)){
      max_sp<-f_max_sp(valim,alfa[k],beeta[i],gamma[j])
      m_max_sp <- rbind(m_max_sp, c(alfa[k],beeta[i],gamma[j],max_sp))
    }
  }
  m_max_sp<-m_max_sp[-1,]
  k<-which(m_max_sp[,4]==max(m_max_sp[,4],na.rm=T))
  M_MAX_SP <-rbind(M_MAX_SP, c(m_max_sp[k,1],m_max_sp[k,2],m_max_sp[k,3],m_max_sp[k,4]))

}
M_MAX_SP<-M_MAX_SP[-1,]

# maksimaalse väärtuse iteratsioonisamm
which(M_MAX_SP[,4]==max(M_MAX_SP[,4]))
# alfa
a<-M_MAX_SP[,1][which(M_MAX_SP[,4]==max(M_MAX_SP[,4]))]
a
# beeta
b<-M_MAX_SP[,2][which(M_MAX_SP[,4]==max(M_MAX_SP[,4]))]
b
# gamma
g<-M_MAX_SP[,3][which(M_MAX_SP[,4]==max(M_MAX_SP[,4]))]
g
# vahemike väärtus
t<-M_MAX_SP[,4][which(M_MAX_SP[,4]==max(M_MAX_SP[,4]))]

data<-data.frame(M_MAX_SP[,1],M_MAX_SP[,4])

ggplot(data=data, aes(x=M_MAX_SP[,1], y=M_MAX_SP[,4])) +
  geom_line() +
  xlab(expression(alpha)) +ylab((expression(paste("S*(",alpha,")")))) +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black")) +
  geom_point(aes(x=a,y=t), color="red") +
  geom_label(aes(a,t-0.05),label=paste(round(a,4),round(t,4),sep=","))
```

# Lisa 2. Kood simulatsiooninäite 6 jaoks

```
# Funktsioonid defineeritud lisas 1.

set.seed(1230)

x<-rweibull(30,0.5,2)+3
x<-sort(x)
x

# Hinnangud paketiga "lmomco"

mle2par(x,type="wei",silent=FALSE)$optim$value
mle2par(x,type="wei")$para

mps2par(x,type="wei",silent=FALSE)$optim$value
mps2par(x,type="wei")$para


# Parameetrite muutumisvahemikud
# asukohaparameeter
alfa<-seq(2.997,3.0001769,by=0.000002)
# kujuparameeter
beeta<-seq(0.1,1.1,length.out=50)
# skaalaparameeter
gamma<-seq(0.5,2.5,length.out=50)



##### SUURIMA TÕEPÄRA HINNANGUD VÕREMEETODIL

M_MAX_LIK <- matrix(data = c(1,2,3,4), nrow = 1, ncol = 4)

for (k in 1:length(alfa)){
 m_max_lik <- matrix(data = c(1,2,3,4), nrow = 1, ncol = 4)
 valim<-x
 for (i in 1:length(beeta)){
  for (j in 1:length(gamma)){
   max_lik<-f_max_lik(valim,alfa[k],beeta[i],gamma[j])
   m_max_lik <- rbind(m_max_lik, c(alfa[k],beeta[i],gamma[j],max_lik))
  }
 }
 m_max_lik<-m_max_lik[-1,]
 l<-which(m_max_lik[,4]==max(m_max_lik[,4],na.rm=T))
 M_MAX_LIK <-rbind(M_MAX_LIK, c(m_max_lik[l,1],m_max_lik[l,2],m_max_lik[l,3],m_max_lik[l,4]))

}
M_MAX_LIK<-M_MAX_LIK[-1,]

data<-data.frame(M_MAX_LIK[,1],M_MAX_LIK[,4])

ggplot(data=data, aes(x=M_MAX_LIK[,1], y=M_MAX_LIK[,4])) +
 geom_line() +
 xlab(expression(alpha)) +ylab((expression(paste("l*(",alpha,")"))))+
 theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
       panel.background = element_blank(),
       axis.line = element_line(colour = "black"))
```

```
##### SUURIMATE VAHEMIKE HINNANGUD VÕREMEETODIL

n<-length(x)
x<-c(NA,x,NA)

M_MAX_SP <- matrix(data = c(1,2,3,4), nrow = 1, ncol = 4)

for (k in 1:length(alfa)){
 m_max_sp<-matrix(data=c(1,2,3,4), nrow=1, ncol=4)
 valim<-x
 for (i in 1:length(beeta)){
  for (j in 1:length(gamma)){
   max_sp<-f_max_sp(valim,alfa[k],beeta[i],gamma[j])
   m_max_sp <- rbind(m_max_sp, c(alfa[k],beeta[i],gamma[j],max_sp))
  }
 }
 m_max_sp<-m_max_sp[-1,]
 k<-which(m_max_sp[,4]==max(m_max_sp[,4],na.rm=T))
 M_MAX_SP <-rbind(M_MAX_SP, c(m_max_sp[k,1],m_max_sp[k,2],m_max_sp[k,3],m_max_sp[k,4]))

}

M_MAX_SP<-M_MAX_SP[-1,]

# funktsiooni maksimaalne väärtus sellise iteratsioonisammu korral
which(M_MAX_SP[,4]==max(M_MAX_SP[,4]))
# alfa
a<-M_MAX_SP[,1][which(M_MAX_SP[,4]==max(M_MAX_SP[,4]))]
a
# beeta
b<-M_MAX_SP[,2][which(M_MAX_SP[,4]==max(M_MAX_SP[,4]))]
b
# gamma
g<-M_MAX_SP[,3][which(M_MAX_SP[,4]==max(M_MAX_SP[,4]))]
g
# funktsiooni väärtus
t<-M_MAX_SP[,4][which(M_MAX_SP[,4]==max(M_MAX_SP[,4]))]


data<-data.frame(M_MAX_SP[,1],M_MAX_SP[,4])


ggplot(data=data, aes(x=M_MAX_SP[,1], y=M_MAX_SP[,4])) +
 geom_line() +
 xlab(expression(alpha)) +ylab((expression(paste("S*(",alpha,")")))) +
 theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
       panel.background = element_blank(),
       axis.line = element_line(colour = "black")) +
 geom_point(aes(x=a,y=t), color="red") +
 geom_label(aes(a-0.001,t-0.001),label=paste(round(a,5),round(t,4),sep=","))
```

# Lisa 3. Kood simulatsiooninäite 7 jaoks

```
### n=50; n=100; n=1000

library(lmomco)
set.seed(1000)
k <- 0; n <- 50; m <-  2000
alfa<-1; beeta<-3; gamma<-2
M <- matrix(data = 1:(n+6), nrow = 1, ncol = n+6)

for (i in 1:m){
 x<-rweibull(n,beeta,gamma)+alfa
 ml_est<-mle2par(x,type="wei")
 mps_est<-mps2par(x,type="wei")
 if (!is.null(ml_est)){ # kontroll, et mõlema meetodiga saaks hinnanguid leida
  k <- k+1
  M<-rbind(M, c(x,ml_est$para[1],ml_est$para[3],ml_est$para[2],
           mps_est$para[1],mps_est$para[3],mps_est$para[2]))

 }
 if (k==1000) break

}

# Katsetuste arv:
i

M<-M[-1,]


# ML

alfa_ml <-M[,n+1]
beeta_ml <-M[,n+2]
gamma_ml <-M[,n+3]

# ml alfa
hist(alfa_ml)
range(alfa_ml)
# hinnangu nihe (bias)
bias_alfa_ml <-1/length(alfa_ml)*(sum((-1*alfa_ml)-alfa))
bias_alfa_ml
# keskmine ruutviga (mse)
mse_alfa_ml <-1/length(alfa_ml)*sum(((-1*alfa_ml)-alfa)**2)
mse_alfa_ml

# ml beeta
hist(beeta_ml)
range(beeta_ml)
# hinnangu nihe (bias)
bias_beeta_ml <-1/length(beeta_ml)*(sum(beeta_ml-beeta))
bias_beeta_ml
# keskmine ruutviga (mse)
mse_beeta_ml <-1/length(beeta_ml)*sum((beeta_ml-beeta)**2)
mse_beeta_ml
```

```
# ml gamma
hist(gamma_ml)
range(gamma_ml)
# hinnangu nihe (bias)
bias_gamma_ml <-1/length(gamma_ml)*(sum(gamma_ml-gamma))
bias_gamma_ml
# keskmine ruutviga (mse)
mse_gamma_ml <-1/length(gamma_ml)*sum((gamma_ml-gamma)**2)
mse_gamma_ml



# MPS


alfa_mps <-M[,n+4]
beeta_mps <-M[,n+5]
gamma_mps <-M[,n+6]


# mps alfa
hist(alfa_mps)
range(alfa_mps)
# hinnangu nihe (bias)
bias_alfa_mps <-1/length(alfa_mps)*(sum((-1*alfa_mps)-alfa))
bias_alfa_mps
# keskmine ruutviga (mse)
mse_alfa_mps <-1/length(alfa_mps)*sum(((-1*alfa_mps)-alfa)**2)
mse_alfa_mps

# mps beeta
hist(beeta_mps)
range(beeta_mps)
# hinnangu nihe (bias)
bias_beeta_mps <-1/length(beeta_mps)*(sum(beeta_mps-beeta))
bias_beeta_mps
# keskmine ruutviga (mse)
mse_beeta_mps <-1/length(beeta_mps)*sum((beeta_mps-beeta)**2)
mse_beeta_mps

# mps gamma
hist(gamma_mps)
range(gamma_mps)
# hinnangu nihe (bias)
bias_gamma_mps <-1/length(gamma_mps)*(sum(gamma_mps-gamma))
bias_gamma_mps
# keskmine ruutviga (mse)
mse_gamma_mps <-1/length(gamma_mps)*sum((gamma_mps-gamma)**2)
mse_gamma_mps
```

**Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks**

Mina, Liis Simmul,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

   **„Pidevate jaotuste parameetrite hindamisest suurima tõepära ja suurimate vahemike meetodil“**,

   mille juhendaja on Kristi Kuljus,
   reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Liis Simmul
**08.05.2019**