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

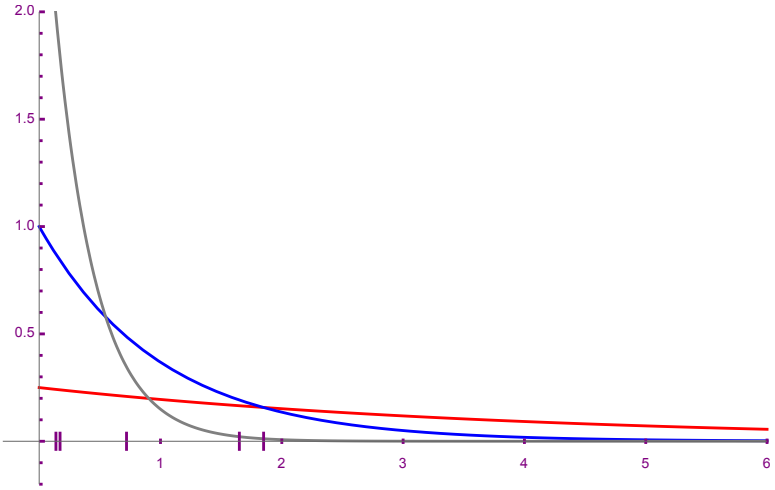
$$f_\lambda(x) \;=\; \begin{cases} \lambda e^{-\lambda x}, & x > 0, \\ 0, & x \le 0, \end{cases}$$

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 λ arvoilla 0.25 (punainen), 1.00 (sininen) ja 3.00 (harmaa). Arvioi silmämääräisesti, mikä arvoista sopisi parhaiten havaintoihin (merkitty vaaka-akselille).

(b) Estimoi parametri λ suurimman uskottavuuden menetelmällä.

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



Ratkaisu.

(a) Yksi tapa ajatella: "sopiva" tiheysfunktio vastaa pisteiden empiiristä jakaumaa, joten sen pitäisi olla korkea sillä, missä on paljon pisteitä, ja matala siellä missä pisteitä on vähän tai ei ollenkaan. Datapisteistä kolme viidestä osuu välille $[0, 1]$, joten tiheysfunktion alle jäävän alueen välillä [0,1] pitää olla kohtuullisen kokoinen (mieluiten noin 3/5). Datapisteistä kaksi viidestä osuu välille $[1.5, 2]$, joten tiheysfunktion alle jäävän alueen välillä $[1.5, 2]$ pitää pitää myös olla kohtuullisen kokoinen. Kuvasta katsottuna sininen käyrä ($\lambda = 1.00$) näyttäisi olevan sopusoinnussa näiden havaintojen kanssa.

Punainen käyrä ($\lambda = 0.25$) sopii dataan huonommin kuin sininen, koska se on suunnilleen kaikissa havaituissa pisteissä alempana. (Sen sijaan se on korekalla kaukana oikealla, missä ei ole havaittuja pisteitä.)

Harmaa ($\lambda = 3.00$) tuntuisi keskittyvän liikaa origon lähelle; kaksi suurinta havaittua datapistettä sopivat huonosti harmaaseen jakaumaan.

(b) Merkitään havaittuja väliaikoja $x_1, \ldots, x_5$. Havaittuun datajoukkoon liittyvä uskottavuusfunktio on

$$L(\lambda) \;=\; \prod_{i=1}^{5} \lambda e^{-\lambda x_i} \;=\; \lambda^5 e^{-\lambda(x_1+\cdots+x_5)}$$

Uskottavuusfunktion $L(\lambda)$ maksimi löytyy samasta pisteestä kuin logaritmisen uskottavuusfunktion $\ell(\lambda) = \log L(\lambda)$. Logaritminen uskottavuusfunktio on

$$\ell(\lambda) \;=\; \log\left(\lambda^5 e^{-\lambda(x_1+\cdots+x_5)}\right) \;=\; 5\log(\lambda) - \lambda(x_1 + \cdots + x_5),$$

ja kaksi ensimmäistä derivaattaa ovat

$$\ell'(\lambda) \;=\; 5\lambda^{-1} - (x_1 + \cdots + x_5).$$

ja

$$\ell''(\lambda) \;=\; -5\lambda^{-2}.$$

Derivaatan ainoa nollakohta on

$$\lambda \;=\; \frac{5}{x_1 + \cdots + x_5}.$$

Koska $\ell''(\lambda) \le 0$ kaikilla $\lambda > 0$, löytyy logaritmisen uskottavuusfunktion maksimi derivaatan nollakohdasta. Tämä nollakohta on myös uskottavuusfunktion $L(\lambda)$ maksimikohta. Näin ollen suurimman uskottavuuden estimaatti on

$$\hat{\lambda}(\vec{x}) \;=\; \frac{5}{x_1 + \cdots + x_5} \;=\; \frac{5}{0.16 + 1.85 + 0.15 + 0.72 + 1.65} \;\approx\; 1.104.$$

Huom. Yleiselle n alkion datajoukolle $\vec{x} = (x_1, \ldots, x_n)$ sama laskelma tuottaa tulokseksi $\hat{\lambda}(\vec{x}) = 1/m(\vec{x})$, missä $m(\vec{x}) = \frac{1}{n}(x_1 + \cdots + x_n)$. Tämä siis pätee *eksponenttijakaumalle*. Jollekin muulle jakaumalle samankaltainen tulos voi päteä tai olla pätemättä (kuten nähdään tehtävässä 4B2).

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?

Ratkaisu.

(a) Koska olemme nähneet vaunun numero 145, varmastikaan b ei voi olla sitä pienempi, esim. 140. Se voi kyllä olla 200.

(b) Jos $b < 145$, niin on mahdotonta nähdä vaunua 145, joten todennäköisyys on nolla.

(c) Joka kerta kun nähdään vaunu, sen sarjanumero on diskreetin tasajakauman mukaisesti jokin luvuista $1, \ldots, b$, jolloin kunkin luvun tn on $1/b$. Erityisesti $P(x_1 = 13) = P(x_2 = 77) = P(x_3 = 111) = P(x_4 = 145) = 1/b$. (Huom. oletus $b \geq 145$, joten nämä neljä lukua ovat joukossa $\{1, \ldots, b\}$ eli mahdollisia havaintoja.) Siten havaitun numerojonon tn on $1/b^4$.

(d)
$$L(b) = \begin{cases} 0 & \text{jos } b < 145 \\ 1/b^4 & \text{jos } b \geq 145 \end{cases}$$

(e) Edellisen kohdan lausekkeesta nähdään suoraan, että $L(b) = 0$ kun $b < 145$, joten maksimi ei voi olla siellä, vaan jossain alueella $b \geq 145$. Lausekkeesta myös nähdään, että b:n kasvaessa $L(b)$ pienenee, joten suurin arvo on kohdassa $b = 145$. (Voit piirtää kuvan nähdäksesi L:n muodon.) SU-estimaatti on siis $\hat{b} = 145$.

(f) Olkoon $M = \max\{x_1, \ldots, x_n\}$ suurin näkemistämme sarjanumeroista. Jos $b < M$, niin $L(b) = 0$ koska emme voi nähdä numeroa M jos vihollisella on vain b vaunua.

Jos $b \geq M$, niin $L(b) = 1/b^n$, joka on positiivinen mutta pienenee kun b kasvaa. Siten SU-estimaatti on $\hat{b} = M$.

(g) Tässä tapauksessa $\hat{b} = x_1$. Esimerkiksi jos näemme yhden vaunun ja sen sarjanumero on 130, niin SU-estimaattimme on, että vihollisella on juuri 130 vaunua, ja näkemämme vaunu on juuri se, jolla on suurin sarjanumero. Tämä vaikuttaa vähän oudolta: miksi satuimme näkemään juuri suurimman numeron?

Asiaa voi ajatella myös näin: *Jos* vihollisella on b vaunua, niin on melko varmaa (todennäköisyys $1 - 1/b$), että havaitsemme jonkin muun kuin suurimman numeron, joten SU-estimaattimme on suurella todennäköisyydellä liian pieni ($\hat{b} < b$).

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 μ: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?

Ratkaisu.

(a) Uskottavuusfunktio on $L(b) = f_b(x_1)f_b(x_2)\dots f_b(x_5)$, eli datapisteiden tiheyksien tulo. Olkoon $M = \max\{x_1, \dots, x_n\}$ eli suurin datapisteistä. Jos $M > b$, niin ainakin yksi datapisteistä (nimittäin M) on välin $[0, b]$ ulkopuolella, joten $L(b) = 0$. Jos $M \le b$, niin kaikki n pistettä ovat välillä $[0, b]$, tiheydet ovat kukin $1/b$ joten $L(b) = (1/b)^n$.

Annetussa datassa on $n = 5$ ja $M = 5.1$, joten

$$L(b) = \begin{cases} 0 & \text{kun } b < 5.1 \\ 1/b^5 & \text{kun } b \ge 5.1 \end{cases}$$

Sanallinen kuvaus: Uskottavuus on nolla parametriarvoille $b < 5.1$ (ne ovat mahdottomia), ja hyppää sitten arvoon $1/5.1^5$ kun $b = 5.1$. Sen jälkeen se laskee kun b kasvaa (koska tasajakauman levetessä tiheysfunktion arvot pienenevät).

(b) Edellisen kohdan perusteella on selvää, että $L(b)$ saa suurimman arvonsa pisteessä $b = M = 5.1$. (Jos haluat, voit tutkia L:n derivaattaa, mutta päätelmään riittää jo sen huomaaminen, että $1/b^n$ on b:n suhteen laskeva funktio.)

(c) Kuten edellä, $L(b) = 0$ kun $b < M$, ja $L(b) = 1/b^n$ kun $b \ge M$. Siis $\hat{b}(\vec{x}) = M = \max\{x_1, \dots, x_n\}$. Toisin sanoen b:n SU-estimaatti on yksinkertaisesti suurin havaituista datapisteistä.

(d) Koska X_1 on tasajakautunut välillä $[0, b]$, niin $\mathrm{E}(X_1) = b/2$. Jos $n = 1$ niin SU-estimaattorimme on yksinkertaisesti $\hat{b}(X_1) = X_1$, joten $\mathrm{E}(\hat{b}(X_1)) = \mathrm{E}(X_1) = b/2$. Estimaattorimme on pahasti alaspäin harhainen, koska $b/2 < b$.

(e) Lasketaan uuden estimaattorimme odotusarvo.

$$\begin{aligned} \mathrm{E}(\tilde{b}(\vec{X})) &= \mathrm{E}\left(\frac{2}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{2}{n} \cdot \sum_{i=1}^{n} \mathrm{E}(X_i) \\ &= \frac{2}{n} \cdot n \cdot \frac{b}{2} \\ &= b. \end{aligned}$$

Estimaattori on siis harhaton, mikä on mukavaa. Ainakin *tässä mielessä* estimaattori tuntuu järkevältä.

Tällaista estimaattoria kutsutaan *momenttiestimaattoriksi*. Ideana on, että ensin estimoidaan joitakin tuntemattoman jakauman ns. momentteja (tässä vain ns. ensimmäinen momentti eli odotusarvo). Sitten *valitaan* se jakaumaperheen (tässä tasajakaumien) jäsen, jolla on kyseiset momentit.

(f) $\tilde{b}(2,3,16) = \frac{2}{3}(2+3+16) = 14$. Estimaatin arvo on mahdoton, koska se on pienempi kuin eräs havaituista datapisteistä! Emme ole mitenkään voineet saada pistettä 16 välin $[0, 14]$ tasajakaumasta.

Tässä ja edellisessä tehtävässä havaittiin, että tasajakauman (diskreetin tai jatkuvan) päätepisteen estimoiminen ei vaikuta aivan helpolta tehtävältä. Ensimmäinen yrityksemme (SU-estimaattori) on alaspäin harhainen, mutta ei ainakaan anna mahdottomia arvoja. Toinen yrityksemme (momenttiestimaattori) on harhaton, mutta voi antaa mahdottomia arvoja. Tähän on kyllä olemassa ratkaisuja (ks. esim. Rossin kirjan Example 7.7c, ja Wikipedia: German tank problem). Eräs helpohko ratkaisu olisi ottaa SU-estimaattori ja *korjata* sitä ylöspäin sopivalla kertoimella, niin että harha poistuu!

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.

Ratkaisu. Uskottavuusfunktio on

$$\begin{aligned} L(p) &= f_p(5) \cdot f_p(2) \cdot f_p(0) \\ &= p(1-p)^5 \cdot p(1-p)^2 \cdot p(1-p)^0 \\ &= p^3(1-p)^7, \end{aligned}$$

ja sen logaritmi on

$$\ell(p) \;=\; \log L(p) \;=\; 3\log p + 7\log(1-p).$$

Logaritmin derivaatta on

$$\ell'(p) \;=\; \frac{3}{p} - \frac{7}{1-p}.$$

Kyseinen derivaatta on nolla, kun

$$\frac{3}{p} \;=\; \frac{7}{1-p} \qquad \text{eli} \qquad p \;=\; 3/10.$$

Ylläoleva arvo on logaritmisen uskottavuusfunktion maksimi, sillä toinen derivaatta $\ell''(p) = -\frac{3}{p^2} - \frac{7}{(1-p)^2} \leq 0$ kaikilla $p > 0$. Koska logaritmi säilyttää järjestyksen (= mitä suurempi on luku, sitä suurempi on sen logaritmikin), niin ylläoleva p on myös varsinaisen uskottavuusfunktion maksimi. Parametrin p suurimman uskottavuuden estimaatti on siis $3/10$.

Saatu arvo sopii sikälikin arkijärkeen, että jos kolmessa heittosarjassa tehtiin yhteensä $5+2+0 = 7$ hukkaheittoa ja $1 + 1 + 1 = 3$ onnistunutta heittoa, niin heittäjä onnistui kolme kertaa kymmenestä heitosta.