Ilkka Mellin

Tilastolliset menetelmät

Osa 2: Otokset, otosjakaumat ja estimointi

Estimointi

Estimointi

>> **Todennäköisyysjakaumien parametrit ja niiden estimointi**

Hyvän estimaattorin ominaisuudet

Todennäköisyysjakaumat tilastollisten aineistojen kuvaajina

- **Tilastollinen aineisto** koostuu tutkimuksen kohteita kuvaavien muuttujien *havaituista arvoista*.
- Tilastollisissa tutkimusasetelmissa havaintoarvoihin liittyy aina *epävarmuutta* ja *satunnaisuutta*.
- Tilastollisissa tutkimusasetelmissa tutkimuksen kohteita kuvaavat muuttujat tulkitaan *satunnaismuuttujiksi*, jotka *generoivat* muuttujien havaitut arvot.
- *Havaintoarvot generoineiden satunnaismuuttujien todennäköisyysjakauma* muodostaa **tilastollisen mallin** sille satunnaisilmiölle, jota havainnot koskevat.

Todennäköisyysjakaumien parametrit 1/2

- Tarkastellaan jotakin tutkimuksen kaikkien mahdollisten kohteiden muodostaman perusjoukon S alkioiden ominaisuutta kuvaavaa *satunnaismuuttujaa* X.
- Oletetaan, että satunnaismuuttuja X noudattaa *todennäköisyysjakaumaa*, jonka *pistetodennäköisyys*- tai *tiheysfunktio*

$$f(x\,;\,\theta)$$

riippuu **parametrista** θ.

- Merkintä:

$$X \sim f(x\,;\theta)$$

Todennäköisyysjakaumien parametrit 2/2

- Satunnaismuuttujan X pistetodennäköisyys- tai tiheysfunktio

  $$f(x\,;\,\theta)$$

  kuvaa satunnaismuuttujan X *todennäköisyysjakaumaa* ja parametri θ kuvaa jotakin jakauman *karakteristista ominaisuutta*.
- Koska parametrin θ arvoa *ei yleensä tunneta*, tilastollisen tutkimuksen tärkeimpiä osatehtäviä on **estimoida** eli *arvioida* parametrin θ tuntematon arvo jakaumasta $f(x\,;\,\theta)$ *poimitun otoksen perusteella*.

Yksinkertainen satunnaisotos

- Olkoon

  $$X_1 , X_2 , \ldots , X_n$$

  (**yksinkertainen**) **satunnaisotos** jakaumasta, jonka pistetodennäköisyys- tai tiheysfunktio $f(x\ ;\ \theta)$ riippuu parametrista θ.
- Tällöin havainnot $X_1 , X_2 , \ldots , X_n$ ovat *riippumattomia*, *identtisesti jakautuneita satunnaismuuttujia*, joilla on *sama pistetodennäköisyys*- tai *tiheysfunktio* $f(x\ ;\ \theta)$:

  $$X_1, X_2, \ldots, X_n \perp$$

  $$X_i \sim f(x\,;\theta)\ ,\ i = 1, 2, \ldots, n$$

Havainnot ja havaintoarvot

- Oletetaan, että satunnaismuuttujat (havainnot)

  $$X_1\ , X_2\ , \ldots\ , X_n$$

  saavat *poimitussa otoksessa* **havaituiksi arvoikseen** luvut

  $$x_1\ , x_2\ , \ldots\ , x_n$$

- **Havaintoarvot**

  $$x_1\ , x_2\ , \ldots\ , x_n$$

  vaihtelevat satunnaisesti otoksesta toiseen jakaumasta

  $$f(x\ ;\ \theta)$$

  saatavin todennäköisyyksin.

Estimaattorit ja estimaatit 1/2

- Oletetaan, että todennäköisyysjakauman $f(x\ ;\ \theta)$ **parametrin θ estimoimiseen** käytetään satunnaismuuttujien $X_1, X_2, \ldots, X_n$ funktiota eli *tunnuslukua*

$$T = g(X_1, X_2, \ldots, X_n)$$

- Tällöin funktiota $T = g(X_1, X_2, \ldots, X_n)$ kutsutaan parametrin θ **estimaattoriksi** ja funktion g *havaintoarvoista*

$$x_1, x_2, \ldots, x_n$$

laskettua arvoa

$$t = g(x_1, x_2, \ldots, x_n)$$

kutsutaan parametrin θ **estimaatiksi**.

Estimaattorit ja estimaatit 2/2

- Olkoon

$$T = g(X_1 , X_2 , \ldots , X_n)$$

jakauman $f(x\ ;\ \theta)$ parametrin θ *estimaattori*.

- Tällöin estimaattorin T havaintoarvoista

$$x_1 , x_2 , \ldots , x_n$$

laskettu arvo eli *estimaatti*

$$t = g(x_1 , x_2 , \ldots , x_n)$$

on satunnaismuuttujan T arvon realisaatio otoksessa.

Estimaattorit ja estimaatit: Kommentti

- Todennäköisyysjakauman $f(x\ ;\ \theta)$ parametrin θ *estimaattorilla*

  $$T = g(X_1\ , X_2\ , \ldots\ , X_n)$$

  tarkoitetaan siis sellaista jakaumaa $f(x\ ;\ \theta)$ noudattavien *satunnaismuuttujien*

  $$X_1\ , X_2\ , \ldots\ , X_n$$

  funktiota, *joka generoi* muuttujien $X_1\ , X_2\ , \ldots\ , X_n$ havaittuihin arvoihin $x_1\ , x_2\ , \ldots\ , x_n$ sovellettuna *estimaatteja* eli *arvioita*

  $$t = g(x_1\ , x_2\ , \ldots\ , x_n)$$

  parametrille θ.

Estimaattorin otosjakauma

- Estimaattorin

  $$T = g(X_1 , X_2 , \ldots , X_n)$$

  havaintoarvoista

  $$x_1 , x_2 , \ldots , x_n$$

  lasketut arvot eli estimaatit

  $$t = g(x_1 , x_2 , \ldots , x_n)$$

  vaihtelevat satunnaisesti otoksesta toiseen.
- Estimaattorin T arvojen satunnaista vaihtelua otoksesta toiseen voidaan kuvata *estimaattorin T* **otosjakaumalla**.

Estimaattoreiden johtaminen

- *Hyvien estimaattoreiden johtaminen* todennäköisyys-jakaumien tuntemattomille parametreille *on teoreettisen tilastotieteen keskeisiä ongelmia*.
- Tärkeimmät estimaattoreiden johtamiseen käytettävät menetelmät:
  - **Suurimman uskottavuuden menetelmä**
  - **Momenttimenetelmä**

  Ks. lukua **Estimointimenetelmät**.

Piste-estimointi ja väliestimointi

- Todennäköisyysjakauman *parametrin* ***arvon*** *estimointia* kutsutaan usein **piste-estimoinniksi**.
- Parametrin estimaattiin on aina syytä liittää **luottamus-väliksi** kutsuttu *väli*, *joka sisältää estimoidun parametrin todellisen*, *mutta tuntemattoman arvon tietyllä*, *soveltajan valittavissa olevalla todennäköisyydellä*.
- Luottamusvälin määräämistä kutsutaan **väliestimoinniksi**.

  Ks. lukua **Väliestimointi.**

Estimointi

Todennäköisyysjakaumien parametrit ja niiden estimointi

>> Hyvän estimaattorin ominaisuudet

Hyvä estimaattori

- Todennäköisyysjakauman parametreille on tavallisesti tarjolla useita *vaihtoehtoisia estimaattoreita*.
- Estimaattorin valintaa ohjaavat **hyvyyskriteerit**, joilla pyritään takamaan se, että valittu estimaattori tuottaa järkeviä arvoja estimoitavalle parametrille.
- Estimaattoreiden hyvyyskriteereitä:
  - **Tyhjentävyys**
  - **Harhattomuus**
  - **Tehokkuus**
  - **Tarkentuvuus**

Tyhjentävyys

- Olkoon $\hat{\theta}$ parametrin θ estimaattori
- Estimaattori $\hat{\theta}$ on **tyhjentävä** parametrille , jos se *käyttää* parametrin arvon estimoimiseen *kaiken otoksessa olevan informaation parametrista* θ.

Hyvän estimaattorin ominaisuudet

Harhattomuus ja harha

- Olkoon $\hat{\theta}$ parametrin θ estimaattori.
- Estimaattori $\hat{\theta}$ on **harhaton** parametrille θ, jos sen odotusarvo yhtyy parametrin θ arvoon:

$$\mathrm{E}(\hat{\theta}) = \theta$$

- Harhaton estimaattori tuottaa *keskimäärin* oikean kokoisia arvoja parametrille.
- Estimaattorin $\hat{\theta}$ **harha** on

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

- Jos estimaattori $\hat{\theta}$ on harhaton parametrille θ, niin

$$\mathrm{Bias}(\hat{\theta}) = 0$$

Estimaattorin keskineliövirhe ja tarkkuus

- Parametrin θ estimaattorin $\hat{\theta}$ **keskineliövirhe** on

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}) &= \mathrm{E}\left[(\hat{\theta}-\theta)^2\right] \\ &= \mathrm{Var}(\hat{\theta}) + \left[\mathrm{Bias}(\hat{\theta})\right]^2\end{aligned}$$

- Jos $\hat{\theta}$ on *harhaton* parametrille θ eli

$$\mathrm{Bias}(\hat{\theta}) = \theta - \mathrm{E}(\hat{\theta}) = 0$$

niin

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta})$$

- Estimaattoria sanotaan **tarkaksi**, jos se on *harhaton* ja sen *varianssi* on *pieni*.

Tehokkuus

- Olkoot $\hat{\theta}_1$ ja $\hat{\theta}_2$ kaksi parametrin θ estimaattoria.
- Estimaattori $\hat{\theta}_1$ on **tehokkaampi** kuin estimaattori $\hat{\theta}_2$, jos
$$\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2)$$
- Parametrin θ estimaattori $\hat{\theta}$ **täystehokas**, jos sen varianssi on *pienempi kuin minkä tahansa muun parametrin* θ *estimaattorin*.

Harhattomuus ja tehokkuus: Esimerkki 1/2

- Olkoon

  $$X_1 , X_2 , \ldots , X_n$$

  satunnaisotos normaalijakaumasta N(μ, σ^2).
- Tällöin sekä havaintoarvojen *aritmeettinen keskiarvo*

  $$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  että havaintoarvojen *mediaani Me* ovat *harhattomia* normaalijakauman odotusarvolle μ :

  $$\mathrm{E}(\bar{X}) = \mathrm{E}(Me) = \mu$$

Harhattomuus ja tehokkuus: Esimerkki 2/2

- Sen sijaan

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n} < \frac{\pi}{2} \cdot \frac{\sigma^2}{n} = \mathrm{Var}(Me)$$

  joten normaalijakautuneiden havaintojen aritmeettisen keskiarvon $\bar{X}$ varianssi on *pienempi* kuin niiden mediaanin Me varianssi.
- Siten havaintoarvojen aritmeettinen keskiarvo $\bar{X}$ on normaalijakauman odotusarvon μ estimaattorina *tehokkaampi* kuin havaintoarvojen mediaani Me.
- Voidaan osoittaa, että havaintoarvojen aritmeettinen keskiarvo $\bar{X}$ on normaalijakauman odotusarvoparametrin μ estimaattorina *täystehokas*.

Tarkentuvuus

- Olkoon $\hat{\theta}$ parametrin θ estimaattori.
- Estimaattori $\hat{\theta}$ on θ **tarkentuva** parametrille θ, jos se *konvergoi* melkein varmasti *kohti parametrin oikeata arvoa*, kun otoskoon n annetaan kasvaa rajatta:

$$\Pr(T_n \to \theta) = 1, \text{ kun } n \to +\infty$$

Ks. monisteen **Todennäköisyyslaskenta** lukua **Stokastiikan konvergenssikäsitteet ja raja-arvolauseet.**

Tarkentuvuus: Esimerkki

- Olkoon

$$X_1\,,X_2\,,\ldots\,,X_n$$

satunnaisotos normaalijakaumasta N(μ, σ^2).

- Tällöin havaintoarvojen *aritmeettinen keskiarvo*

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

on normaalijakauman odotusarvoparametrin μ *tarkentuva* estimaattori:

$$\Pr(\bar{X}\to\mu)=1\,,\ \text{kun}\ n\to\infty$$