Jakauman keskilukujen robustista estimoinnista

Toni Rissanen

LuK-tutkielma
Joulukuu 2023

MATEMATIIKAN JA TILASTOTIETEEN LAITOS

TURUN YLIOPISTO
Matematiikan ja tilastotieteen laitos

Toni Rissanen: Jakauman keskilukujen robustista estimoinnista
LuK-tutkielma, 18 s., 4 liites.
Tilastotiede
Joulukuu 2023

Tämä työ keskittyy tutkimaan tuntemattoman jakauman keskikohdan estimoinnin robustisuutta, käyttäen yleisesti käytettyjä keskilukuja, kuten otoskeskiarvoa ja mediaania. Työ alkaa klassisten tilastollisten menetelmien toimintaperiaatteiden esittelyllä ja tarkastelee niiden käytössä ilmeneviä käytännön ongelmia. Havaitaan, että klassiset menetelmät, jotka perustuvat normaalijakaumaoletukseen, voivat aiheuttaa ongelmia jakauman keskikohdan estimoinnissa tilanteissa, joissa jakauma ei ole normaalisti jakautunut. Seuraavaksi työssä esitellään robustit menetelmät, joiden tarkoituksena on estimoida jakauman keskikohtaa tarkemmin kuin klassiset menetelmät tilanteissa, joissa normaalijakaumaoletuksesta poiketaan.

Otoskeskiarvon ja mediaanin käyttäytymistä jakauman keskikohdan estimoinnissa tutkitaan työssä ensin 16 havainnon aineistolla, jossa yksi arvo on poikkeava. Lisäksi otoskeskiarvon ja mediaanin käyttäytymistä tarkastellaan simuloimalla havaintojen poimintaa jakaumaseoksesta ja Cauchy-jakaumasta. Havaitaan, että poikkeava arvo tai muu normaalijakaumasta eroava tilanne vaikuttaa enemmän otoskeskiarvon estimointitarkkuuteen kuin mediaaniin.

Työssä esitellään jakaumaseos ja Cauchy-jakauma. Yksi esitetyistä keskeisistä tuloksista on, että Cauchy-jakaumalla ei ole odotusarvoa. Lisäksi esitellään M-estimaattori ja sen toimintaperiaate. Työssä konkreettisena esimerkkinä M-estimaattoreista esitellään Huber-estimaattori, jonka arvo lasketaan samassa 16 havainnon aineistossa kuin otoskeskiarvo ja mediaani. Tuloksista ilmenee, että Huber-estimaattori on robustimpi kuin otoskeskiarvo. Lopuksi työssä esitellään lähdekirjallisuutta, jota lukija voi käyttää syventyäkseen M-estimointiin.

Asiasanat: jakauman keskikohdan estimointi, robustisuus, M-estimaattori.

Sisällys

1 Johdanto

Kaikki tilastolliset menetelmät riippuvat suorasti tai epäsuorasti oletuksista. Näitä oletuksia käytetään yleisesti muotoilemaan tilastollinen mallinnus- tai analyysiongelma teoreettisesti ja laskennallisesti käytettävään muotoon. Kuitenkin yleisesti tiedetään, että näin muodostetut mallit ovat todellisuuden yksinkertaistuksia ja ne ovat likimääräisesti paikkaansapitäviä. Tutkijalla pitää olla riittävä luottamus siihen, että mallit ovat likimääräisesti paikkaansapitäviä. Tämän takia tässä työssä tutkitaan kuinka paljon mahdolliset poikkeamat vaikuttavat käytettyihin menetelmiin ja sitä kautta tuloksiin. Tämä johdanto seuraa Maronnan et al. kirjaa.[5, s. xxi-xxii]

Yleisin oletus on, että havaittu aineisto noudattaa normaalijakaumaa. Tätä oletusta on käytetty tilastotieteessä kaksi vuosisataa ja normaalijakaumaoletus on klassisten tilastollisten menetelmien taustalla. Suurin oikeutus normaalijakaumaoletukselle on, että monet oikeat aineistot noudattavat likimääräisesti normaalijakaumaa. Tämän lisäksi normaalijakaumasta voidaan johtaa kaavoja optimaalisiin tilastomenetelmiin kuten esimerkiksi suurimman uskottavuuden menetelmään. Tällaisia menetelmiä kutsutaan klassisiksi tilastollisiksi menetelmiksi. Nämä menetelmät kuitenkin vaativat, että aineisto noudattaa normaalijakaumaa ilman merkittäviä poikkeamia.

Usein käytännön aineistot noudattavat oletettua normaalijakaumaa likimääräisesti niin, että suurin osa havainnoista noudattavat oletettua jakaumaa ja loput jotain toista jakaumaa. Monet aineistot käytännössä vaikuttavat varsin normaalisti jakautuneilta, mutta pieni osa havainnoista ovat epätyypillisiä ollen kaukana suurimmasta osasta muuta aineistoa. Tällaisia epätyypillisiä havaintoja kutsutaan oudokeiksi. Näitä havaitaan yleisesti aineiston analyysien ja tilastollisten mallinnusten sovelluksissa. Jopa yksittäinen oudokki aineistossa voi aiheuttaa suuria muutoksia klassisiin tilastollisiin menetelmiin, jotka nojautuvat normaalijakaumaoletukseen. Oudokit aiheuttavat likimääräisesti normaalijakautuneen aineiston näyttävän jakauman keskeltä kuin normaalijakaumalta, mutta jakauman hännät ovat pidemmät ja paksut kuin normaalijakaumalla.

Aineiston noudattaessa likimääräisesti normaalijakaumaa voitaisiin odottaa, että klassiset tilastolliset menetelmät toimisivat likimääräisesti. Normaalijakaumaan nojautuvat tilastolliset menetelmät eivät kuitenkaan toimi tällöin sellaisenaan ja niitä on muokattava. Pitää tietää millaista poikkeama on normaalijakaumasta, jotta menetelmää voidaan muokata. Jos oletetaan, että aineisto on normaalisti jakautunut, mutta sen jakaumalla on todellisuudessa paksut hännät, klassiset menetelmät eivät toimi toivotulla tavalla. Esimerkiksi klassisten estimaattien luottamusväleistä voi muodostua leveitä tai luottamustasot voivat olla alhaisia.

Robusti lähestymistapa tilastollisessa mallintamisessa ja data-analyysissä pyrkii johtamaan menetelmiä, jotka tuottavat luotettavia estimaatteja parametreille ja niihin liittyviä testejä. Robusteja menetelmiä voidaan luonnehtia niin, että ne antavat likimääräisesti saman vastauksen kuin klassiset menetelmät, kun aineisto noudattaa tarkasti jakaumaa eikä siinä ole oudokkeja. Toisaalta, jos aineistossa on vähäinen määrä oudokkeja, niin robustit menetelmät antavat likimääräisesti saman vastauksen kuin klassiset menetelmät jakaumaa noudattavalla aineistolla. Tämän seurauksena robusteja menetelmiä käyttämällä voidaan luotettavasti havaita oudokkeja mo-

nissa erilaisissa aineistoissa. Robusti lähestymistapa toimii myös muille jakaumille, jotka ovat lähellä nominaalimallia.

Tässä työssä esitellään ensin motivaatio robustien estimaattien käyttöön. Myös esitellään klassisten tilastollisten estimaattien ongelma jakauman keskilukujen estimoinnissa, kun aineisto ei noudata tarkasti mitään tunnettua jakaumaa. Lisäksi tässä työssä simuloidaan keskiarvon ja mediaanin jakaumia, kun otokset otetaan jakaumista, jotka eivät ole normaalijakaumia. Lopuksi esitellään M-estimaattori. Tämä estimaattori toimii likimäärin yhtä hyvin tilanteissa, joissa aineisto noudattaa tarkasti jotain jakaumaa ja kun aineisto sisältää oudokkeja.

Työn lähteinä on käytetty tilastotieteen kirjallisuutta ja internetsivustoja. Motivaatio työn tekemiseen saatiin Maronnan et al. kirjottamasta kirjasta *Robust statistics : theory and methods (with R)*[5] ja varsinkin sen jaksosta 2. Lisätietoja robustiseen tilastotieteeseen ja varsinkin M-esimaattoreihin hankittiin Bergerin ja Casellan kirjasta *Statistical inference*[1] jaksosta 10.2.2.

2 Jakauman keskikohdan estimointiongelma

Tuntemattoman jakauman keskikohtaa tai keskimmäistä arvoa kuvataan käyttämällä keskilukuja, joista tärkeimmät ovat odotusarvo ja mediaani. Mediaani on määritelmänsä mukaan jakauman keskimmäinen arvo ja tämän takia se on määritelty hyvin kaikille jakaumille. Joillain jakaumilla ei välttämättä ole odotusarvoa. Tästä esimerkkinä on työssä myöhemmin esiteltävä Cauchy-jakauma, jolla on paksut hännät. Tässä työssä tarkastellaan esimerkiksi jakaumia, jotka ovat lähellä jotakin normaalijaukaumaa ja symmetrisiä. Tällöin mediaani ja mahdollisesti olemassa oleva odotusarvo ovat likimain yhtä suuret.

Jakauman keskikohdan estimointi siitä poimitulla satunnaisotoksella on tärkeä ja klassinen tehtävä tilastollisessa päättelyssä. Otoskeskiarvo on tähän tehtävään todella paljon käytetty estimaattori. Esimerkiksi, kun oletetaan jakauman noudattavan jotain normaalijakaumaa $N(\mu, \sigma^2)$, tilastollisen päättelyn kursseilla on opittu, että otoskeskiarvo on suurimman uskottavuuden estimaattori μ:lle. Tämä on myös todistettu DeGrootin ja Schervishin kirjassa jaksossa 7.5. [4, s. 420-421] Toinen luonteva estimaattori jakauman keskikohdan estimointiin on otosmediaani. Tässä työssä tutkitaan otoksen mediaanin ja otoskeskiarvon ominaisuuksia ja käyttäytymistä tilanteissa, joissa taustalla oleva jakauma poikkeaa jotenkin normaalijakaumasta.

2.1 Empiirinen esimerkki estimointiongelmasta

Tarkastellaan seuraavaksi 16 luvun aineistoa, jossa 15 lukua on satunnaisesti otettu normaalijakaumasta $N(0, 1)$ ja yksi luku normaalijakaumasta $N(20, 1)$. Kyseisellä aineistolla voidaan nähdä onko yhdellä oudokilla vaikutusta otoskeskiarvoon, kun muut havainnot noudattavat samaa normaalijakaumaa. Tutkitaan sen lisäksi myös aineistosta lasketun mediaanin käyttäytymistä ja verrataan sitä otoskeskiarvoon.

Taulukossa 1 esitellään esimerkkiaineisto, jossa on 16 lukua. Havainnoista 15 on otettu satunnaisesti jakaumasta $N(0, 1)$ ja yksi havainto jakaumasta $N(20, 1)$:

Huomataan, että arvo 20.04 eroaa huomattavasti muista aineiston arvoista. Tämä arvo on aineiston oudokki. Aineiston otoskeskiarvo $\overline{x} = 1.24$. Tämä arvo on suu-

-1.79	-1.78	-1.70	-1.14	-0.48	-0.31	-0.11	-0.05
0.07	0.49	0.90	0.97	1.28	1.31	2.18	20.04

Taulukko 1: Esimerkkiaineisto, jossa on 16 lukua.

rempi kuin 12 aineiston arvoista, joten otoskeskiarvo ei ole hyvä estimaatti jakauman keskikohdalle. Jos aineistosta poistetaan oudokki, on otoskeskiarvo $\overline{x} = -0.01$. Huomataan, että oudokki vaikuttaa merkittävästi otoskeskiarvoon, jonka arvo lähestyy oudokin arvoa.

Lasketaan aineistoille tyypilliset tunnusluvut otosvarianssi ja otoskeskihajonta. Huomataan, että otoskeskihajontaa ja otosvarianssia laskettaessa jakajana toimii otoskoon n sijasta tekijä $n-1$. Aineiston, jossa on oudokki, otosvarianssi $s^2 \approx 26.5$ ja ilman oudokkia otosvarianssi $s^2 \approx 1.5$. Vastaavasti otoskeskihajonnat ovat oudokilla $s \approx 5.2$ ja ilman $s \approx 1.2$. Huomataan, että otoskeskihajonta on huomattavasti suurempi, kun aineistossa on oudokki. Tämä vaikuttaa tietenkin myös otosvarianssiin, koska otosvarianssi on otoskeskihajonnan toinen potenssi.

Sama vaikutus huomataan laskemalla aineistoista 95 prosentin luottamusväli aineistojen odotusarvolle. Luottamusväli on laskettu Studentin t-jakaumasta kaavalla $\overline{x} \pm c \cdot \frac{s}{\sqrt{n}}$, jossa $\overline{x}$ on otoskeskiarvo, n on otoskoko ja c on t-jakauman arvo 0.025-yläkvantiili, kun vapausasteluku on $n-1$. Huomataan, että aineistoissa on 15 ja 16 havaintoa, joten käytetään c:n arvoja 2.145 ja 2.131.[6] Aineiston, jossa on oudokki, luottamusväli on $(-1.50, 3.99)$. Kun oudokki poistetaan luottamusväli on $(-0.69, 0.67)$. Huomataan, että oudokki leventää luottamusväliä huomattavasti ja epäsymmetrisesti.

Tarkastellaan seuraavaksi kuinka suuri vaikutus yhdellä oudokilla voi olla teoreettisesti keskiarvoon. Vaihdetaan oudokki 20.04 johonkin arvoon x. Otoskeskiarvon määritelmän $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ avulla voidaan huomata, kun x annetaan käydä kaikki luvut välillä $(-\infty, \infty)$, että sen arvot vaihtelevat välillä $(-\infty, \infty)$. Tätä kutsutaan rajoittamattomaksi vaikutukseksi.

Toinen yleinen menetelmä jakauman keskikohdan estimoimiseen on käyttää mediaania. Muistetaan, että otosmediaani lasketaan kahdella hieman eri tavalla riippuen onko aineistossa parillinen vai pariton määrä havaintoja. Aineisto järjestetään suuruusjärjestykseen ja merkitään järjestysluvuin alaindeksiin. Jos aineistossa x on n kappaletta havaintoja ja n on pariton on aineiston x mediaani havainto $x_{(n+1)/2}$.[7] Jos ainestossa x on parillinen määrä havaintoja, on mediaani kahden keskimmäisen arvon keskiarvo ja se lasketaan

$$\frac{x_{(n/2)} + x_{((n/2)+1)}}{2}. \tag{1}$$

Tutkitaan miten mediaani käyttäytyy taulukon 1 aineistossa. Huomataan, että aineistossa on parillinen määrä havaintoja ja oudokin poiston jälkeen pariton määrä havaintoja.

Huomataan, että taulukon 1 aineiston mediaani on 0.01 ja oudokin poistamisen jälkeen aineiston mediaani on -0.05. Toisin sanoen oudokin olemassaolo ei juurikaan vaikuta mediaanin arvoon. Lisäksi huomataan, että keskiarvo on likimäärin sama kuin mediaani, kun aineistosta poistetaan oudokki.

Tutkitaan seuraavaksi onko oudokin arvolla vaikutusta mediaanin arvoon. Vaihdetaan oudokin 20.04 arvo mielivaltaiseen lukuun x. Huomataan, kun annetaan luvun x käydä arvot välillä $(-\infty, \infty)$, että mediaanin arvo ei vaihtele välillä $(-\infty, \infty)$ kuten keskiarvo vaihteli samassa tilanteessa. Kun x lähestyy $-\infty$, vaihtuu mediaani arvosta 0.01 arvoon -0.08. Tällöin mediaanin arvo on lukujen $x_{(8)} = -0.11$ ja $x_{(9)} = -0.05$ keskiarvo. Aineiston kaikkien lukujen järjestystunnusluvut kasvavat yhdellä, koska oudokin järjestystunnusluku vaihtuu viimeisestä ensimmäiseksi.

Taulukosta 2 huomataan, että aineistossa yksikin oudokki vaikuttaa suuresti keskiarvoon mutta ei juurikaan mediaaniin. Tämä johtuu siitä, että mediaania laskettaessa ei oteta huomioon havaintojen arvoja vaan niiden järjestys. Yhden oudokin olemassaolo heikentää myös muiden tilastollisten tunnuslukujen kuten esimerkiksi varianssin ja keskihajonnan tarkkuutta suurentamalla niiden arvoa. Tämän takia olisi hyvä löytää estimaatti näille kaikille tunnusluvuille, jotka pystyvät estimoimaan niitä myös epäoptimaalisissa tilanteissa. Tässä työssä kuitenkin keskitytään etsimään robusti estimaatti jakauman keskikohdan estimoimiselle.

	Esimerkkiaineisto oudokilla	Esimerkkiaineisto ilman oudokkia
Keskiarvo	1.24	-0.01
Mediaani	0.01	-0.05
Varianssi	26.5	1.5

Taulukko 2: Esimerkkiaineistosta lasketut tunnusluvut oudokilla ja ilman.

3 Jakaumaseos ja Cauchy-jakauma

Tässä työssä simuloidaan keskiarvon ja mediaanin otosjakaumaa, kun otokset otetaan jakaumaseoksesta ja Cauchy-jakaumasta. Seuraavaksi esitellään mitä jakaumaseoksella ja Cauchy-jakaumalla tarkoitetaan.

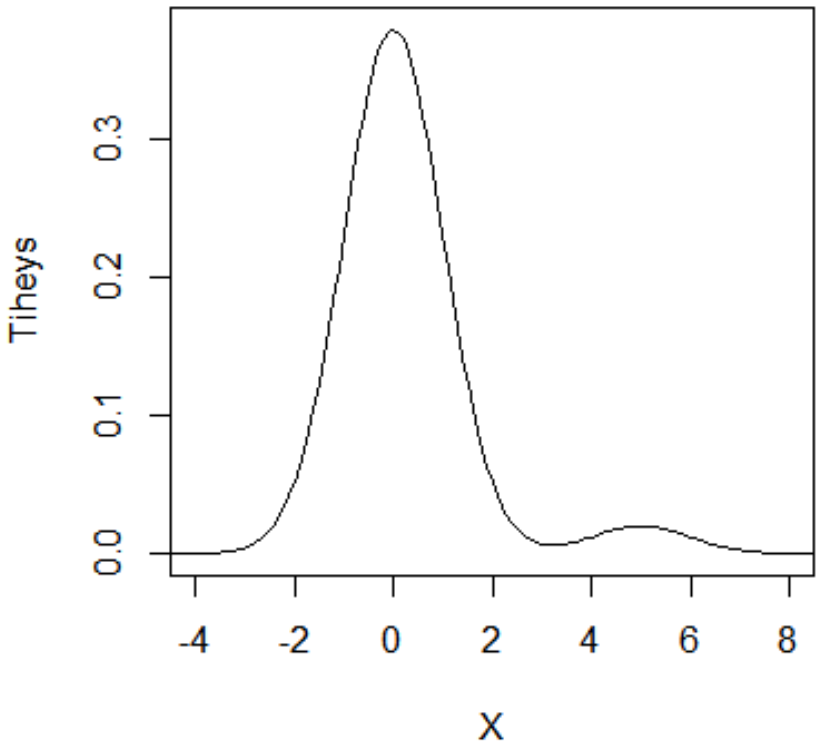
3.1 Jakaumaseos

Jakaumien sekoite voidaan esittää ajattelemalla, että havainnoista $1-\epsilon$:n suuruinen osuus noudattaa normaalijakaumaa ja ϵ:n suuruinen osuus noudattaa jotain tuntematonta mekanismia. Esimerkiksi jokin mittalaite tuottaa oikeita mittaustuloksia 95 % ajasta ja 5 % ajasta virheellisiä tuloksia. Tämä voidaan esittää jakaumana F:

$$F = (1 - \epsilon)G + \epsilon H, \tag{2}$$

jossa G on normaalijakauma $N(\mu, \sigma^2)$ ja H voi olla mikä tahansa jakauma, esimerkiksi normaalijakauma suuremmalla varianssilla tai mahdollisesti eri odotusarvolla. Yleisesti jakaumaa F sanotaan jakaumien G ja H seokseksi. Jakaumaa F kutsutaan normaaliksi jakaumaseokseksi tai saastuneeksi normaalijakaumaksi, jos G ja H ovat kummatkin normaalijakautuneita.[5, s. 19-20]

Kuvassa 1 on jakaumaseos, johon havainto on poimittu 95 prosentin todennäköisyydellä jakaumasta $N(0, 1)$ ja 5 prosentin todennäköisyydellä jakaumasta $N(5, 1)$. Huomataan, että havaintoja kerääntyy kummankin käytetyn normaalijakauman odotusarvon ympärille.



Kuva 1: Jakaumaseos, jossa 95 % todennäköisyydellä havainto on poimittu jakaumasta $N(0,1)$ ja 5 % todennäköisyydellä jakaumasta $N(5,1)$

3.2 Cauchy-jakauma

Cauchy-jakauma edustaa äärimmäistä tapausta ja toimii vastaesimerkkinä monille tilastotieteessä vakiintuneille tuloksille. Esimerkiksi Cauchy-jakaumalla ei ole odotusarvoa. Cauchy-jakauman tiheysfunktio määritellään sijaintiparametrilla a ja skaalausparametrilla b seuraavasti:

$$f(x|a,b) = \frac{1}{\pi b[1+(x-a)^2/b^2]}, \quad -\infty < a < \infty, 0 < b. \tag{3}$$

ja tähän jakaumaan viitataan merkinnällä Cauchy(a,b). Tiheysfunktion standardimuoto saadaan sijoittamalla $a = 0$ ja $b = 1$. Sijaintiparametri a on Cauchy-jakauman mediaani.[2, s. 343] Kuvasta 2 huomataan, että Cauchy-jakauman huippu laskee ja sen hännistä tulee paksummat, kun skaalausparametrin b arvo kasvaa.

Todistetaan seuraavaksi DeGrootin ja Schervishin kirjan *Probability and Statistics* sivua 210 mukaillen miksi Cauchy-jakaumalla ei ole odotusarvoa.[4] Määritellään ensin odotusarvo. Oletetaan, että jatkuvalle satunnaismuuttujalle X kumpikin

integraaleista 4 ovat äärellisiä:

$$\int_0^\infty xf(x)dx, \int_{-\infty}^0 xf(x)dx. \tag{4}$$

Tällöin odotusarvo on määritelty satunnaismuuttujalle X seuraavasti:

$$E(X) = \int_{-\infty}^\infty xf(x)dx = \int_{-\infty}^0 xf(x)dx + \int_0^\infty xf(x)dx. \tag{5}$$

Jos kummatkin integraalit (4) ovat äärettömiä, niin $E(X)$ ei ole olemassa.

Tätä määritelmää hyväksikäyttäen voidaan todistaa, ettei Cauchy-jakaumalla ole odotusarvoa. Oletetaan, että satunnaismuuttuja X on standardisesti Cauchy-jakautunut ja sillä on tiheysfunktio

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad \text{kun} \quad -\infty < x < \infty. \tag{6}$$

Voidaan varmistaa, että tiheysfunktio f kelpaa tiheysfunktioksi eli tulos $\int_{-\infty}^\infty f(x)dx = 1$ seuraavalla tunnetulla tuloksella:

$$\int \frac{1}{1+x^2}dx = tan^{-1}(x), \quad \text{kun} \quad -\infty < x < \infty.$$

Kun tiheysfunktio $f(x)$ on (6), niin integraalit (4) ovat:

$$\int_0^\infty \frac{x}{\pi(1+x^2)}dx = \infty \quad \text{ja} \int_{-\infty}^0 \frac{x}{\pi(1+x^2)}dx = -\infty,$$

joten odotusarvoa $E(X)$ ei ole olemassa Cauchy-jakaumalle.

Lisäksi Cauchy-jakaumalla on joitain erikoisominaisuuksia ja niistä muutama seuraavaksi. Jos X ja Y ovat riippumattomia standardinormaalisia satunnaismuuttujia, niin silloin pätee $X/Y \sim$ Cauchy(0,1). Jakauma Cauchy(0,1) on Studentin t -jakauman erikoistapaus vapausasteella $df = 1$.[2, s. 345]

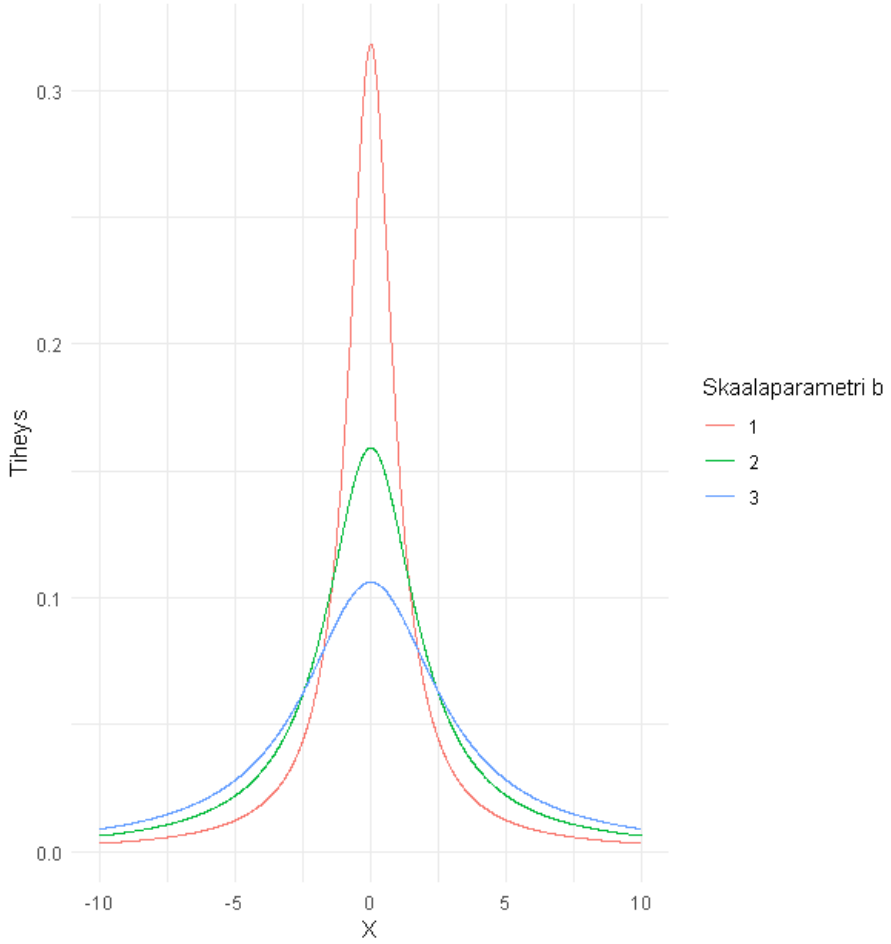
4 Otoskeskiarvon ja mediaanin otosjakauman simulointia

4.1 Jakaumaseos

Tässä jaksossa simuloidaan otoskeskiarvon ja mediaanin jakaumaa, kun otokset otetaan jakaumaseoksesta. Simuloinnit on suoritettu R-kielellä. Simuloinnissa käytetyt R-koodit on saatavilla liittessä 1.

4.1.1 Otoskeskiarvo

Tutkitaan, minkälainen vaikutus jakaumaseoksesta poimimisella tai otoksen koolla on otoskeskiarvon otosjakaumaan. Simuloidaan otoskeskiarvon otosjakaumaa, kun poimitaan 10 ja 100 havainnon otos jakaumaseoksesta. Toistetaan näiden otosten



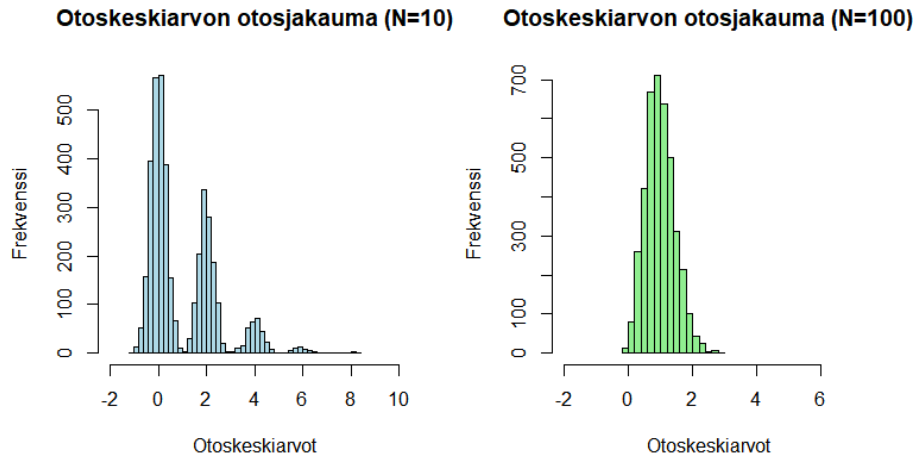
Kuva 2: Cauchy-jakauman tiheysfunktio eri parametrin b arvoilla, kun $a = 0$.

poimintaa jakaumaseoksesta 4000 kertaa ja lasketaan jokaiselle otokselle keskiarvo. Lopuksi tehdään näistä keskiarvoista histogrammi. Havainto otokseen poimitaan 95 prosentin todennäköisyydellä normaalijakaumasta $N(0, 1)$ ja 5 prosentin todennäköisyydellä normaalijakaumasta $N(20, 1)$.

Kuvasta 3 huomataan, että kummankin otoskoon histogrammit ovat vasemmalle vinoja. Lisäksi kuvasta 3 huomataan sama tulos kuin aikaisemmin esitellyn esimerkkiaineiston keskiarvon laskennassa. Otoskeskiarvon arvo lähestyy oudokin arvoa ja sen jakauma on vino kohti jakaumaseoksen toista normaalijakaumaa $N(20, 1)$. Vinous kuitenkin vähenee otoskoon kasvaessa, mikä huomataan kuvan 3 oikeanpuoleisesta histogrammista.

Lasketaan kuvan 3 kummallekin jakaumalle tunnusluvut otoskeskiarvo ja otosvarianssi. Otoskeskiarvo 10 otoksen otoskeskiarvojen jakaumalle on noin 1.0 ja otosvarianssi noin 2.0. Otoskeskiarvo 100 otoksen otoskeskiarvojen jakaumalle on myös noin 1.0, mutta varianssi on vain noin 0.2. Huomataan, että otoskoon kasvaessa keskiarvot keskittyvät uuden keskiarvon ympärille. Tämä otoskeskiarvo on siirtynyt jakauman $N(0, 1)$ odotusarvosta kohti jakauman $N(20, 1)$ odotusarvoa, vaikka kunkin havainnon poimintatodennäköisyys jakaumasta $N(20, 1)$ on 5 prosenttia.

Klassisissa normaalijakaumaoletukseen perustuvissa meneltelmissä käytetään otos-



Kuva 3: Otoskeskiarvon otosjakaumat jakaumaseoksesta, kun havainto on poimittu 95 %:n todennäköisyydellä jakaumasta $N(0,1)$ ja 5 %:n todennäköisyydellä jakaumasta $N(20,1)$. Otoskoot $n = 10$ ja $n = 100$. Otoksien lukumäärä on 4000.

keskiarvoa muodostettaessa esimerkiksi parametrien luottamusvälejä. Huomataan näistä tuloksista, että otoskeskiarvon otosjakauma jakaumaseoksesta poikkeaa jopa 100 havainnon otoksissa huomattavasti normaalijakaumasta $N(0,1)$. Näistä tuloksista huomataan otoskeskiarvon olevan epätarkka estimaattori odotusarvolle, kun havaintoja ei poimita normaalijakaumasta. Tilanne ei parane suurentamalla otoskokoa. Tämän takia ei ole suositeltavaa käyttää otoskeskiarvoa odotusarvon estimaattorina, kun havainnot poimitaan jakaumaseoksesta. Jos näin kuitenkin toimitaan, lasketut luottamusvälit voivat olla liian leveitä ja epäsymmetrisiä eikä niiden luotettavuudesta voida olla varmoja.

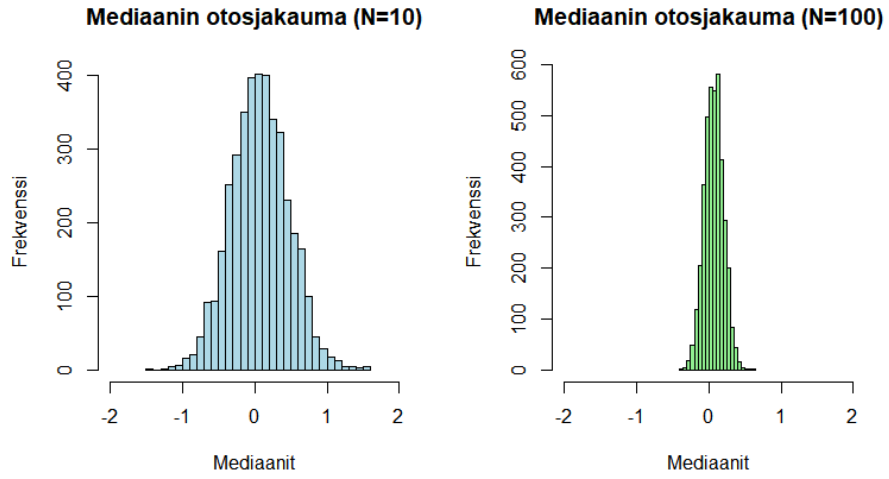
4.1.2 Mediaani

Simuloidaan mediaanin otosjakaumaa, kun otokset poimitaan samasta jakaumasta kuin jaksossa 4.1.1. Otoksia, joissa on 10 ja 100 havaintoa, poimitaan 4000 ja jokaisesta otoksesta lasketaan mediaani. Lopuksi tehdään näistä mediaaneista histogrammi.

Huomataan kuvasta 4, että suurin osa mediaaneista ovat kerääntyneet normaalijakauman $N(0,1)$ odotusarvon ympärille. Tästä voidaan päätellä, että mediaanin otosjakaumaan ei vaikuta jakaumaseoksen toinen normaalijakauma $N(20,1)$ yhtä paljon kuin otoskeskiarvon otosjakaumaan jaksossa 4.1.1. Otoskoon kasvaessa huomataan, että mediaanin otosjakauma keskittyy jakauman $N(0,1)$ odotusarvon ympärille.

Kuvan 4 kummankin jakauman otoskeskiarvo on noin 0.07. Huomataan, että mediaanin otosjakauman otoskeskiarvo on lähes yhtäsuuri kuin normaalijakauman $N(0,1)$ odotusarvo. Kun otoskoko on 10, niin mediaanin otosjakauman otosvarianssi on noin 0.16. Kun otoskoko kasvaa ja on 100, niin otosvarianssi on noin 0.02. Huomataan, että mediaanit keskittyvät otosjakaumansa otoskeskiarvon ympärille, kun otoskoko kasvaa. Mediaani vaikuttaa olevan robustimpi estimaattori jakauman

keskikohdalle tässä tilanteessa kuin keskiarvo.



Kuva 4: Mediaanin otosjakaumat jakaumaseoksesta, kun havainto on poimittu 95 %:n todennäköisyydellä jakaumasta $N(0,1)$ ja 5 %:n todennäköisyydellä jakaumasta $N(20,1)$. Otoskoot $n = 10$ ja $n = 100$. Otoksien lukumäärä on 4000.

4.2 Cauchy-jakauma

Tässä luvussa esitellään simuloinnin tulokset otoskeskiarvon ja mediaanin otosjakaumille, kun otokset poimitaan Cauchy-jakaumasta Cauchy(0,1). Simulointi on suoritettu R-kielellä. Simuloinnissa käytetyt R-koodit on saatavilla liitteessä 2.

4.2.1 Otoskeskiarvo

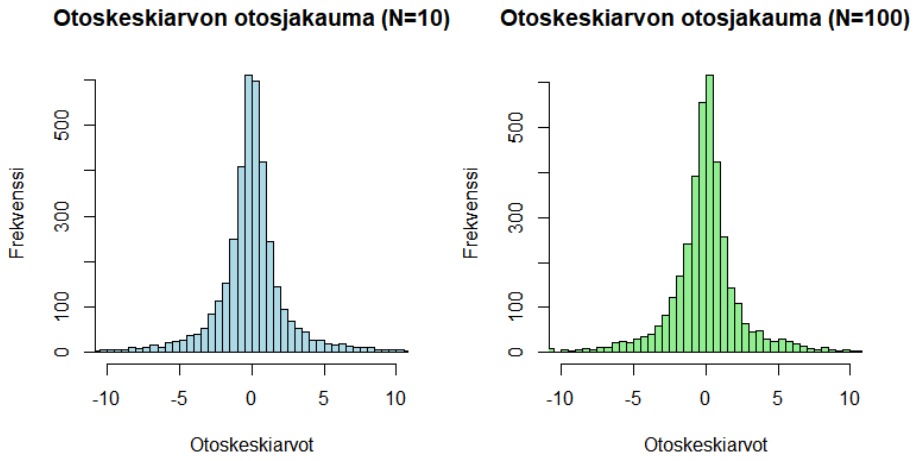
Simuloidaan otoskeskiarvon otosjakaumaa Cauchy-jakaumasta Cauchy(0,1). Poimitaan sekä 10 että 100 havainnon otoksia 4000 kappaletta ja lasketaan jokaisesta otoksesta keskiarvo. Lopuksi piirretään otoskeskiarvoista histogrammi.

Huomataan, että merkittävä osa otoskeskiarvoista on Cauchy(0,1)-jakauman sijaintiparametrin ympärillä kuvan 5 kummassakin kuvaajassa. Toisaalta kummassakin kuvaajassa otoskeskiarvoja on levittäytynyt laajalti poispäin kuvajan huipusta. Kuvaajista on jouduttu leikkaamaan pois äärimmäisiä arvoja, ettei niistä tulisi järjettömän leveitä. Huomataan, ettei otoskoon kasvattamisella ole mitään vaikutusta otoskeskiarvon otosjakauman muotoon. Mitään keskittymistä jonkin arvon ympärille ei tapahdu.

Kuten huomataan kuvan 5 histogrammeista, että otoskeskiarvot eivät otoskoon suurentuessa ole normaalisti jakautuneet keskeisen raja-arvolauseen mukaisesti. Syy tähän on se, että keskeinen raja-arvolause olettaa satunnaismuuttujilla, joista otoskeskiarvo lasketaan, olevan hyvin määritelty odotusarvo ja varianssi.[8] Kuten jaksossa 3.2 todistettiin, niin Cauchy-jakaumalla ei ole odotusarvoa. Tämän takia keskeinen raja-arvolauseella ei ole vaikutusta Cauchy-jakaumasta lasketuille otoskeskiarvoille. Lisäksi todennäköisyyslaskennan avulla voidaan osoittaa, että otoskes-

kiarvo noudattaa otoskoosta riippumatta samaa Cauchy-jakaumaa kuin yksittäinen havainto. [2, s. 279]

Lasketaan kuvan 5 jakaumille tunnusluvut otoskeskiarvo ja otosvarianssi. Otoskoon ollessa 10 otoskeskiarvo on noin 4.5. Toisaalta otoskoon ollessa 100 otoskeskiarvo on noin -1.8. Näistä huomataan, että kuvaajien ulkopuolella on huomattava määrä otoskeskiarvoja, jotka vaikuttavat jakaumien otoskeskiarvoihin siirtämällä niitä jakaumassa olevasta huipusta. Kummankin otoskoon jakauman otosvarianssi on noin 6000, joka erittäin suuri. Näistä syistä vaikuttaa siltä, että otoskeskiarvo ei ole käyttökelpoinen estimaatti otettaessa otoksia Cauchy-jakaumasta.



Kuva 5: Otoskeskiarvon otosjakaumat Cauchy(0,1)-jakaumasta, kun otoskoot ovat 10 ja 100. Otoksien lukumäärä on 4000. Äärimmäisiä arvoja on jätetty pois kuvaajista, jottei kuvaajista tulisi tavattoman leveitä.

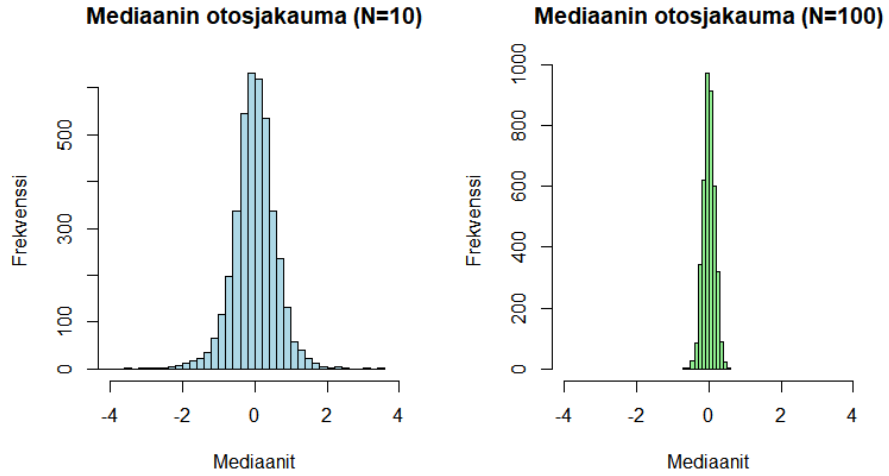
4.2.2 Mediaani

Simuloidaan mediaanin otosjakaumaa Cauchy(0,1)-jakaumasta. Poimitaan sekä 10 että 100 havainnon otoksia Cauchy-jakaumasta 4000 kappaletta ja lasketaan jokaisesta mediaani. Piirretään mediaaneista histogrammit.

Toisin kuin otoskeskiarvon otosjakauma, mediaanin otosjakauma vaikuttaa tiivistyvän jonkin arvon ympärille, mikä huomataan kuvan 6 kuvaajista. Kuvasta 6 päätellen mediaanin otosjakauma tiivistyy Cauchy(0,1)-jakauman sijaintiparametrin $a = 0$ ympärille. Lisäksi huomataan, että otoskoon kasvaessa mediaanit keskittyvät huipun ympärille eli niiden satunnaisvaihtelu pienenee.

Kummankin otoskoon jakaumassa mediaanin otoskeskiarvo on hyvin lähellä nollaa. Otoskoon ollessa 10 otoskeskiarvo on noin -0.002. Kun otoskoko on 100, otoskeskiarvo on sen sijaan noin -0.004. Mediaanin otosvarianssi toisaalta pienenee, kun otoskoko kasvaa. Otoskoon ollessa 10 otosvarianssi on noin 0.35. Kun otoskoko on 100, otosvarianssi on noin 0.03. Vaikuttaa näiden tunnuslukujen mukaan, että mediaanin otosjakauman keskiarvo lähenee Cauchy(0,1) sijaintiparametria a asymptoottisesti. Kun otoskoot ovat 10 ja 100, mediaanin otosjakauman otosvarianssi on nollan ja ykkösen välillä ja se pienenee otoskoon kasvaessa. Tämän takia vaikuttaa

siltä, että mediaani olisi tarkentuva estimaatti Cauchy-jakauman keskikohdalle tai keskimmäiselle arvolle.



Kuva 6: Mediaanin otosjakaumat Cauchy(0,1)-jakaumasta, kun otoskoot ovat 10 ja 100. Otoksien lukumäärä on 4000.

4.3 Simuloinnin yhteenveto

Simulointien tuloksista voidaan huomata, että mediaani on robustimpi estimaattori kuin otoskeskiarvo tutkituissa jakaumatilanteissa.

Taulukossa 3 on jakaumaseoksesta simuloitujen otosjakaumien tunnusluvut otoskeskiarvo ja otosvarianssi. Taulukon tuloksista huomataan, että otoskeskiarvoon vaikuttaa jakaumaseoksen $N(20, 1)$-jakauma enemmän kuin mediaaniin. Otoskeskiarvon otosjakauman simuloinnissa huomataan samanlaista käyttäytymistä kuin jakson 2 aineistossa, jossa oli oudokki. Otoskeskiarvoa simuloitaessa, sen otosjakauman otoskeskiarvo siirtyy kohti $N(20, 1)$-jakauman odotusarvoa. Mediaanin otoskeskiarvo ei siirry $N(0, 1)$-jakauman odotusarvosta juurikaan. Sekä otoskeskiarvoa ja mediaanin simuloitaessa huomataan, että otoskoon kasvaessa niiden otosvarianssi pienenee. Toisin sanoen ne keskittyvät otoskeskiarvonsa ympärille. Mediaani on tulosten perusteella tässä tapauksessa robustimpi estimaattori jakauman keskikohdan estimointiin.

	Otoskeskiarvo		Mediaani	
	$n = 10$	$n = 100$	$n = 10$	$n = 100$
Otoskeskiarvo	1.01	0.99	0.07	0.07
Otosvarianssi	2.02	0.20	0.16	0.02

Taulukko 3: Jakaumaseoksen simuloinnissa saatujen otosjakaumien tunnusluvut. Tunnusluvut on pyöristetty kahden desimaalin tarkkuudelle.

Taulukossa 4 on Cauchy(0,1)-jakaumasta simuloitujen otosjakaumien tunnusluvut otoskeskiarvo ja otosvarianssi. Huomataan, että mediaanin otosjakauman otoskeskiarvo vaikuttaa lähestyvän Cauchy(0,1)-jakauman sijaintiparametria a. Lisäksi

mediaanin otosvarianssi pienenee ja mediaanit keskittyvät otoskeskiarvon ympärille, kun otoskoko kasvaa. Keskiarvoa simuloitaessa huomataan, että sen otoskeskiarvo vaihtelee. Otoskeskiarvo vaihtelee positiivisen ja negatiivisen arvon välillä. Tämä vaikuttaa kertovan siitä, että keskiarvon otosjakaumassa on huomattava määrä äärimmäisiä arvoja, jotka vaikuttavat otoskeskiarvoon. Vaikuttaa myös, että näiden äärimmäisten arvojen määrä ja paikka vaihtelevat. Keskiarvon otosvarianssi on suuri ja se pysyy yhtä suurena otoskoon kasvaessa. Näistä tuloksista huomataan, että keskiarvo ei ole käyttökelpoinen estimaatti Cauchy-jakauman keskikohdan estimoimiseen. Mediaani on tähän robustimpi estimaatti kuin keskiarvo.

	Keskiarvo		Mediaani	
	$n = 10$	$n = 100$	$n = 10$	$n = 100$
Otoskeskiarvo	4.55	-1.79	0.00	0.00
Otosvarianssi	59527.95	6020.57	0.35	0.03

Taulukko 4: Cauchy(0,1)-jakauman simuloinnissa saatujen otosjakaumien tunnusluvut. Tunnusluvut on pyöristetty kahden desimaalin tarkkuudelle.

5 M-estimaattorit

Monet käytössä olevat estimaattorit ovat tuloksia jonkin kohdefunktion minimoinnista. Esimerkiksi keskiarvo ja mediaani ovat tällaisia. Kun $x_1, x_2, \ldots, x_n$ on poimittu aineisto, niin keskiarvo minimoi lausekkeen $\sum_{i=1}^{n}(x_i - \theta)^2$ ja mediaani lausekkeen $\sum_{i=1}^{n}|x_i - \theta|$. Keskiarvon tilanteessa funktio on neliöity, minkä takia se on herkkä vaihteluille, koska oudokit saavat suuren painoarvon. Mediaanin tilanteessa funktio on itseisarvoistettu, mikä johtaa siihen, että arvojen vaikutus estimaattoriin on lineaarista.[1, s. 484-485]

Tilastotieteilijä Peter Huber pyrki määrittelemään robustin estimaattorin käyttämällä keskiarvon ja mediaanin kompromissia. Hän määritteli kohdefunktion

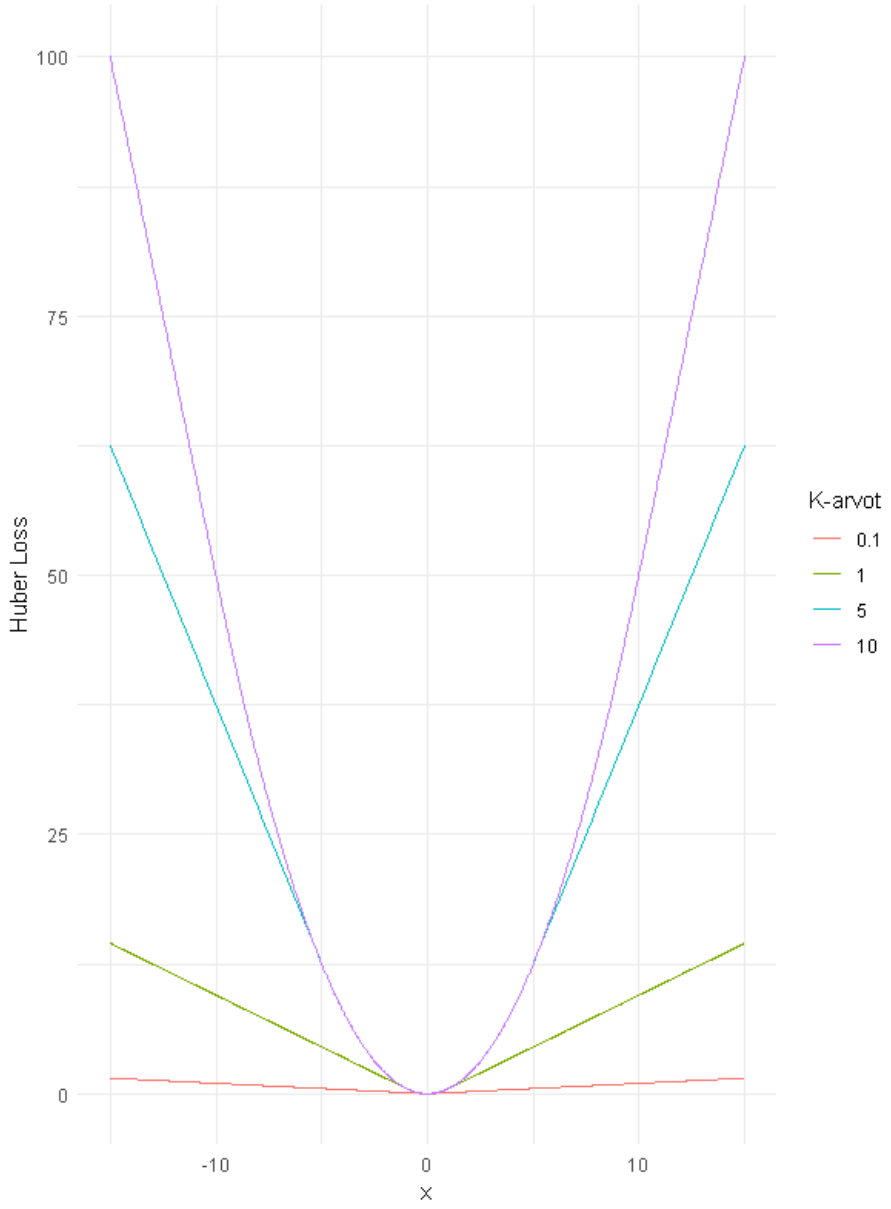
$$\sum_{i=1}^{n} \rho(x_i - \theta), \tag{7}$$

jossa ρ on

$$\rho(x) = \begin{cases} \frac{1}{2}x^2, & \text{kun } |x| \leq k \\ k|x| - \frac{1}{2}k^2, & \text{kun } |x| \geq k. \end{cases} \tag{8}$$

Funktio $\rho(x)$ käyttäytyy kuten funktio x^2 kaikilla $|x| \leq k$ ja kuten $|x|$ kaikilla $|x| > k$. Huomataan lisäksi, että funktio on jatkuva, koska $\frac{1}{2}k^2 = k|k| - \frac{1}{2}k^2$. Tämän lisäksi ρ on derivoituva. Vakiota k kutsutaan säätöparametriksi, joka säätää funktion ρ suhdetta. Pienet vakion k arvot tuottavat mediaanin kaltaisen estimaattorin. Tämä huomataan kuvan 7 kuvaajasta. Arvon k lähestyessä nollaa suurempi osa funktiosta $\rho(x)$ on lineaarista. Estimaattoria, joka minimoi funktion (7) ja (8), kutsutaan Huber-estimaattoriksi.

Tutkitaan miten Huber-estimaattori käyttäytyy esimerkkiaineistossa. Lasketaan Huber-estimaatti taulukon 1 aineistolle, kun säätöparametri k saa kokonaislukuarvot



Kuva 7: Kuvaajaa Huberin $\rho(x)$ funktiosta neljällä eri arvolla k.

väliltä $[1, 10]$. Huomataan taulukosta 5, että Huber-estimaatin arvot ovat nollan ja yhden välillä. Kun aineistosta laskettiin otoskeskiarvo, se oli 1.24 ja kaikki lasketut Huber-estimaatit olivat lähempänä nollaa. Tämä kertoo siitä, että aineistossa oleva oudokki ei vaikuta Huber-estimaatin arvoon yhtä paljon kuin keskiarvoon. Huber-estimaattori on robustimpi kuin keskiarvo. Toisaalta tässä tilanteessa mediaani on robustein, koska sen arvo oli tässä aineistossa 0.01.

k	1	2	3	4	5	6	7	8	9	10
Huber-estimaatti	0.12	0.20	0.30	0.41	0.51	0.61	0.72	0.82	0.93	1.03

Taulukko 5: Taulukon 1 aineistosta lasketut Huber-estimaatit, kun k saa arvot väliltä $[1, 10]$.

Estimaattoria, joka minimoi $\sum_i \rho(x_i - \theta)$ yleisemmällä funktiolla ρ, kutsutaan M-estimaattoriksi. Huomataan, jos valitaan funktioksi x^2, niin estimaattori on keskiarvo. Toisaalta valinta $|x|$ tuottaa mediaanin. M-estimaattoreita erilaisilla ominaisuuksilla voidaan johtaa vaihtamalla minimoitavaa funktiota ρ. Funktion minimointi toteutetaan yleisesti ratkaisemalla derivaatan nollakohdat, kun funktio on derivoituva. Määrittelemällä $\psi = \rho'$ voidaan kirjoittaa M-estimaattori seuraavan yhtälön ratkaisuna:

$$\sum_{i=1}^{n} \psi(x_i - \theta) = 0. \tag{9}$$

Tässä työssä ei perehdytä muihin M-estimaattoreihin tai niiden teoriaan syvällisemmin. Lisätietoa Huber-estimaattorista on saatavilla esimerkiksi Bergerin ja Casellan kirjan *Statistical inference*[1] jaksossa 10.2. Muita M-estimaattoreita on esitelty Maronna et al. kirjan *Robust statistics : theory and methods (with R)*[5] jaksossa 3.2. Kirjassa on esitelty esimerkiksi winsorkeskiarvo ja leikattu keskiarvo. Syvällistä tietoa robustista tilastotieteestä ja M-estimaattoreista on saatavilla aihetta alunperin tutkineen Peter J. Huberin kirjan *Robust statistics* toisessa painoksessa.[3]

Viitteet

[1] G. Casella ja R. L. Berger, *Statistical inference*, 2. laitos. Duxbury, CA: Thomson Learning, 2001.

[2] K. Krishnamoorthy, *Handbook of Statistical Distributions with Applications*. Chapman Hall/CRC, 2006.

[3] P. Huber ja E. Ronchetti, *Robust Statistics*, 2. laitos. Wiley, 2009.

[4] M. H. Degroot ja M. J. Schervish, *Probability and Statistics*. Pearson, 2012.

[5] R. A. Maronna, *Robust statistics : theory and methods (with R)*, 2. laitos. Hoboken, New Jersey: Wiley, 2019.

[6] *Student's t-distribution*, Wikipedia, lokakuu 2019. url: `https://en.wikipedia.org/wiki/Student%27s_t-distribution` (viitattu 17. 10. 2023).

[7] *Median*, Wikipedia, huhtikuu 2020. url: https://en.wikipedia.org/wiki/Median (viitattu 17. 10. 2023).

[8] *Keskeinen raja-arvolause*, Wikipedia, elokuu 2023. url: https://fi.wikipedia.org/wiki/Keskeinen_raja-arvolause (viitattu 27. 11. 2023).

Liitteet

Liite 1: R-koodiliite, Keskiarvon ja mediaanin otosjakauman simulointi jakauma-seoksesta

```
set.seed(66)

# Parametrit
N <- 10      # Otoksen koko
M <- 4000    # Toistojen määrä

# Alustetaan tyhjä vektori keskiarvoille
keskiarvot <- numeric(M)
keskiarvot_100 <- numeric(M)  # Uusi vektori otoksen koolla
   100

# Alustetaan tyhjä vektori medianeille
medianit <- numeric(M)
medianit_100 <- numeric(M)  # Uusi vektori otoksen koolla 100

# Toistetaan otoksen ottaminen, keskiarvon ja medianin
   laskeminen M kertaa
for (i in 1:M) {
  # Alustetaan tyhjä otos-vektori
  otos <- numeric(N)
  otos_100 <- numeric(100)  # Uusi otosvektori otoskoolla 100

  # Generoidaan kunkin havainnon alkuperäinen jakauma
     satunnaisesti
  for (j in 1:N) {
    if (runif(1) <= 0.05) {
      otos[j] <- rnorm(1, mean = 20)
    } else {
      otos[j] <- rnorm(1)
    }
  }

  # Toinen otos koolla 100
  for (j in 1:100) {
    if (runif(1) <= 0.05) {
      otos_100[j] <- rnorm(1, mean = 20)
    } else {
      otos_100[j] <- rnorm(1)
```

```
    }
  }

  # Lasketaan otoksen keskiarvo ja tallennetaan se vektoriin
  keskiarvot[i] <- mean(otos)
  keskiarvot_100[i] <- mean(otos_100)

  # Lasketaan otoksen mediaani ja tallennetaan se vektoriin
  medianit[i] <- median(otos)
  medianit_100[i] <- median(otos_100)
}

# Piirretään histogrammit samaan kuvaan
par(mfrow=c(1,2))  # Asetetaan kaksi histogrammia samaan
   kuvaan

hist(keskiarvot, breaks = 40, col = "lightblue", main = "
   Keskiarvon␣otosjakauma␣(N=10)", ylab = "Frekvenssi", xlab =
    "Keskiarvot", xlim=c(-2, 10))
hist(keskiarvot_100, breaks = 20, col = "lightgreen", main = "
   Keskiarvon␣otosjakauma␣(N=100)", ylab = "Frekvenssi", xlab
   = "Keskiarvot", xlim=c(-2, 7))

#Lasketaan jakaumien tunnusluvut keskiarvoille

#Otoskeskiarvo
kaka10 <- mean(keskiarvot)
kaka100 <- mean(keskiarvot_100)
#Otosvarianssi
kavar10 <- var(keskiarvot)
kavar100 <- var(keskiarvot_100)

# Piirretään histogrammit mediaineille samaan kuvaan
par(mfrow=c(1,2))  # Asetetaan kaksi histogrammia samaan
   kuvaan

hist(medianit, breaks = 40, col = "lightblue", main = "
   Mediaanin␣otosjakauma␣(N=10)", ylab = "Frekvenssi", xlab =
   "Medianit", xlim=c(-2, 2))
hist(medianit_100, breaks = 15, col = "lightgreen", main = "
   Mediaanin␣otosjakauma␣(N=100)", ylab = "Frekvenssi", xlab =
    "Medianit", xlim=c(-2, 2))

#Lasketaan jakaumien tunnusluvut mediaaneille
#Otoskeskiarvo
medka10 <- mean(medianit)
medka100 <- mean(medianit_100)
#Otosvarianssi
medvar10 <- var(medianit)
```

```
medvar100 <- var(medianit_100)
```

Liite 2: R-koodiliite, Keskiarvon ja Mediaanin otosjakauman simulointi Cauchy(0,1)-jakaumasta

```
set.seed(66)

# Parametrit
N <- 10      # Otoksen koko
M <- 4000    # Toistojen määrä

# Alustetaan tyhjä vektori keskiarvoille
keskiarvotc <- numeric(M)
keskiarvot_100c <- numeric(M)  # Uusi vektori otoksen koolla
   100

# Alustetaan tyhjä vektori medianeille
medianitc <- numeric(M)
medianit_100c <- numeric(M)  # Uusi vektori otoksen koolla 100

# Toistetaan otoksen ottaminen, keskiarvon ja medianin
   laskeminen M kertaa
for (i in 1:M) {
  # Alustetaan tyhjä otos-vektori
  otosc <- numeric(N)
  otos_100c <- numeric(100)  # Uusi otos vektori koolla 100

  # Generoidaan kukin havainto Cauchy-jakaumasta satunnaisesti
  for (j in 1:N) {
    otosc[j] <- rcauchy(1, location = 0, scale = 1)
  }

  # Toinen otos koolla 100
  for (j in 1:100) {
    otos_100c[j] <- rcauchy(1, location = 0, scale = 1)
  }

  # Lasketaan otoksen keskiarvo ja tallennetaan se vektoriin
  keskiarvotc[i] <- mean(otosc)
  keskiarvot_100c[i] <- mean(otos_100c)

  # Lasketaan otoksen mediaani ja tallennetaan se vektoriin
  medianitc[i] <- median(otosc)
  medianit_100c[i] <- median(otos_100c)
}

# Piirretään histogrammit samaan kuvaan
par(mfrow=c(1,2))  # Asetetaan kaksi histogrammia samaan
   kuvaan
```

```
hist(keskiarvotc, breaks = 50000, col = "lightblue", main = "
   Keskiarvon␣otosjakauma␣(N=10)", ylab = "Frekvenssi", xlab =
    "Keskiarvot", xlim = c(-10,10))
hist(keskiarvot_100c, breaks = 15000, col = "lightgreen", main
    = "Keskiarvon␣otosjakauma␣(N=100)", ylab = "Frekvenssi",
   xlab = "Keskiarvot", xlim = c(-10,10))

# Piirretään histogrammit mediaineille samaan kuvaan
par(mfrow=c(1,2))  # Aseta kaksi histogrammia samaan kuvaan

hist(medianitc, breaks = 50, col = "lightblue", main = "
   Mediaanin␣otosjakauma␣(N=10)", ylab = "Frekvenssi", xlab =
   "Medianit", xlim = c(-4,4))
hist(medianit_100c, breaks = 10, col = "lightgreen", main = "
   Mediaanin␣otosjakauma␣(N=100)", ylab = "Frekvenssi", xlab =
    "Medianit", xlim = c(-4,4))

#Lasketaan jakaumien tunnusluvut keskiarvoille

#Otosmediaani
kaka10c <- mean(keskiarvotc)
kaka100c <- mean(keskiarvot_100c)
#Otosvarianssi
kavar10c <- var(keskiarvotc)
kavar100c <- var(keskiarvot_100c)

#Lasketaan jakaumien tunnusluvut mediaaneille

#Otoskeskiarvo
medka10c <- mean(medianitc)
medka100c <- mean(medianit_100c)
#Otosvarianssi
medvar10c <- var(medianitc)
medvar100c <- var(medianit_100c)
```