Johdatus tilastotieteeseen

Estimointi

Estimointi

Todennäköisyysjakaumien parametrit ja niiden estimointi

Hyvän estimaattorin ominaisuudet

Estimointi: Mitä opimme? – 1/4

- **Tilastollisen tutkimuksen** tavoitteena on tehdä *johtopäätöksiä* **prosesseista**, *jotka* **generoivat** *reaalimaailman ilmiöitä koskevia havaintoja*.
- Tavoitteeseen pyritään rakentamalla **tilastollinen malli** sille prosessille, joka on generoinut tutkimuksen kohteena olevaa ilmiötä koskevat havainnot.
- Koska tilastollisissa tutkimusasetelmissa havaintoihin liittyy aina **satunnaisuutta** tai **epävarmuutta**, tilastolliset mallit ovat luonteeltaan **todennäköisyysmalleja**.
- Tilastollinen malli on täysin määrätty, *jos havaintojen* **todennäköisyysjakauma** *tunnetaan*.

Estimointi:
Mitä opimme? – 2/4

- *Havaintojen todennäköisyysjakauman* määräävät jakauman karakteristisia ominaisuuksia kuvaavat **parametrit**, *joiden arvoja ei sovellustilanteessa yleensä tunneta*.
- Jos jakauman tuntemattomille parametreille ei löydetä hyviä **estimaatteja** eli *arvioita*, jakaumaa ei voida käyttää mallina sille prosessille, joka on generoinut tutkimuksen kohteena olevaa ilmiötä koskevat havainnot.
- Tilastollisen tutkimuksen tärkeimpiä osatehtäviä on **estimoida** eli *arvioida havainnot generoineen prosessin mallina käytettävän todennäköisyysjakauman tuntemattomat parametrit ilmiötä koskevien havaintojen perusteella*.

Estimointi:
Mitä opimme? – 3/4

- *Havaintojen funktiota*, joka tuottaa **estimaatteja** eli *arvioita* todennäköisyysjakauman tuntemattoman parametrin todelliselle arvolle, kutsutaan parametrin **estimaattoriksi**.
- Tilastotieteen tärkeimpiä osatehtäviä on hyvien estimaattoreiden *johtaminen* todennäköisyysjakauman parametreille.

  Ks. lukua **Estimointimenetelmät**.
- Koska todennäköisyysjakauman parametreille voidaan muodostaa erilaisia estimaattoreita, estimaattoreille on esitetty erilaisia **hyvyys-kriteereitä**, joita käytetään apuna estimaattorin valinnassa.
- *Vaatimukset hyvälle estimaattorille*:

  Hyvän estimaattorin on oltava **harhaton**, **tehokas**, **tyhjentävä** ja **tarkentuva**.

Estimointi:
Mitä opimme? – 4/4

- Todennäköisyysjakauman tuntemattomien parametrien *arvojen määräämistä* kutsutaan tavallisesti **piste-estimoinniksi**.
- Todennäköisyysjakauman parametrin estimaattiin on aina syytä liittää **luottamusväliksi** kutsuttu väli, *joka sisältää parametrin todellisen arvon*, *tietyllä*, *soveltajan valittavissa olevalla todennäköisyydellä*.

  Ks. lukua **Väliestimointi**.

Estimointi: Esitiedot

- Esitiedot: ks. seuraavia lukuja:

  Tilastollisten aineistojen kerääminen ja mittaaminen

  Tilastollisten aineistojen kuvaaminen

  Otos ja otosjakaumat

- Tarvitset esitietoja myös seuraavista kalvokokoelman **Johdatus todennäköisyyslaskentaan** luvuista:

  Satunnaismuuttujat ja todennäköisyysjakaumat

  Jakaumien tunnusluvut

Estimointi: Lisätiedot

- Todennäköisyysjakaumien *parametrien estimaattoreiden johtamis-periaatteita* käsitellään luvussa

  Estimointimenetelmät

- *Luottamusvälien määräämistä* todennäköisyysjakaumien *parametreille* käsitellään luvussa

  Väliestimointi

- Todennäköisyysjakaumien *parametreja koskevien tilastollisten hypoteesien testaamista* käsitellään luvuissa

  Tilastolliset testit

  Testit suhdeasteikollisille muuttujille

- *Jakaumaoletuksien testaamista* käsitellään luvussa

  Yhteensopivuuden, homogeenisuuden ja riippumattomuuden testaaminen

Estimointi

>> **Todennäköisyysjakaumien parametrit ja niiden estimointi**

Hyvän estimaattorin ominaisuudet

Todennäköisyysjakaumien parametrit ja niiden estimointi

Avainsanat

Estimaatti
Estimaattori
Estimointi
Havainto
Havaintoarvo
Otosjakauma
Parametri
Piste-estimointi
Tilastollinen aineisto
Tilastollinen malli
Todennäköisyysjakauma
Yksinkertainen satunnaisotos
Luottamusväli

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.
- **Tilastollisella mallilla** tarkoitetaan *havaintoarvot generoineiden satunnaismuuttujien todennäköisyys-jakaumaa*.

Todennäköisyysjakaumien parametrit 1/2

- Tarkastellaan jotakin tutkimuksen kaikkien mahdollisten kohteiden muodostaman perusjoukon S alkioiden ominaisuutta kuvaavaa *satunnaismuuttujaa* X.
- Oletetaan, että satunnaismuuttuja X noudattaa *toden-näköisyysjakaumaa*, 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

Hyvän estimaattorin ominaisuudet

Avainsanat

- **Estimaattori**
- **Harha**
- **Harhattomuus**
- **Hyvyyskriteeri**
- **Keskineliövirhe**
- **Parametri**
- **Tarkentuvuus**
- **Tehokkuus**
- **Tyhjentävyys**

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:
  - **Harhattomuus**
  - **Tyhjentävyys**
  - **Tehokkuus**
  - **Tarkentuvuus**

Harhattomuus ja tyhjentävyys

- **Harhattomuus**:

  Estimaattori T on parametrin θ *harhaton estimaattori*, jos sen odotusarvo yhtyy parametrin θ arvoon:

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

- **Tyhjentävyys**:

  Estimaattori T on *tyhjentävä* parametrille θ, jos se *käyttää* parametrin arvon estimoimiseen *kaiken otoksessa olevan informaation*.

Tehokkuus ja tarkentuvuus

- **Tehokkuus**:

  Estimaattori T on parametrin θ *tehokas estimaattori*, jos sen *varianssi on pienempi kuin minkä tahansa muun estimaattorin*.

- **Tarkentuvuus**:

  Estimaattori T on parametrin θ *tarkentuva estimaattori*, 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. lukua **Konvergenssikäsitteet ja raja-arvolauseet.**

Estimaattorin harha

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

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

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

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

niin

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

Estimaattorin keskineliövirhe

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

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}\left[(\hat{\theta}-\theta)^2\right]$$

$$= \mathrm{Var}(\hat{\theta}) + \left[\mathrm{Bias}(\hat{\theta})\right]^2$$

- Jos $\hat{\theta}$ on parametrin θ *harhaton* estimaattori 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*.