3 Estimointiteoriaa

- Tarkastelemme frekventististä tilastollista mallia eli jakaumaperhettä

$$\{f(\mathbf{y};\theta), \theta \in \Theta\}$$

  sekä aineistoa, jonka ajattelemme tulleen jostakin tähän perheeseen kuuluvasta jakaumasta.
- Menetelmiä, joilla tuntemattoman parametrin "todellista" arvoa voidaan arvioida eli estimoida. Tämä tarkoittaa sitä, että parametriavaruudesta valitaan yksi arvo $\hat{\theta}$, joka on (jonkin kriteerin mielessä) paras arvaus parametrin todelliselle arvolle.
- Ts. jakaumaperheestä valitaan estimaattia $\hat{\theta}$ vastaava jakauma $\mathbf{y} \mapsto f(\mathbf{y};\hat{\theta})$, joka on mielestämme paras arvaus sillä jakaumalle, joka havainnot tuotti.

Kylvän epäilyksen siemenen:

- Sana *todellinen* laitettiin edellä lainausmerkkeihin fraasissa parametrin "todellinen" arvo.
- Saattaa olla, että havainnot on tuottanut sellainen prosessi, jota analyysissä käyttämämme malli $f(\mathbf{y}; \theta)$ ei kuvaa hyvin.
- Kuuluisaa tilastotieteilijää George E. P. Boxia lainaten

  All models are wrong, but some are useful.

- Voimme olla aivan varmoja parametrisen mallin oikeellisuudesta vain harvoissa tapauksissa, kuten silloin, jos olemme aineiston simuloineet tietokoneella ko. parametrisesta mallista. Tällaisessa tapauksessa parametrin todellinen arvo on se arvo, jota käytettiin simuloinnissa.

3.1 Parametri ja tunnusluku

- Sanaa parametri voi tarkoittaa tilastotieteessä eri yhteyksissä eri asioita.
- Tähän asti sillä on tarkoitettu sitä parametrisessa mallissa $f(\mathbf{y}; \theta)$ esiintyvää lukua (tai luvuista koostuvaa vektoria) θ, jonka tunteminen kiinnittäisi havaintosatunnaisvektorin $\mathbf{Y}$ jakauman.
- Toisaalta sana parametri voi tarkoittaa mitä tahansa vektorin $\mathbf{Y}$ jakauman ominaisuutta kuvaavaa lukua.

Parametrejä pallot kulhossa -esimerkissä

- Yksittäisen heiton 0/1-esityksen Y_i odotusarvo tai varianssi,

$$EY_i = \frac{\theta}{N}, \qquad \operatorname{var} Y_i = \frac{\theta}{N}\left(1 - \frac{\theta}{N}\right).$$

- Summan $X = t(\mathbf{Y}) = Y_1 + \cdots + Y_n$ odotusarvo ja varianssi

$$EX = n\frac{\theta}{N}, \qquad \operatorname{var} X = n\frac{\theta}{N}\left(1 - \frac{\theta}{N}\right).$$

- Kaikkia näitä suureita voidaan kutsua parametreiksi. Parametri on yleisesti ottaen jokin mallin parametrista θ riippuva lauseke $\tau = k(\theta)$.
- Parametreja merkitään mielellään kreikkalaisilla kirjaimilla.

Populaatioparametri

- Parametrista käytetään myös nimitystä populaatioparametri.
- Tällöin ajatellaan, että aineisto on (jollakin menetelmällä muodostettu) otos joko jostakin äärellisestä populaatiosta tai jostakin (kuvitteellisesta) äärettömästä populaatiosta.
- Estimoinnin tavoitteena on tehdä johtopäätöksiä ko. populaatiosta (ts. populaatioparametreista) havaintojen avulla.
- Soveltajan tulee tarkoin miettiä, mitä populaatiota havaintoaineisto edustaa, eli mihin populaatioon tilastolliset johtopäätökset voidaan yleistää.

Tunnusluku

Määritelmä (Tunnusluku)

Tunnusluku (engl. *statistic*) tarkoittaa mitä tahansa lukua, joka voidaan laskea aineistosta ilman, että tarvitsee tuntea mitään tilastollisen mallin tuntematonta parametria.

- Binomikokeessa onnistumisten lukumäärä $t(\mathbf{y}) = \sum_{i=1}^{n} y_i$ on eräs tunnusluku.
- Kaikki tunnusluvut voidaan esittää kaavalla $t(\mathbf{y})$ jossa funktio t valitaan kulloisenkin tilanteen mukaan, ja funktio t ei saa riippua mistään mallin tuntemattomasta parametrista.

3.2 Estimaatti, estimaattori ja otantajakauma

Määritelmä (Estimaatti)

Joitakin tunnuslukuja käytetään parametrien arvioina, jolloin niitä kutsutaan vastaavien parametrien estimaateiksi.

- Nasta purkissa -esimerkissä onnistumistodennäköisyyttä θ voidaan arvioida laskemalla onnistumisten suhteellinen osuus n kokeessa, eli

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{1}$$

- Estimaatteja on tapana merkitä laittamalla hattu vastaavan parametrin päälle.
- Jos tarjolla on monta erilaista estimaattia samalle parametrille, niin ne voidaan erottaa toisistaan esimerkiksi lisäämällä merkintöihin ala- tai yläindeksejä.

Parametrin estimaatin pitää kuulua parametriavaruuteen

- Minimaalinen järkevyysvaatimus estimaatille on se, että mallin parametrin θ estimaatin $\hat{\theta}$ pitää kuulua parametriavaruuteen Θ.
- Vastaavasti parametrin $\tau = k(\theta)$ estimaatin $\hat{\tau}$ pitää kuulua joukkoon

$$\{k(\theta) : \theta \in \Theta\}.$$

Kuuluuko estimaatti parametriavaruuteen? (Nasta purkissa)

- Tarkastellaan estimaattia onnistumisten suhteellinen osuus (1).
- Nasta purkissa -esimerkissä estimaatti kuuluu parametriavaruuteen automaattisesti, mikäli parametriavaruudeksi on valittu $[0, 1]$.
- Mikäli parametriavaruudeksi valitaan avoin väli $(0, 1)$, niin estimaatti (1) ei täytä tätä minimaalista järkevyysvaatimusta, mikäli nasta ei päädy kertaakaan selälleen (jolloin $\sum_i y_i = 0$) tai mikäli nasta ei päädy kertaakaan kyljelleen (jolloin $\sum y_i = n$).

Kuuluuko estimaatti parametriavaruuteen? (Pallot kulhossa)

- Pallot kulhossa -esimerkissä onnistumisten suhteellista osuutta (1) voitaisiin ehdottaa onnistumistodennäköisyyden $\phi = \theta/N$ estimoimiseen, mutta tällöin törmätään ongelmaan.
- Parametrin ϕ arvot kuuluvat mallissa joukkoon

$$\{0, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}, 1\}$$

- Sen estimaatti $\hat{\phi}$ voi saada arvoja joukosta

$$\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1\}.$$

  eikä näillä joukoilla välttämättä ole edes kovin montaa yhteistä alkiota.
- Tämä ongelma pitäisi käytännössä kiertää pyöristämällä suhteellinen osuus jollakin tavalla diskreettiin parametriavaruuteen.

Tunnusluku satunnaismuuttujana

- Frekventistisessä päättelyssä tunnusluvun $t(\mathbf{y})$ lisäksi tarkastellaan sitä vastaavaa satunnaismuuttujaa $t(\mathbf{Y})$.
- Tällöin tunnuslukua ei lasketa havaitusta aineistosta, vaan se lasketaan aineistoa vastaavasta satunnaisvektorista $\mathbf{Y}$, jolla oletetaan olevan jokin todennäköisyysjakauma.
- Niin kauan kuin pysytään mallin $\{f(\mathbf{y};\theta) : \theta \in \Theta\}$ puitteissa, oletetaan että satunnaisvektorilla $\mathbf{Y}$ on todellista parametrinarvoa θ vastaava todennäköisyysjakauma.

Otantajakauma

Määritelmä (Otantajakauma)

Satunnaismuuttujan $t(\mathbf{Y})$ jakaumaa kutsutaan tämän tunnusluvun otantajakaumaksi (engl. *sampling distribution*). Vastaavasti, muotoa $h(\mathbf{Y}, \theta)$ olevan suureen jakaumaa kutsutaan tämän suureen otantajakaumaksi. Oletamme, että $\mathbf{Y}$ noudattaa jakaumaa $f(\mathbf{y}; \theta)$ todellisella parametrinarvolla θ.

Miksi termi otantajakauma?

- Taustalla on ajatus otannan tai aineiston keruun toistamisesta.
- Jos aineiston keruu voitaisiin toistaa samoissa olosuhteissa riippumattomasti r kertaa, ja saataisiin aineistot $\mathbf{y}_1, \ldots, \mathbf{y}_r$ (jossa kukin $\mathbf{y}_i$ on n-vektori), niin tällöin arvot $t(\mathbf{y}_1), \ldots t(\mathbf{y}_r)$ olisivat otos satunnaismuuttujaksi ymmärretyn tunnusluvun $t(\mathbf{Y})$ jakaumasta.
- Tämä ajatus voidaan toteuttaa konkreettisesti tietokoneella. Annetaan parametrisessa mallissa parametrille θ jokin lukuarvo, ja simuloidaan otos $\mathbf{y}_1, \ldots, \mathbf{y}_r$ jakaumasta $f(\mathbf{y}; \theta)$. Tällaisia simulointimenetelmiä on saatavilla lukuisille yhteisjakaumille $f(\mathbf{y}; \theta)$.

Kaksi frekventististä päättelyä koskevaa huomautusta

- Parametri θ on frekventistisessä päättelyssä kiinteä mutta tuntematon luku, jolla ei ole todennäköisyysjakaumaa.
- Frekventistisessä päättelyssä tarkastellaan parametriavaruudessa määriteltyjä jakauma. Ne ovat aina otantajakaumia.

Estimaattori ja sen otantajakauma

- Frekventistisessä tilastotieteessä erityisen kiinnostava asia on estimaattorin otantajakauma.
- Estimaattori tarkoittaa sitä, että estimaatin ei ajatella olevan konkreettinen luku, vaan sen ajatellaan olevan satunnaismuuttuja.
- Estimaattia $\hat{\theta} = t(\mathbf{y})$ vastaa estimaattori $t(\mathbf{Y})$, joka on satunnaismuuttuja. Voimme merkitä sitä myös kaavalla

$$\hat{\theta}(\mathbf{Y}) = t(\mathbf{Y}).$$

- Tässä tekstissä $\hat{\theta}$ tai $\hat{\theta}(\mathbf{y})$ tarkoittaa estimaattia (ts. konkreettista lukua), ja $\hat{\theta}(\mathbf{Y})$ vastaavaa estimaattoria, joka on satunnaismuuttuja.
- Jos $\hat{\theta}$ lasketaan aineistosta $\mathbf{y}$ kaavalla $t(\mathbf{y})$, niin $\hat{\theta}(\mathbf{Y}) = t(\mathbf{Y})$.

Estimaattorin otantajakauma — nasta purkissa

Estimaattia

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

vastaa estimaattori

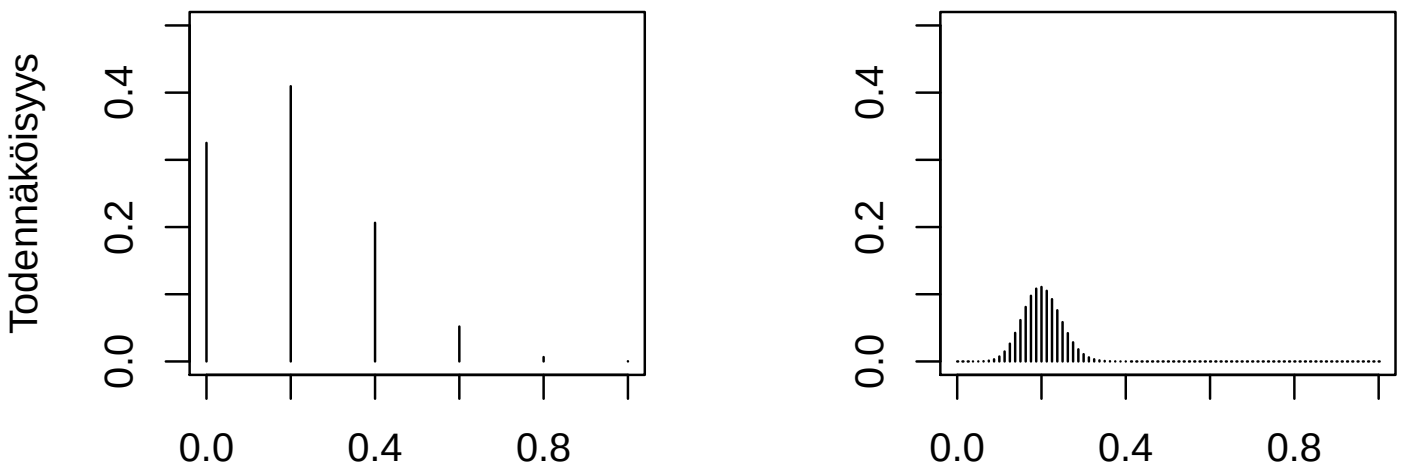
$$\hat{\theta}(\mathbf{Y}) = \frac{1}{n} \sum_{i=1}^{n} Y_i, \tag{2}$$

jonka otantajakauma on skaalausta vaille sama kuin tunnusluvun $\sum Y_i$ jakauma, joka puolestaan on binomijakauma $\mathrm{Bin}(n, \theta)$. Estimaattorin (2) otantajakauma on tällä perusteella

$$P_\theta \left(\hat{\theta}(\mathbf{Y}) = \frac{k}{n} \right) = \binom{n}{k} \theta^k (1-\theta)^{n-k}, \qquad k = 0, 1, \ldots, n.$$

Kuva estimaattorin otantajakaumasta — nasta purkissa

Kuva: Estimaattorin "onnistumisten suhteellinen osuus binomikokeessa" otantajakauma, kun $\theta = 0.2012$ ja $n = 5$ (vasemmalla) ja $n = 80$ (oikealla).



3.3 Todennäköisyyslaskennan kertausta

- Jos X on satunnaismuuttuja, jonka odotusarvo on $\mu = EX$, niin X:n varianssi on luku

$$\operatorname{var}(X) = E(X - \mu)^2.$$

- Jos X on satunnaismuuttuja, ja a ja b ovat vakiota, niin satunnaismuuttujan $aX + b$ odotusarvo ja varianssi ovat

$$E(aX + b) = aEX + b, \qquad \operatorname{var}(aX + b) = a^2 \operatorname{var} X. \quad (3)$$

- Jos X_1 ja X_2 ovat satunnaismuuttujia, niin niiden summan odotusarvo on odotusarvojen summa,

$$E(X_1 + X_2) = EX_1 + EX_2. \quad (4)$$

- Jos X_1 ja X_2 ovat riippumattomia satunnaismuuttujia, niin niiden summan tai erotuksen varianssi saadaan laskemalla yhteen muuttujien varianssit, eli

$$\operatorname{var}(X_1 \pm X_2) = \operatorname{var} X_1 + \operatorname{var} X_2. \quad (5)$$

Todennäköisyyslaskennan kertausta (jatkoa)

- Jos $\mu = EX$, ja a on vakio, niin helpolla laskulla nähdään, että

$$E(X-a)^2 = E(X-\mu)^2 + (\mu - a)^2 = \operatorname{var} X + (\mu - a)^2 \quad (6)$$

- Tšebyševin epäyhtälön mukaan mille tahansa vakiolle a

$$P(|X-a| > \epsilon) \leq \frac{E(X-a)^2}{\epsilon^2}, \qquad \text{kaikille } \epsilon > 0. \quad (7)$$

3.4 Otantajakauman ominaisuuksia

Binomikokeessa onnistumisten suhteellisen osuuden (2) (otantajakauman) odotusarvo ja varianssi ovat helppo selvittää:

$$E_\theta[\hat{\theta}(\mathbf{Y})] = \frac{1}{n}\sum_{i=1}^{n} E_\theta(Y_i) = \frac{1}{n}\, n\,\theta = \theta,$$

$$\mathrm{var}_\theta[\hat{\theta}(\mathbf{Y})] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}_\theta(Y_i) = \frac{1}{n^2}\, n\,\theta\,(1-\theta) = \frac{1}{n}\,\theta\,(1-\theta)$$

Alaindeksillä θ korostetaan sitä, että satunnaisvektorilla $\mathbf{Y} = (Y_1, \ldots, Y_n)$ oletetaan olevan jakauma $f(\mathbf{y}; \theta)$.
Otantajakauman varianssin määrittäminen perustui siihen oletukseen, että satunnaismuuttujat Y_i ovat riippumattomia.
Vaihtoehtoisesti voimme johtaa odotusarvon ja varianssin käyttämällä hyväksi tunnettuja kaavoja binomijakauman $\mathrm{Bin}(n, \theta)$ odotusarvolle ja varianssille.

Harhaton estimaattori

Määritelmä (Harhattomuus)

Jos estimaattorin odotusarvo on sama kuin parametrin todellinen arvo, eli

$$E_\theta[\hat{\theta}(\mathbf{Y})] = \theta, \quad \text{kaikilla } \theta,$$

niin sanotaan, että estimaattori $\hat{\theta}(\mathbf{Y})$ on *harhaton* (engl. *unbiased*). Muussa tapauksessa sanotaan, että estimaattori on harhainen (engl. *biased*).

Tarkemmin sanoen edellinen asia voidaan ilmaista niin, että estimaattori on *odotusarvon mielessä* harhaton; odotusarvon sijasta voisimme toki tarkastella muitakin otantajakauman keskikohtaa kuvailevia suureita, kuten mediaania.

Estimaattorin harha

Määritelmä (Harha)

Estimaattorin $\hat{\theta}(\mathbf{Y})$ harha on

$$\text{bias}_\theta(\hat{\theta}(\mathbf{Y})) = E_\theta(\hat{\theta}(\mathbf{Y})) - \theta. \tag{8}$$

Harhaa pidetään usein estimaattorin systemaattisena virheenä. Harhattoman estimaattorin harha on nolla koko parametriavaruudessa. Harhainen estimaattori ei kuitenkaan välttämättä ole huono estimaattori eikä harhaton estimaattori ole välttämättä hyvä estimaattori. Nasta purkissa -esimerkissä estimaattori (2) (onnistumisten suhteellinen osuus) on harhaton.

Estimaattorin harha (jatkoa)

Mallin parametrin θ sijasta voitaisiin tarkastella myös jotakin muuta parametria $\tau = k(\theta)$ estimoivan estimaattorin $\hat{\tau}(\mathbf{Y})$ harhaa. Tämä tietenkin määritellään edellistä vastaavalla kaavalla

$$\mathrm{bias}_\theta(\hat{\tau}(\mathbf{Y})) = E_\theta(\hat{\tau}(\mathbf{Y})) - k(\theta). \tag{9}$$

Harha voidaan määritellä samalla kaavalla myös silloin, jos parametri on vektori.

Merkinnöistä

Merkinnät näyttävät raskailta, joten avaan seuraavaksi niiden merkitystä estimaattorin harhan määritelmän eli kaavan (8)

$$\mathrm{bias}_\theta(\hat{\theta}(\mathbf{Y})) = E_\theta(\hat{\theta}(\mathbf{Y})) - \theta$$

kohdalla.

- Siinä puhutaan estimaattorista $\hat{\theta}(\mathbf{Y})$, jota siis käsitellään satunnaismuuuttujana.
- Estimaattori $\hat{\theta}(\mathbf{Y})$ on funktio satunnaisvektorista $\mathbf{Y}$, joten estimaattorin jakauma riippuu satunnaisvektorin $\mathbf{Y}$ jakaumasta.
- Alaindeksi θ kertoo, että satunnaisvektorin $\mathbf{Y}$ jakaumalla on yptnf tai ytf $f(\mathbf{y}; \theta)$.

Onko pakko käyttää näin raskaita merkintöjä?

- Näissä luentomuistiinpanoissa käytetään tällaisia pedanttisia merkintöjä, jotta lukija pystyisi kaavoista heti näkemään, mitä suureita pidetään kiinteinä ja mitä satunnaisina ja mitä jakaumia satunnaisille suureille oletetaan.
- Sen jälkeen, kun nämä asiat alkavat olla itsestään selviä, voit rauhassa tiputtaa kaavoista ylimääräiset koristeet, ja kirjoittaa vaikkapa

$$\mathrm{bias}(\hat{\theta}) = E\hat{\theta} - \theta,$$

millä tyylillä nämä asiat monessa oppikirjassa esitetään.
- Kirjallisuudessa ei välttämättä tehdä eroa termien estimaatti ja estimaattori välillä, vaan termi estimaatti saattaa tarkoittaa niistä kumpaa tahansa. Lisäksi merkintä $\hat{\theta}$ saattaa kontekstista riippuen tarkoittaa yhtä hyvin estimaattia tai estimaattoria.

Keskineliövirhe, MSE

Määritelmä (Keskineliövirhe)

Estimaattorin $\hat{\theta}(\mathbf{Y})$ keskineliövirhe (engl. *mean squared error*) on

$$\mathrm{mse}_\theta(\hat{\theta}(\mathbf{Y})) = E_\theta\left[(\hat{\theta}(\mathbf{Y}) - \theta)^2\right] \tag{10}$$

Keskineliövirhe kuvaa estimaattorin tarkkuutta: mitä pienempi keskineliövirhe, sitä tarkempia arvioita keskimäärin saadaan. Keskineliövirhe riippuu tyypillisesti voimakkaasti otoskoosta n siten, että suuremmalla otoskoolla saavutetaan pienempi keskineliövirhe.

Keskineliövirhe (jatkoa)

Mikäli estimaattori on harhaton, niin sen keskineliövirhe on sama kuin sen varianssi.
Helpolla laskulla (vrt. kaava (6)) nähdään, että keskineliövirhe voidaan esittää laskemalla yhteen estimaattorin varianssi ja sen harhan neliö, eli

$$\mathrm{mse}_\theta(\hat{\theta}(\mathbf{Y})) = \mathrm{var}_\theta(\hat{\theta}(\mathbf{Y})) + \Big(\mathrm{bias}_\theta(\hat{\theta}(\mathbf{Y}))\Big)^2. \qquad (11)$$

Tämä on ns. *bias-variance decomposition*.

Keskineliövirheen neliöjuuri, RMSE

Keskineliövirheen sijasta usein tarkastellaan sen neliöjuurta, koska se on samalla skaalalla kuin itse estimaattori.

Määritelmä (Keskineliövirheen neliöjuuri, RMSE)

Estimaattorin $\hat{\theta}(\mathbf{Y})$ keskineliövirheen neliöjuuri (engl. *root mean squared error*) on

$$\mathrm{rmse}_\theta(\hat{\theta}(\mathbf{Y})) = \sqrt{\mathrm{mse}_\theta(\hat{\theta}(\mathbf{Y}))}. \tag{12}$$

Estimaatin keskivirhe

Estimaattien yhteydessä usein kerrotaan niiden keskivirhe. Tämä on yksi tapa arvioida estimointiin liittyvää epävarmuutta.

Määritelmä (Keskivirhe)

Estimaatin $\hat{\theta}$ keskivirhe (engl. *standard error*, *s.e.*, *se*) tarkoittaa otoksesta (jollakin järkevällä tavalla) muodostettua estimaattia vastaavan estimaattorin $\hat{\theta}(\mathbf{Y})$ keskineliövirheen neliöjuurelle (eli RMSE:lle).

Miten estimaatin keskivirhe muodostetaan?

- Estimaattorin keskineliövirheen neliöjuuri (eli RMSE) riippuu yleensä jollakin tavalla parametrinarvosta θ, ja kun tähän kaavaan sijoitetaan tuntemattoman parametrin tai tuntemattomien parametrien tilalle niiden estimaatit, niin saadaan estimaatin keskivirhe.
- Tyypillisesti keskivirheestä puhutaan silloin, kun vastaava estimaattori on harhaton. Tällöin sen keskineliövirheen neliöjuuri on sama asia kuin estimaattorin (otantajakauman) varianssin neliöjuuri. Varianssin neliöjuuresta käytetään nimitystä keskihajonta (engl. *standard deviation*).
- Harhatonta estimaattoria vastaavan estimaatin keskivirhe on kyseisen estimaattorin otantajakauman estimoitu keskihajonta.

Miksi puhutaan näin pitkään otantajakaumasta?

- Frekventistisessä tilastotieteessä erilaisia estimaattoreja verrataan keskenään niiden otantajakaumien ominaisuuksien (kuten esimerkiksi harhan ja varianssin) avulla.
- Kun estimaatti sitten lasketaan aineistosta, niin (epämuodollisesti) ajatellaan, että kyseinen estimaatti on tarkka, mikäli vastaavalla estimaattorilla on suotuisa otantajakauma (esim. pieni harha ja pieni varianssi).

3.5 Tarkentuvuus

- Tarkentuvuus on yksinkertainen esimerkki estimaattorin asymptoottisesta ominaisuudesta.
- Tällöin otoskoon kuvitellaan kasvavan rajatta, vaikka todellisuudessa havaintoja tietenkin on täsmälleen vain niin monta kuin mitä niitä on.
- Estimaattorien asymptoottisia ominaisuuksia tarkastellaan sen takia, että riittävän suurella otoskoolla asymptoottisten ominaisuuksien ajatellaan toteutuvan likimäärin.
- Teoreettisessa tilastotieteessä ei tyypillisesti pystytä kertomaan, milloin otoskoko on riittävän suuri jotta asymptotiikasta saatava arvio olisi käytännön kannalta riittävällä tarkkuudella voimassa. Tämä on taas kerran sellainen asia, jota on helpointa yrittää selvittää tietokonesimuloinneilla.

Estimaattorin asymptoottinen ominaisuus

- Kun puhutaan estimaattorien asymptoottisista ominaisuuksista, niin kutakin otoskokoa n oikeastaan vastaa yksi estimaattori.
- Itse asiassa tarkastelemme näiden estimaattorien muodostamaa jonoa.
- Yksinkertaisuuden vuoksi emme merkitse otoskokoa näkyviin estimaattorin yhteyteen.
- Käytämme tavanomaista puhetapaa, jossa ei tehdä eroa yksittäistä otoskokoa vastaavan estimaattorin ja eri otoskokoja vastaavien estimaattorien muodostaman estimaattorijonon välillä.

Tarkentuvuus

- Parametrin θ estimaattori $\hat{\theta}(\mathbf{Y})$ on tarkentuva (engl. *consistent*), mikäli se suppenee kohti parametrin θ todellista arvoa otoskoon n kasvaessa rajatta.
- Tällöin tietenkin myös havaintosatunnaisvektorin

$$\mathbf{Y} = (Y_1, \ldots, Y_n)$$

  pituus kasvaa rajatta.
- Tarkentuvuus on oikeastaan edellytys sille, että estimaattorin $\hat{\theta}(\mathbf{Y})$ voidaan sanoa estimoivan parametria θ. Jos otokokoa pystytään kasvattamaan rajatta, niin parametrin arvo saadaan rajalla selvitettyä tarkentuvan estimaattorin avulla.

Stokastinen suppeneminen

Tämän kurssin puitteissa tarkentuvuuden yhteydessä vaaditaan ns. stokastinen suppeneminen, joka määritellään ensin satunnaismuuttujajonolle $X_1, X_2, \ldots$.

Määritelmä (Stokastinen suppeneminen)

Jono satunnaismuuttujia $X_1, X_2, \ldots$ suppenee stokastisesti (engl. *converges in probability*) kohti vakiota a, mikäli

$$P(|X_n - a| > \epsilon) \to 0, \qquad \text{kaikilla } \epsilon > 0.$$

Tämä asia voidaan ilmaista merkinnällä

$$X_n \xrightarrow{\mathrm{p}} a.$$

Mitä stokastinen suppeneminen tarkoittaa?

- Oletetaan, että $X_n \xrightarrow{\mathrm{p}} a$.
- Jos $\epsilon > 0$ on mielivaltaisen pieni luku, niin todennäköisyys, että X_n ei satu välille $[a - \epsilon, a + \epsilon]$ suppenee kohti nollaa n:n kasvaessa.
- Yhtäpitävästi: todennäköisyys, että X_n sattuu välille $[a - \epsilon, a + \epsilon]$ suppenee kohti ykköstä, kun n kasvaa rajatta.
- Satunnaismuuttujien X_n jakauma keskittyy yhtä tiukemmin luvun a läheisyyteen, kun n kasvaa.

Kriteeri stokastiselle suppenemiselle

- Stokastisen suppenemisen voi usein todistaa seuraavan kriteerin avulla.

$$E(X_n - a)^2 \to 0 \quad \Rightarrow \quad X_n \xrightarrow{\mathrm{p}} a. \tag{13}$$

- Tämä seuraa Tšebyševin epäyhtälöstä (7). Jos nimittäin $\epsilon > 0$, niin

$$P(|X_n - a| > \epsilon) \leq \frac{E(X_n - a)^2}{\epsilon^2},$$

ja tämä yläraja suppenee oletuksen mukaan kohti nollaa n:n kasvaessa.

Tarkentuva estimaattori

Määritelmä (Tarkentuvuus)

Jos $\hat{\theta}(\mathbf{Y}) \xrightarrow{\mathrm{p}} \theta$ kaikilla $\theta \in \Theta$, kun otoskoko n ja sen mukana havaintosatunnaisvektorin

$$\mathbf{Y} = (Y_1, \ldots, Y_n)$$

pituus kasvavat rajatta, niin tällöin sanotaan, että estimaattori(jono) $\hat{\theta}(\mathbf{Y})$ on *tarkentuva* (engl. *consistent*).

Tarkentuvuuden tarkistaminen

- Tyypillisesti estimaattorin keskineliövirhe $\mathrm{mse}_\theta(\hat{\theta}(\mathbf{Y}))$ suppenee kohti nollaa, kun otoskoko n kasvaa rajatta.
- Tällöin siis kaikilla θ pätee

$$E_\theta \left(\hat{\theta}(\mathbf{Y}) - \theta\right)^2 \to 0, \qquad \text{kun } n \to \infty.$$

- Tuloksen (13) mukaan $\hat{\theta}(\mathbf{Y}) \xrightarrow{\mathrm{p}} \theta$ kaikilla θ joten tällöin ko. estimaattori on tarkentuva.

Nasta purkissa — tarkentuvuus

Nasta purkissa -esimerkissä estimaattorin (2) keskineliövirhe on harhattomuuden ansiosta sama kuin sen varianssi, joten

$$\mathrm{mse}_\theta(\hat{\theta}(\mathbf{Y})) = \mathrm{var}_\theta(\hat{\theta}(\mathbf{Y})) = \frac{1}{n}\,\theta\,(1-\theta),$$

ja koska tämä suppenee otoskoon kasvaessa kohti nollaa kaikilla θ, on estimaattori tarkentuva.