# 4B Parametrien estimointi

Merkinnöistä: Kun tiheysfunktion merkinnässä on alaindeksi, se voi eri tilanteissa tarkoittaa eri asioita (tämä pitää tarvittaessa selventää). Alaindeksi voi ilmaista, minkä satunnaismuuttujan tiheysfunktio on kyseessä, kuten $f_X$, $f_Y$, $f_{X,Y}$ jne. Tai se voi olla parametri, joka kertoo mikä tietyn jakaumaperheen jakaumista on kyseessä, esim. jos puhutaan eksponenttijakaumista, niin $f_5(x)$ voi tarkoittaa että kyseessä on eksponenttijakauma nimenomaan taajuusparametrilla 5.

Parametria merkitään myös muilla tavoilla, mm. ehdollista tiheysfunktiota vastaavalla pystyviivamerkinnällä $f(x \mid \lambda)$ tai puolipisteellä $f(x\,;\,\lambda)$. Vaihteleviin merkintöihin joutuu matematiikassa ja tilastotieteessä valitettavasti tottumaan.

Lyhenne $SU-estimaattori$ tarkoittaa suurimman uskottavuuden estimaattoria (engl. ML estimator, maximum likelihood estimator).

## Tuntitehtävät

**4B1** (Palvelupyyntöjen väliajat) Palvelimelle saapuvien palvelupyyntöjen väliajat (yksikkönä sekunti) ovat toisistaan riippumattomat ja noudattavat tiheysfunktiota

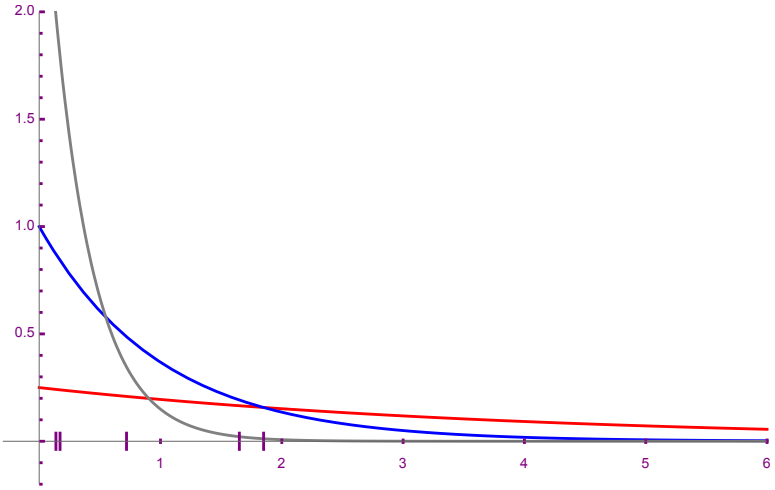
$$f_\lambda(x) \;=\; \begin{cases} \lambda e^{-\lambda x}, & x > 0, \\ 0, & x \le 0, \end{cases}$$

missä $\lambda > 0$ on tuntematon parametri. Palvelimen käynnistymisen jälkeen on mitattu väliajat $0.16, 1.85, 0.15, 0.72, 1.65$.

(a) Alla on hahmoteltu tiheysfunktion kuvaaja parametrin $\lambda$ arvoilla 0.25 (punainen), 1.00 (sininen) ja 3.00 (harmaa). Arvioi silmämääräisesti, mikä arvoista sopisi parhaiten havaintoihin (merkitty vaaka-akselille).

(b) Estimoi parametri $\lambda$ suurimman uskottavuuden menetelmällä.

(**Vihje:** Uskottavuusfunktio $L(\lambda)$ maksimoituu samassa pisteessä, missä logaritminen uskottavuusfunktio $\ell(\lambda) = \log(L(\lambda))$. Jälkimmäistä saattaa olla mukavampi derivoida.)



**4B2** (Sarjanumerot) Vihollisen panssarivaunuissa on sarjanumerot $1, 2, \ldots, b$. Tiedustelijamme ovat tehneet neljä vaunuhavaintoa ja nähneet sarjanumerot $x_1 = 13$, $x_2 = 77$, $x_3 = 111$ ja $x_4 = 145$. Oletamme, että kukin havainto tapahtui satunnaisesti, diskreetin tasajakauman mukaisesti kaikkien sarjanumeroiden joukosta. Parametri $b$ on tuntematon. (Ks. luento 4A.)

(a) Päättele havaintojen perusteella, onko mahdollista, että $b = 140$? Entä voiko olla $b = 200$?

(b) Jos vihollisella on $b$ vaunua ja $b < 145$, mikä on todennäköisyys havaita juuri tämä neljän sarjanumeron jono?

(c) Jos vihollisella on $b$ vaunua ja $b \geq 145$, mikä on todennäköisyys havaita juuri tämä neljän sarjanumeron jono? (Vastaus on jokin $b$:tä sisältävä lauseke.)

(d) Kirjoita uskottavuusfunktio $L(b)$ lausekkeena, joka on pätevä kaikilla positiivisilla kokonaisluvuilla $b$. (Vihje: Jaa tapauksiin.)

(e) Tutki lauseketta ja etsi se $b$:n arvo, jolla $L(b)$ on suurimmillaan. Toisin sanoen etsi suurimman uskottavuuden estimaatti $\hat{b}$.

(f) Yleistä edelliset havainnot: Mikä on SU-estimaatti $\hat{b}(\vec{x})$ jos olemme nähneet $n$ vaunua $(x_1, \ldots, x_n)$?

(g) Jos olemme nähneet vain yhden vaunun ($n = 1$), jonka sarjanumero on $x_1$, mikä on edellisen kohdan mukaisesti $b$:n SU-estimaatti? Onko se mielestäsi hyvä estimaatti?

## Kotitehtävät

**4B3** (Jatkuva tasajakauma) Datalähteenä on jatkuva tasajakauma välillä $[0, b]$, tiheysfunktiolla

$$f_b(x) = \begin{cases} \frac{1}{b}, & 0 \leq x \leq b, \\ 0, & \text{muuten.} \end{cases}$$

Parametri $b$ on tuntematon positiivinen reaaliluku.

(a) Datalähteestä on saatu viisi lukua $(1.3, 1.9, 3.6, 1.1, 5.1)$. Kirjoita uskottavuusfunktio $L(b)$ (vihje: jaa tapauksiin). Piirrä funktio käsin tai tietokoneella esim. välillä $b \in [1, 10]$. Selitä sanallisesti, millainen on funktion muoto ja miksi. Vihje: Tehtävä muistuttaa panssarivaunuongelmaa, mutta nyt parametri ja data eivät ole kokonaislukuja.

(b) Määritä datan perusteella SU-estimaatti $\hat{b}$.

(c) Yleistä mihin tahansa datajoukkoon: Jos on saatu luvut $\vec{x} = (x_1, x_2, \ldots, x_n)$, mikä on parametrin $b$ SU-estimaatti?

(d) Olkoon parametrilla $b$ eräs arvo (jota emme tiedä). Oletetaan, että havaitsemme vain yhden datapisteen. Pidetään sitä satunnaismuuttujana $X_1$, joka noudattaa tasajakaumaa välillä $[0, b]$. Mikä on sen odotusarvo? Mikä on SU-estimaattorin $b(X_1)$ odotusarvo? Onko SU-estimaattori harhaton vai harhainen, ja jos harhainen niin mihin suuntaan?

(e) Kokonaan toisenlainen estimaattori (joka ei ole SU-estimaattori) voidaan muodostaa seuraavasti. Generoivan jakauman odotusarvo on $\mu = b/2$. Jos käytämme datan keskiarvoa $m(\vec{x})$ estimoimaan $\mu$:ta, niin tuntuisi luontevalta, että $2m(\vec{x})$ olisi hyvä estimaattori suureelle $2\mu = b$. Määritellään siis uusi estimaattori

$$\tilde{b}(\vec{x}) = 2m(\vec{x}) = \frac{2}{n}\sum_{i=1}^{n} x_i.$$

Tutki onko uusi estimaattorimme $\tilde{b}(\vec{X})$ harhainen vai harhaton, kun kukin havainto $X_i$ tulee tasajakaumasta välillä $[0, b]$. (Vihje: Odotusarvo termeittäin.) Vaikuttaako uusi estimaattori järkevältä?

(f) Laske e-kohdan mukainen estimaatti $b$:lle, kun data on $\vec{x} = (2, 3, 16)$. Onko estimaatti mielestäsi järkevä?

**4B4** (Geometrisen jakauman sovittaminen) Satunnaismuuttujalla $X$ on geometrinen jakauma parametrilla $p$, tiheysfunktiolla

$$f_p(x) = \begin{cases} p(1-p)^x, & x = 0, 1, 2, \ldots \\ 0, & \text{muuten.} \end{cases}$$

Tulkinta: $X$ saadaan kun koetta (esim. nopanheittoa tai roskan heittämistä roskakoriin), joka onnistuu tn:llä $p$, toistetaan kunnes koe onnistuu, ja lasketaan sitä edeltäneiden epäonnistumisten lukumäärä. Yhdestä tällaisesta koesarjasta siis tulee tulokseksi vain *yksi* satunnaisluku, ja sillä on geometrinen jakauma.

Tästä jakaumasta on havaittu riippumattomasti datapisteet $x_1 = 5$, $x_2 = 2$ ja $x_3 = 0$. Määritä suurimman uskottavuuden estimaatti jakauman parametrille $p$.

Toistokoetulkinta: Roskakoriin on heitetty kolme roskaa, kukin omana sarjanaan. Ensimmäinen roska saatiin koriin $x_1$:llä hukkaheitolla (ja yhdellä onnistuneella). Toiseen roskaan meni $x_2$ hukkaheittoa ja kolmanteen $x_3$ hukkaheittoa. Emme tarkastele yksittäisiä heittoja, vaan satunnaismuuttujia $X_1$ "ensimmäisen sarjan hukkaheittojen määrä", ja $X_2$, $X_3$ vastaavasti. Nämä kolme lukua ovat geometrisesti jakautuneita. Parametri $p$ kuvaa, millä todennäköisyydellä kukin yksittäinen heitto onnistuu.

Vihje. Muodosta ensin uskottavuusfunktio. Seuraavaksi kannattaa ehkä ottaa logaritmi, jotta maksimoiminen on helpompaa (voit kyllä yrittää ilmankin). Jos et muista, miten logaritmia ja yhdistettyä funktiota derivoidaan, palauta mieleen.