

MS-A0502 Todennäköisyyslaskennan ja tilastotieteen peruskurssi

4A Parametrien estimointi

Emilia Blåsten

Matematiikan ja systeemianalyysin laitos
Perustieteiden korkeakoulu
Aalto-yliopisto

Lukuvuosi 2021–2022
Periodi II

Sisältö

Tilastollinen päättely

Parametriset tilastolliset jakaumat

Suurimman uskottavuuden estimaattorit

Estimaattoreiden ominaisuuksia

Tilastollinen päättely

Tavoitteena tehdä päätelmiä havaitun datan pohjalta.

1. Valitaan tilanteeseen sopiva stokastinen **malli**
jakaumaperhe, esim. "kaikki normaalijakaumat" tai "kaikki tasajakaumat $[0, m]$ "
2. **Sovitetaan** malli havaittuun dataan
(estimoidaan mallin parametrit)
3. Lasketaan sovitetusta mallista tarvittavat tunnusluvut
4. Tehdään johtopäätökset

Johtopäätökset ovat yleensä (valistuneita) arvauksia:

- Mikä on kirahvin todellinen paino, kun kolmen punnituksen tulokset olivat 1250 kg, 1300 kg, 1360 kg?
- Kannattaako espoolaisten enemmistö Espoon pysymistä itsenäisenä, kun mielipidekyselyssä 509 tuhannesta kannatti itsenäisyyttä?
- Pysykö raakaöljyn hinta nykytasollaan vuoden loppuun asti?

Sisältö

Tilastollinen päättely

Parametriset tilastolliset jakaumat

Suurimman uskottavuuden estimaattorit

Estimaattoreiden ominaisuuksia

Tuntemattoman jakauman parametrit

Tarkastellaan tuntematonta datalähdettä, jonka tutkittavan suureen jakauma $f(x)$ tunnetaan **parametreja** vaille.

Esim. (yksi tuntematon parametri):

- Bernoullijakauma: $f_p(1) = p$ ja $f_p(0) = 1 - p$
- Eksponenttijakauma: $f_\lambda(x) = \lambda e^{-\lambda x}$, $x > 0$
- Välin $[0, b]$ tasajakauma: $f_b(x) = \frac{1}{b}$

Esim. (2 tuntematonta parametria):

- Välin $[a, b]$ tasajakauma: $f_{a,b}(x) = \frac{1}{b-a}$
- Normaalijakauma: $f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Mikä on havaitun datan (x_1, \dots, x_n) perusteella paras arvaus tuntemattoman parametrin arvoksi?

Huom. merkinnöistä: Tässä alaindeksillä täsmennetään *eräs* tietty tiheysfunktio (tietyillä parametreilla), eikä satunnaismuuttujaa. Voitaisiin myös käyttää pystyviivamerkintää, esim. $f(x | \lambda)$.

Parametrien estimointi

Tarkastellaan tuntematonta datalähdettä, jonka tutkittavan suureen jakauma on $f_\theta(x)$, parametri θ tuntematon.

Datalähteestä on saatu n riippumatonta ja samoin (tiheydellä f_θ) jakautunutta havaintoa x_1, \dots, x_n .

Parametrin θ :

- **estimaatti** on datan $\vec{x} = (x_1, \dots, x_n)$ pohjalta laskettu arvaus $\hat{\theta} = g(\vec{x})$
- **estimaattori** on funktio $(x_1, \dots, x_n) \mapsto g(x_1, \dots, x_n)$, joka kuvaa datan estimaatiksi

Tietyn parametrin estimaattoriksi ei yleensä ole yksikäsitteistä “parasta” valintaa.

On olemassa (useitakin) ominaisuuksia, joita estimaattorilla olisi mukava olla.

Muutamia käsitellään tässä luennossa: **suurin uskottavuus** ja **harhattomuus**.

Valitettavasti nämä voivat olla ristiriidassa.

Esimerkki: Viallisten osuus

Tuotantolinja tuottaa komponentteja, joista osuus p on viallisia, toisistaan riippumattomasti. Kun tarkastettiin 200 komponenttia, havaittiin 22 viallista. Määritä estimaatti tuntemattoman parameterin p arvolle.

Intuitiivisesti luonteva estimaatti on

$$\hat{p} = \frac{22}{200} = 11\%$$

Onko tämä paras estimaatti? Onko muita luonnollisia vaihtoehtoja?

Huom: Hatullisella kirjaimella merkitään yleensä estimaattia ja ilman hattua "oikeaa" parametrin arvoa (populaatiossa tai generoivassa mallissa).

Esimerkki: Diskreetin tasajakauman parametri

Vieraan vallan sotilaskoneissa on sarjanumerot $1, 2, \dots, n$.

Tiedustelijat ovat havainneet kolmen sotilaskoneen sarjanumerot $x_1 = 63$, $x_2 = 17$, $x_3 = 203$. Määritä havaintojen pohjalta estimaatti sotilaskoneiden lukumäärälle n .

Tiedustelutietoa tuottava datalähde noudattaa tasajakaumaa

$$f_{1,n}(k) = \begin{cases} \frac{1}{n}, & k = 1, \dots, n, \\ 0, & \text{muuten.} \end{cases}$$

Mikä on luonteva estimaattori $\hat{n}(\vec{x})$ parametrille n ?

Tätä estimointitehtävää voi lähestyä eri tavoin ja saada erilaisia estimaattoreita. Ks. harjoitus 4B.

Sisältö

Tilastollinen päättely

Parametriset tilastolliset jakaumat

Suurimman uskottavuuden estimaattorit

Estimaattoreiden ominaisuuksia

Uskottavuusfunktio

[engl. likelihood function]

Datalähteen stokastinen malli: (X_1, \dots, X_n) , jonka komponentit f_θ -jakautuneet ja toisistaan riippumattomat.

Mallin ennustama todennäköisyys havaita arvot (x_1, \dots, x_n) on diskreetille jakaumalle

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = f_\theta(x_1) \cdots f_\theta(x_n)$$

ja jatkuvalle jakaumalle (likimain, pienellä ε)

$$\mathbb{P}(X_1 = x_1 \pm \frac{\varepsilon}{2}, \dots, X_n = x_n \pm \frac{\varepsilon}{2}) \approx \varepsilon^n f_\theta(x_1) \cdots f_\theta(x_n).$$

Uskottavuusfunktio $L(\theta) = f_\theta(x_1) \cdots f_\theta(x_n)$ kertoo f_θ -mallin ennustaman todennäköisyyden havaita (likimain) sama data, mitä oikeasti havaittiin.

Suurimman uskottavuuden estimaatti

[engl. maximum likelihood estimate, ML estimate]

Uskottavuusfunktio $L(\theta) = f_{\theta}(x_1) \cdots f_{\theta}(x_n)$ kertoo f_{θ} -mallin ennustaman todennäköisyyden havaita (likimain) sama data, mitä oikeasti havaittiin.

Mitä suurempi uskottavuusfunktion arvo on pisteessä θ , sen uskottavampana voidaan pitää oletusta, että havaittu data on peräisin f_{θ} -jakautuneesta datalähteestä.

Parametrin θ suurimman uskottavuuden estimaatti $\hat{\theta} = \hat{\theta}(\vec{x})$ on se parametrin arvo, joka maksimoi uskottavuusfunktion.

Näin aseteltuna kyseessä on *optimointitehtävä*, tarkemmin sanottuna funktion L maksimikohdan etsimistehtävä. Funktion maksimointiin löytyy työkaluja differentiaalilaskennasta.

Esimerkki: Viallisten osuuden estimointi

Tuotantolinja tuottaa komponentteja, joista kukin on toisista riippumatta viallinen tn:llä p . Kun tarkastettiin 200 komponenttia, havaittiin 22 viallista. Määritä suurimman uskottavuuden estimaatti tuntemattomalle parametrille p .

Kun tarkastetaan $n = 200$ komponentin erä, noudattaa viallisten komponenttien lukumäärä N binomijakaumaa

$$f_p(x) = \mathbb{P}(N = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, 1, \dots, 200$$

Millä parametrin p arvolla uskottavuusfunktio

$$L(p) = \binom{200}{22} p^{22} (1 - p)^{200-22}$$

saavuttaa suurimman arvonsa?

Tämä on yhden muuttujan (p) *optimointitehtävä*. Kaikki muut suureet (n ja x) ovat tunnettuja.

Esimerkki: Viallisten osuuden estimointi

$$L(p) = \binom{200}{22} p^{22} (1-p)^{178}$$

maksimoituu silloin, kun $\ell(p) = \log L(p)$ maksimoituu:

$$\ell(p) = \log f_p(22) = \log \binom{200}{22} + 22 \log p + 178 \log(1-p)$$

$$\ell'(p) = 22 \frac{1}{p} - 178 \frac{1}{1-p}$$

$$\ell''(p) = -22 \frac{1}{p^2} - 178 \frac{1}{(1-p)^2} \leq 0$$

Parametrin p suurimman uskottavuuden estimaatti löytyy derivaatan nollakohdasta:

$$\ell'(p) = 0 \quad \iff \quad \frac{22}{p} = \frac{178}{1-p} \quad \iff \quad p = \frac{22}{200}$$

Logaritmin ottaminen oli laskennallinen temppu derivoinnin helpottamiseksi. Vaihtoehtoisesti oltaisiin voitu etsiä suoraankin funktion L maksimikohtaa.

Binomijakauman tn-parametrin ML-estimointi

Fakta

Bin(n, p)-jakauman tuntemattoman parametrin p suurimman uskottavuuden estimaatti havaitun datapisteen x suhteen on

$$\hat{p} = \frac{x}{n}.$$

Todistus.

Toista edellinen laskelma korvaamalla $200 \mapsto n$ ja $22 \mapsto x$. □

Normaalijakauman parametrien ML-estimaatit

Normaalijakauman tiheysfunktio

$$f_{(\mu, \sigma)}(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$$

on parametreja μ ja σ vaille tunnettu.

Fakta

Normaalijakauman parametrien (μ, σ) suurimman uskottavuuden estimaatit datajoukolle $\vec{x} = (x_1, \dots, x_n)$ ovat

$$m(\vec{x}) = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{ja} \quad \text{sd}(\vec{x}) = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - m(\vec{x}))^2}$$

eli datajoukon \vec{x} keskiarvo ja keskihajonta.

Tämä voidaan todistaa kirjoittamalla uskottavuuden **molemmat osittaisderivaatat** (kummankin parametrin suhteen) ja ratkaisemalla, milloin ne ovat nollia, ks. luentomoniste.

Sisältö

Tilastollinen päättely

Parametriset tilastolliset jakaumat

Suurimman uskottavuuden estimaattorit

Estimaattoreiden ominaisuuksia

Harhaton estimaattori

[engl. unbiased estimator]

Jakauman f_θ parametrin θ estimaattori $\hat{\theta}(\vec{x})$ on **harhaton**, jos f_θ -jakaumaa vastaavalle stokastiselle mallille $\vec{X} = (X_1, \dots, X_n)$ pätee

$$\mathbb{E}\hat{\theta}(\vec{X}) = \theta$$

eli jos “odotusarvo on kohdallaan”.

Tulkinta: Jos tuntematon datalähde todella noudattaa f_θ -jakaumaa, ja datalähteestä tehdään n riippumatonta havaintoa ja lasketaan estimaatti harhattomalla estimaattorilla, ja jos sama toistettaisiin monta kertaa, niin toistojen keskiarvo on lähellä oikeaa parametria.

Esimerkki: Viallisten osuus

Kun n on tunnettu, on $\text{Bin}(n, p)$ -jakauman tuntemattoman parametrin p suurimman uskottavuuden estimaattori

$$\hat{p}(x) = \frac{x}{n}.$$

Jos N on $\text{Bin}(n, p)$ -jakaumaa noudattava satunnaismuuttuja, niin

$$\mathbb{E}(\hat{p}(N)) = \mathbb{E}\left(\frac{N}{n}\right) = \frac{1}{n}\mathbb{E}(N) = \frac{1}{n} \times np = p.$$

Näin ollen funktio

$$x \mapsto \hat{p}(x)$$

on parametrin p harhaton estimaattori.

Esim: Normaalijakauman odotusarvon ML-estimaattori

Normaalijakauman odotusarvoparametrin μ suurimman uskottavuuden estimaattori on

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

Stokastiselle mallille $\vec{X} = (X_1, \dots, X_n)$

$$\mathbb{E}[m(\vec{X})] = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \mu,$$

joten funktio $\vec{x} \mapsto m(\vec{x})$ on parametrin μ harhaton estimaattori.

Esim: Normaalijakauman varianssin ML-estimaattori

Normaalijakauman varianssiparametrin σ^2 suurimman uskottavuuden estimaattori on

$$\text{var}(\vec{x}) = \frac{1}{n} \sum_{i=1}^n (x_i - m(\vec{x}))^2.$$

Stokastiselle mallille $\vec{X} = (X_1, \dots, X_n)$

$$\mathbb{E}[\text{var}(\vec{X})] = \mathbb{E} \left(\frac{1}{n} \sum_{i=1}^n (X_i - m(\vec{X}))^2 \right) = \dots = \frac{n-1}{n} \sigma^2,$$

joten $\text{var}(\vec{x})$ on harhainen. Varianssiparametrin harhaton estimaattori on niin sanottu otosvariassi

$$\text{var}_s(\vec{x}) = \frac{1}{n-1} \sum_{i=1}^n (x_i - m(\vec{x}))^2.$$

Suurilla n :n arvoilla näissä ei ole kovin suurta eroa.

Seuraavalla kerralla puhutaan luottamusväleistä...