Tilastolliset menetelmät: Otokset, otosjakaumat ja estimointi

Sisällys

4. Otokset ja otosjakaumat

Tilastollinen aineisto koostuu tutkimuksen kohteita ja niiden olosuhteita kuvaavien muuttujien **havaituista arvoista**. Koska tilastollisissa tutkimusasetelmissa havaintoarvoihin liittyy aina **epävarmuutta** tai **satunnaisuutta**, havaintoarvojen ajatellaan olevan jonkin **satunnaismuuttujan generoimia**. Tilastollisen aineiston **tilastollinen malli** tarkoittaa tämän satunnaismuuttujan **todennäköisyysjakaumaa**.

Voimme ajatella, että *havaintoarvoihin liittyvä satunnaisuus* on seurausta siitä, että havaintoarvot on saatu **arpomalla** käyttämällä *arvontatodennäkköisyyksinä* todennäköisyyksiä siitä todennäköisyysjakaumasta, joka toimii havaintoarvojen vaihtelua kuvaavana tilastollisena mallina. Koska siten havaintoarvot vaihtelevat satunnaisesti arvonnasta toiseen, myös *kaikki havaintoarvoista johdetut suureet* – kuten otostunnusluvut – *vaihtelevat satunnaisesti arvonnasta toiseen.*

Käsittelemme tässä luvussa *tilastollisia aineistoja kuvaavien* **tunnuslukujen** eli **otossuureiden** todennäköisyysjakaumia eli **otosjakaumia**. Tarkastelun kohteena ovat erityisesti **aritmeettisen keskiarvon** ja **otosvarianssin** sekä **suhteellisen frekvenssin** otosjakaumat.

Avainsanat:

Aritmeettinen keskiarvo, Havainto, Havaintoarvo, Keskeinen raja-arvolause, χ^2-jakauma, Normaalijakauma, Otos, Otosjakauma, Otostunnusluku, Otosvarianssi, Riippumattomuus, Satunnaisotos, Suhteellinen frekvenssi, t-jakauma, Tilastollinen aineisto, Tilastollinen malli, Todennäköisyysjakauma

4.1. Satunnaisotos

Tilastolliset aineistot

Tilastollinen aineisto koostuu tutkimuksen kohteita kuvaavien muuttujien *havaituista arvoista*. Siitä, että tilastollisissa tutkimusasetelmissa havaintoarvoihin liittyy aina *epävarmuutta* tai *satunnaisuutta*, seuraa seuraavat kaksi seikkaa:

(i) Tilastollisissa tutkimusasetelmissa ajatellaan, että *havaintoarvot on generoinut ilmiö, joka on luonteeltaan satunnainen*.

(ii) Tilastollisen tutkimuksen kohteita kuvaavat muuttujat tulkitaan *satunnaismuuttujiksi* ja havaintoarvot tulkitaan näiden *satunnaismuuttujien realisoituneiksi arvoiksi*.

Tilastolliset mallit

Tilastollisen aineiston **tilastollisella mallilla** tarkoitetaan niiden satunnaismuuttujien *todennäköisyysjakaumaa, jonka ajatellaan generoineen havainnot*. Havaintoarvojen ajatellaan syntyneen *arpomalla* käyttäen arvontatodennäköisyyksinä aineiston mallina käytetystä todennäköisyysjakaumasta saatavin todennäköisyyksin.

Huomautus:

- Todennäköisyysjakaumat riippuvat tavallisesti *parametreista* eli vakioista, joiden arvoja ei yleensä tunneta.

Kun tilastollisia malleja sovelletaan reaalimaailman ilmiöitä kuvaavien havaintoaineistojen analysointiin, kohdataan tavallisesti seuraavat mallin **parametreja** koskevat ongelmat:

(i) Parametrien arvoja *ei tunneta* ja ne on **estimoitava** eli *arvioitava* havaintoaineistosta; lisätietoja: ks. lukua **Estimointi**.

(ii) Parametrien arvoista on esitetty *oletuksia* tai *väitteitä*, joita halutaan **testata** eli asettaa koetteelle havaintoaineistosta saatua informaatiota vastaan; lisätietoja: ks. lukua **Tilastollinen testaus**.

Tilastollisten mallien parametrien *estimointi* ja *testaus* muodostavat keskeisen osan **tilastollista päättelyä**.

Satunnaisotanta

Satunnaisotos poimitaan *arpomalla* havaintoyksiköt perusjoukosta otokseen. Arpomisessa käytettävää menetelmää kutsutaan **satunnaisotannaksi**. Satunnaisotannassa *sattuma* määrää mitkä perusjoukon alkioista tulevat otokseen.

Jos havaintoyksiköt poimitaan perusjoukosta satunnaisotannalla, pätee seuraava:

(i) **Havaintoyksiköitä kuvaavien muuttujien *havaitut arvot ovat satunnaisia* siinä mielessä, että ne vaihtelevat satunnaisesti otoksesta toiseen**.

(ii) ***Kaikki* havaintoyksiköitä kuvaavien muuttujien *havaituista arvoista lasketut tunnusluvut ovat satunnaisia* siinä mielessä, että ne vaihtelevat satunnaisesti otoksesta toiseen**.

Olkoot

$$X_1, X_2, \dots, X_n$$

riippumattomia ja *identtisesti jakautuneita* satunnaismuuttujia, joiden *pistetodennäköisyys*- tai *tiheysfunktio* on $f(x)$:

$$X_1, X_2, \ldots, X_n \perp$$
$$X_i \quad f(x)\,,\ i = 1, 2, \ldots, n$$

Sanomme tällöin, että satunnaismuuttujat

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

muodostavat **satunnaisotoksen** jakaumasta, jonka *pistetodennäköisyys*- tai *tiheysfunktio* on $f(x)$ ja kutsumme satunnaismuuttujia $X_1, X_2, \ldots, X_n$ **havainnoiksi**. *Otoksen poimimisen jälkeen* satunnaismuuttujat $X_1, X_2, \ldots, X_n$ saavat havaituiksi arvoikseen **havaintoarvot**

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

Merkitsemme tätä seuraavasti:

$$X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$$

Havaintoarvot ovat kiinteitä lukuja, mutta ne vaihtelevat satunnaisesti otoksesta toiseen. Siten satunnaisuus liittyy satunnaisotannassa siihen, että havaintoarvot vaihtelevat toisistaan riippumatta ja satunnaisesti otoksesta toiseen. **Satunnaisuus ei siis liity otannan tuloksena saatuihin havaintoarvoihin**, **vaan otoksen poimintaan**.

Satunnaisotoksen tilastollinen malli

Oletetaan, että

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

havainnot muodostavat *satunnaisotoksen* jakaumasta, jonka *pistetodennäköisyys*- tai *tiheysfunktio* on $f(x)$. Satunnaismuuttujien $X_1, X_2, \ldots, X_n$ *yhteisjakauma* muodostaa **tilastollisen mallin** *havaintoarvojen satunnaiselle vaihtelulle otoksesta toiseen*.

Koska satunnaismuuttujat $X_1, X_2, \ldots, X_n$ on oletettu riippumattomiksi, niin niiden **yhteisjakauma** on muotoa

$$f(x_1, x_2, \ldots, x_n) = f(x_1) \times f(x_2) \times \cdots \times f(x_n)$$

jossa

$$X_i \quad f(x_i)\,,\ i = 1, 2, \ldots, n$$

4.2. Otostunnusluvut ja otosjakaumat

Otostunnusluvut

Oletetaan, että havainnot

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

muodostavat *satunnaisotoksen* jakaumasta, jonka *pistetodennäköisyys*- tai *tiheysfunktio* on $f(x)$ ja olkoon

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

jokin *satunnaismuuttujien* $X_1, X_2, \ldots, X_n$ (mitallinen) *funktio*. *Satunnaismuuttujaa T* kutsutaan (*otos*-) **tunnusluvuksi**.

Oletetaan, että otoksen poimimisen jälkeen satunnaismuuttujat $X_1, X_2, \ldots, X_n$ saavat havaituiksi arvoikseen *havaintoarvot* $x_1, x_2, \ldots, x_n$:

$$X_1 = x_1 \,,\; X_2 = x_2, \ldots, X_n = x_n$$

Tällöin tunnusluku

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

saa havaituksi arvokseen t funktion g arvon pisteessä $(x_1, x_2, \ldots, x_n)$:

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

Otosjakauma

Oletetaan, että havainnot

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

muodostavat *satunnaisotoksen*, jonka *pistetodennäköisyys*- tai *tiheysfunktio* on $f(x)$ ja olkoon

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

jokin *otostunnusluku*. Koska tunnusluku T on *satunnaismuuttuja*, sillä on *todennäköisyysjakauma*, jota kutsutaan *tunnusluvun T* **otosjakaumaksi**. Tunnusluvun T otosjakauma muodostaa *tilastollisen mallin* eli *todennäköisyysmallin tunnusluvun T arvojen satunnaisvaihtelulle otoksesta toiseen*.

Huomautus:

- Otosjakauma on tavallisesti *epäoperationaalinen* siinä mielessä, että se riippuu *tuntemattomista* parametreista. Otostunnuslukujen otosjakaumien johtaminen on kuitenkin teoreettisesti tärkeää, koska niillä on tärkeä rooli todennäköisyysjakaumien parametreja koskevassa *estimointi*- ja *testiteoriassa*.

Otosjakaumat: Esimerkkejä

Tutkimme alla seuraavia otosjakaumia:

- **Aritmeettisen keskiarvon ja otosvarianssin otosjakaumat**
- **Suhteellisen frekvenssin otosjakauma**

4.3. Aritmeettisen keskiarvon ja otosvarianssin otosjakaumat

Aritmeettinen keskiarvo ja otosvarianssi

Oletetaan, että havainnot

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

muodostavat *satunnaisotoksen* jakaumasta, jonka odotusarvo on μ ja varianssi on σ^2. Tällöin havainnot $X_1, X_2, \ldots, X_n$ ovat *riippumattomia satunnaismuuttujia*, joilla kaikilla on *sama odotus-arvo* ja *varianssi*:

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

$$\mathrm{E}(X_i) = \mu \,,\; i = 1, 2, \ldots, n$$

$$\mathrm{Var}(X_i) = \mathrm{D}^2(X_i) = \sigma^2 \text{ , } i = 1, 2, \ldots , n$$

Otoksen $X_1, X_2, \ldots , X_n$ ominaisuuksia voidaan kuvata havaintoarvojen *aritmeettisella keskiarvolla* ja *varianssilla*: Määritellään havaintojen $X_1, X_2, \ldots , X_n$ *aritmeettinen keskiarvo* kaavalla

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

ja määritellään havaintojen $X_1, X_2, \ldots , X_n$ *otosvarianssi* kaavalla

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Sekä aritmeettinen keskiarvo $\bar{X}$ että otosvarianssi s^2 ovat *satunnaismuuttujia*, joiden saamat arvot vaihtelevat satunnaisesti otoksesta toiseen.

Aritmeettisen keskiarvon odotusarvo ja varianssi

Oletetaan, että havainnot

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

muodostavat *satunnaisotoksen* jakaumasta, jonka odotusarvo on μ ja varianssi on σ^2. Tällöin havaintojen $X_1, X_2, \ldots , X_n$ *aritmeettisella keskiarvolla*

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

on seuraava *odotusarvo* ja *varianssi*:

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

$$\mathrm{Var}(\bar{X}) = \mathrm{D}^2(\bar{X}) = \frac{\sigma^2}{n}$$

Aritmeettisen keskiarvon $\bar{X}$ *standardipoikkeamaa*

$$\mathrm{D}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

kutsutaan tavallisesti **keskiarvon keskivirheeksi** ja se kuvaa aritmeettisen keskiarvon otosvaihtelua oman odotusarvonsa μ ympärillä.

Perustelu:

Olkoot $X_1, X_2, \ldots , X_n$ *riippumattomia* satunnaismuuttujia, joille

$$\mathrm{E}(X_i) = \mu \text{ , } i = 1, 2, \ldots , n$$

$$\mathrm{Var}(X_i) = \sigma^2 \text{ , } i = 1, 2, \ldots , n$$

Odotusarvon yleisten ominaisuuksien perusteella pätee (myös *ilman riippumattomuus-oletusta*):

$$\mathrm{E}(\bar{X}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n\mu = \mu$$

Varianssin yleisten ominaisuuksien perusteella pätee (*koska satunnaismuuttujat* $X_1, X_2, \ldots, X_n$ *on oletettu riippumattomiksi*):

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \overset{\perp}{=} \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

■

Aritmeettisen keskiarvon otosjakauma: Käyttytyminen otoskoon kasvaessa

Koska aritmeettisen keskiarvon $\bar{X}$ odotusarvo on

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

ja varianssi on

$$\mathrm{Var}(\bar{X}) = \mathrm{D}^2(\bar{X}) = \frac{\sigma^2}{n}$$

niin aritmeettisen keskiarvon otosjakauma *keskittyy yhä voimakkaammin havaintojen yhteisen odotusarvon* μ *ympärille, kun otoskoko n kasvaa*.

Aritmeettisen keskiarvon otosjakauma: Normaalijakautunut otos

Oletetaan, että havainnot

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

muodostavat *satunnaisotoksen normaalijakaumasta* $\mathrm{N}(\mu,\sigma^2)$. Tällöin havaintojen $X_1, X_2, \ldots, X_n$ aritmeettinen keskiarvo $\bar{X}$ noudattaa *normaalijakaumaa* parametrein μ ja σ^2/n :

$$\bar{X} \quad \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Perustelu:

Olkoon $X_1, X_2, \ldots, X_n$ *otos normaalijakaumasta*

$$\mathrm{N}(\mu,\sigma^2)$$

Koska oletuksen mukaan havainnot $X_1, X_2, \ldots, X_n$ ovat *riippumattomia*, niin

$$\sum_{i=1}^{n} X_i \quad \mathrm{N}(n\mu, n\sigma^2)$$

ja

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Yksityiskohdat: Ks. todistusta normaalijakautuneen otoksen aritmeettisen keskiarvon ja otosvarianssin s^2 riippumattomuudelle tässä samassa kappaleessa sekä monisteen **Todennäköisyyslaskenta** lukuja **Jatkuvia jakaumia**, **Moniulotteiset satunnaismuuttujat ja todennäköisyysjakaumat** sekä **Satunnaismuuttujien muunnokset ja niiden jakaumat**.

■

Aritmeettisen keskiarvon otosjakauma: Asymptoottinen jakauma

Oletetaan, että havainnot

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

muodostavat *satunnaisotoksen* jakaumasta, jonka odotusarvo on μ ja varianssi on σ^2. Tällöin *keskeisestä raja-arvolauseesta* seuraa, että havaintojen aritmeettinen keskiarvo $\bar{X}$ noudattaa *suurissa otoksissa approksimatiivisesti* (*asymptoottisesti*) *normaalijakaumaa* parametrein μ ja σ^2/n :

$$\bar{X} \underset{a}{\sim} \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Standardoidun aritmeettisen keskiarvon otosjakauma: Odotusarvo ja varianssi

Koska

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

$$\mathrm{Var}(\bar{X}) = \mathrm{D}^2(\bar{X}) = \frac{\sigma^2}{n}$$

niin *standardoidun* satunnaismuuttujan

$$Z = \frac{\bar{X} - \mathrm{E}(\bar{X})}{\mathrm{D}(\bar{X})} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

odotusarvo ja *varianssi* ovat

$$\mathrm{E}(Z) = 0$$

$$\mathrm{Var}(Z) = 1$$

Standardoidun aritmeettisen keskiarvon otosjakauma: Normaalijakautunut otos

Oletetaan, että havainnot

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

muodostavat *satunnaisotoksen normaalijakaumasta* $\mathrm{N}(\mu,\sigma^2)$. Tällöin *standardoitu satunnais-muuttuja*

$$Z = \frac{\bar{X} - \mathrm{E}(\bar{X})}{\mathrm{D}(\bar{X})} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

noudattaa eksaktisti (eli myös äärellisissä otoksissa) *standardoitua normaalijakaumaa* N(0,1):

$$Z \sim \mathrm{N}(0,1)$$

Standardoidun aritmeettisen keskiarvon otosjakauma: Asymptoottinen jakauma

Oletetaan, että havainnot

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

muodostavat *satunnaisotoksen jakaumasta*, jonka odotusarvo on μ ja varianssi on σ^2. Tällöin *keskeisestä raja-arvolauseesta* seuraa, että standardoitu satunnaismuuttuja

$$Z = \frac{\bar{X} - \mathrm{E}(\bar{X})}{\mathrm{D}(\bar{X})} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

noudattaa *suurissa otoksissa approksimatiivisesti* (*asymptootttisesti*) *standardoitua normaalijakaumaa* N(0,1):

$$Z \sim \mathrm{N}(0,1)$$

Aritmeettisen keskiarvon otosjakauma: Kommentteja

Oletukset havaintojen *riippumattomuudesta*, *samasta jakaumasta* ja *normaalisuudesta* ovat välttämättömiä aritmeettisen keskiarvon *eksaktia* eli *tarkkaa* otosjakaumaa koskevalle äärellisen otoskoon tulokselle.

Aritmeettisen keskiarvon otosjakaumaa koskeva asymptoottinen tulos seuraa **keskeisestä raja-arvolauseesta**; ks. monisteen **Todennäköisyyslaskenta** lukua **Jatkuvia jakaumia** tai lukua **Konvergenssikäsitteet ja raja-arvolauseet**.

Aritmeettisen keskiarvon otosjakaumaa koskeva asymptoottinen tulos pätee tietyin lisäehdoin myös monissa sellaisissa tilanteissa, joissa havaintojen riippumattomuutta ja samaa jakaumaa koskevat oletukset eivät päde.

Otosvarianssin odotusarvo ja varianssi

Oletetaan, että havainnot

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

muodostavat *satunnaisotoksen* jakaumasta, jonka odotusarvo on μ ja varianssi on σ^2. Tällöin havaintojen $X_1, X_2, \ldots , X_n$ *otosvarianssilla*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

on seuraava *odotusarvo*:

$$\mathrm{E}(s^2) = \sigma^2$$

Jos lisäksi voidaan olettaa, että havainnot $X_1, X_2, \ldots , X_n$ noudattavat *normaalijakaumaa* N(μ,σ^2)., niin **otosvarianssin** s^2 **varianssi** on

$$\mathrm{Var}(s^2) = \mathrm{D}^2(s^2) = \frac{2\sigma^4}{n-1}$$

Siten otosvarianssin s^2 *keskivirhe* on *normaalisen otoksen tapauksessa*

$$\mathrm{D}(s^2) = \sigma^2\sqrt{\frac{2}{n-1}}$$

Otosvarianssin otosjakauma: Normaalijakautunut otos

Oletetaan, että havainnot

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

muodostavat *satunnaisotoksen normaalijakaumasta* $\mathrm{N}(\mu,\sigma^2)$ ja olkoon s^2 havaintojen $X_1, X_2, \ldots, X_n$ otosvarianssi. Määritellään satunnaismuuttujat

$$Y = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2$$

ja

$$V = \sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 = \frac{(n-1)s^2}{\sigma^2}$$

Tällöon satunnaismuuttuja Y noudattaa χ^2*-jakaumaa* vapausastein n:

$$Y \quad \chi^2(n)$$

ja satunnaismuuttuja $V = (n-1)s^2/\sigma^2$ noudattaa χ^2*-jakaumaa* vapausastein $(n-1)$:

$$V = \frac{(n-1)s^2}{\sigma^2} \quad \chi^2(n-1)$$

Huomautus:

- Satunnaismuuttuja V on saatu satunnaismuuttujasta Y korvaamalla tavallisesti tuntematon parametri μ sitä vastaavalla otossuureella $\bar{X}$.

Perustelu:

Olkoon $X_1, X_2, \ldots, X_n$ *otos normaalijakaumasta*

$$\mathrm{N}(\mu,\sigma^2)$$

Olkoon

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

havaintojen $X_1, X_2, \ldots, X_n$ *aritmeettinen keskiarvo* ja

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

niiden *otosvarianssi*.

Määritellään satunnaismuuttuja Y kaavalla

$$Y = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2$$

Koska havainnot $X_1, X_2, \ldots, X_n$ ovat *riippumattomia* ja noudattavat normaalijakaumaa

$$X_i \quad \mathrm{N}(\mu,\sigma^2),\ i = 1, 2, \mathrm{K}, n$$

niin myös *standardoidut satunnaismuuttujat*

$$Y_i = \frac{X_i-\mu}{\sigma},\ i = 1, 2, \mathrm{K}, n$$

ovat *riippumattomia* ja noudattavat standardoitua normaalijakaumaa N(0,1):

$$Y_i \sim \mathrm{N}(0,1),\ i = 1, 2, \ldots, n$$

Edellä esitetystä seuraa, että satunnaismuuttuja Y on *riippumattomien, standardoitua normaalijakaumaa* N(0,1) *noudattavien satunnaismuuttujien* Y_i, i = 1, 2, … , n *neliösumma*:

$$Y = \sum_{i=1}^{n} Y_i^2$$

Suoraan *χ^2-jakauman määritelmästä* seuraa, että satunnaismuuttuja Y noudattaa *χ^2-jakaumaa vapausastein n*:

$$Y \sim \chi^2(n)$$

Ks. monisteen **Todennäköisyyslaskenta** lukua **Normaalijakaumasta johdettuja jakaumia**.

Määritellään nyt satunnaismuuttuja V kaavalla

$$V = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma} \right)^2$$

Satunnaismuuttuja V saadaan satunnaismuuttujasta

$$Y = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2$$

korvaamalla odotusarvo μ sitä vastaavalla otossuureella $\bar{X}$.

Satunnaismuuttujan V määritelmässä esiintyvän summan termit

$$U_i = \frac{X_i - \bar{X}}{\sigma},\ i = 1, 2, \ldots, n$$

eivät ole riippumattomia. Voidaan kuitenkin osoittaa, että V voidaan esittää *riippumattomien, standardoitua normaalijakaumaa* N(0,1) *noudattavien satunnaismuuttujien* V_i, i = 1, 2, … , n – 1 *neliösummana*; (ks. todistusta normaalijakautuneen otoksen aritmeettisen keskiarvon $\bar{X}$ ja otosvarianssin s^2 riippumattomuudelle tässä samassa kappaleessa):

$$V = \sum_{i=1}^{n-1} V_i^2$$

Suoraan *χ^2-jakauman määritelmästä* seuraa, että satunnaismuuttuja V noudattaa *χ^2-jakaumaa vapausastein* (n – 1):

$$V \sim \chi^2(n-1)$$

Ks. monisteen **Todennäköisyyslaskenta** lukua **Normaalijakaumasta johdettuja jakaumia**.

■

Huomautuksia:

- Satunnaismuuttuja Y noudattaa χ^2-jakaumaa, jossa *vapausasteiden lukumäärä on sama kuin havaintojen lukumäärä n*.
- Kun satunnaismuuttujasta Y siirrytään satunnaismuuttujaan V *menetetään yksi vapausaste*.

- Yhden vapausasteen menetys on seurausta siitä, että *parametrin* μ *korvaaminen vastaavalla otossuureella* $\bar{X}$ riippumattomissa satunnaismuuttujissa

  $$Y_i = \frac{X_i - \mu}{\sigma}, \ i = 1, 2, \ldots, n$$

  luo yhden (*lineaarisen*) *side-ehdon satunnaismuuttujien*

  $$U_i = \frac{X_i - \bar{X}}{\sigma}, \ i = 1, 2, \ldots, n$$

  välille.

Otosvarianssin otosjakauma: Komentteja

Oletukset havaintojen *riippumattomuudesta*, *samasta jakaumasta* ja *normaalisuudesta* ovat välttämättömiä otosvarianssin *eksaktia* eli *tarkkaa* otosjakaumaa koskevalle äärellisen otoskoon tulokselle.

Aritmeettisen keskiarvon ja otosvarianssin riiippumattomuus ja otosjakaumat: Normaalijakautunut otos

Oletetaan, että havainnot

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

muodostavat *satunnaisotoksen normaalijakaumasta* $\mathrm{N}(\mu,\sigma^2)$. Tällöin havaintojen aritmeettinen keskiarvo

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

ja otosvarianssi

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

ovat satunnaismuuttujina *riippumattomia*:

$$\bar{X} \perp s^2$$

Lisäksi

$$\bar{X} \quad \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{(n-1)s^2}{\sigma^2} \quad \chi^2(n-1)$$

Perustelu:

Oletetaan, että havainnot $X_1, X_2, \ldots, X_n$ muodostavat *satunnaisotoksen normaalijakaumasta* $\mathrm{N}(\mu,\sigma^2)$ ja olkoon

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

havaintojen *aritmeettinen keskiarvo* ja

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

niiden *otosvarianssi*.

Otoksen *yhteisjakauman tiheysfunktio* voidaan kirjoittaa havaintojen riippumattomuuden ja normaalisuuden takia seuraavaan muotoon:

$$f(x_1, x_2, \ldots, x_n) = (2\pi)^{-\frac{1}{2}n}\sigma^{-n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}$$

Määritellään *lineaarinen* muunnos

$$\begin{cases} Y_1 = \frac{1}{\sqrt{n}}X_1 + \frac{1}{\sqrt{n}}X_2 + \frac{1}{\sqrt{n}}X_3 + \cdots + \frac{1}{\sqrt{n}}X_n \\ Y_2 = \frac{1}{\sqrt{2}}X_1 - \frac{1}{\sqrt{2}}X_2 \\ Y_3 = \frac{1}{\sqrt{6}}X_1 + \frac{1}{\sqrt{6}}X_2 - \frac{2}{\sqrt{6}}X_3 \\ \vdots \\ Y_n = \frac{1}{\sqrt{n(n-1)}}X_1 + \frac{1}{\sqrt{n(n-1)}}X_2 + \frac{1}{\sqrt{n(n-1)}}X_3 + \cdots - \frac{n-1}{\sqrt{n(n-1)}}X_n \end{cases}$$

Muunnos voidaan esittää *matriisein* muodossa

$$\mathbf{Y} = \mathbf{BX}$$

jossa

$$\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$$
$$\mathbf{X} = (X_1, X_2, \ldots, X_n)$$

ja $n\times n$-matriisi

$$\mathbf{B} = \begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & -\frac{n-1}{\sqrt{n(n-1)}} \end{bmatrix}$$

on *ortogonaalinen*:

B´B = **BB´** = **I**).

Matriisi **B** nähdään ortogonaaliseksi seuraavalla tavalla: Määritellään $n\times n$-matriisi

$$\mathbf{C} = \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 & 1 \\ 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 1 & -2 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 1 & -3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 1 & 1 & 1 & \cdots & -(n-2) & 0 \\ 1 & 1 & 1 & 1 & \cdots & 1 & -(n-1) \end{bmatrix}$$

On helppo nähdä, että matriisin **C** rivit ovat *kohtisuorassa* toisiaan vastaan. Matriisi **B** saadaan matriisista **C** *normeeraamalla* sen rivit niin, että niiden pituudeksi tulee 1.

Koska muunnos

$$\mathbf{Y} = \mathbf{BX}$$

on ortogonaalinen, niin muunnosta vastaavan *Jacobin determinantin* itseisarvo = 1.

Koska

$$Y_1 = \frac{1}{\sqrt{n}}(X_1 + X_2 + \cdots + X_n) = \sqrt{n}\bar{X}$$

ja

$$\begin{aligned} Y_1^2 + Y_2^2 + \cdots + Y_n^2 &= \mathbf{Y'Y} = \mathbf{X'B'BX} = \mathbf{X'X} = X_1^2 + X_2^2 + \cdots + X_n^2 \\ &= \sum_{i=1}^{n}(X_i - \bar{X})^2 + n\bar{X}^2 \end{aligned}$$

niin

$$Y_2^2 + \cdots + Y_n^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 = (n-1)s^2$$

Koska

$$\begin{aligned} \sum_{i=1}^{n}(X_i - \mu)^2 &= \sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(\bar{X} - \mu)^2 \\ &= \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2 \\ &= Y_2^2 + \cdots + Y_n^2 + (Y_1 - \sqrt{n}\mu)^2 \end{aligned}$$

niin satunnaismuuttujien $Y_1, Y_2, \ldots, Y_n$ *yhteisjakauman tiheysfunktioksi* saadaan

$$\begin{aligned} f(y_1, y_2, \ldots, y_n) &= \frac{1}{(2\pi)^{\frac{1}{2}n}\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\left[(Y_1 - \sqrt{n}\mu)^2 + Y_2^2 + \cdots + Y_n^2\right]\right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(Y_1 - \sqrt{n}\mu)^2\right\} \\ &\quad \times\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}Y_2^2\right\}\times\cdots\times\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}Y_n^2\right\} \end{aligned}$$

Edellä esitetystä seuraa, että satunnaismuuttujat $Y_1, Y_2, \ldots, Y_n$ ovat *riippumattomia* ja *normaalijakautuneita*:

$$Y_1, Y_2, \ldots, Y_n \perp$$
$$Y_1 = \sqrt{n}\bar{X} \quad \mathrm{N}(\sqrt{n}\mu, \sigma^2)$$
$$Y_i \quad \mathrm{N}(0, \sigma^2),\ i = 2, \ldots, n$$

Lisäksi

$$s^2 = \frac{1}{n-1}(Y_2^2 + \cdots + Y_n^2) = \frac{\sigma^2}{n-1}\left[\left(\frac{Y_2}{\sigma}\right)^2 + \cdots + \left(\frac{Y_n}{\sigma}\right)^2\right]$$

jossa

$$\frac{Y_i}{\sigma} \quad \mathrm{N}(0,1),\ i = 2, \ldots, n$$

Siten olemme todistaneet, että

$$\bar{X} \perp s^2$$
$$\bar{X} \quad \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$
$$\frac{(n-1)s^2}{\sigma^2} \quad \chi^2(n-1)$$

■

Huomautus:

- Todistuksessa on sovellettu monisteen **Todennäköisyyslaskenta** luvun **Satunnaismuuttujien muunnokset ja niiden jakaumat** teoriaa sekä luvussa **Normaalijakaumasta johdettuja jakaumia** esitettyä χ^2-jakauman määritelmää.

Oletetaan, että havainnot

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

muodostavat *satunnaisotoksen normaalijakaumasta* N(μ,σ^2) ja ja olkoon

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

havaintojen *aritmeettinen keskiarvo* ja

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

niiden *otosvarianssi*. Tällöin

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \quad t(n-1)$$

Perustelu:

Oletetaan, että havainnot X_1, X_2, … , X_n muodostavat *yksinkertaisen satunnaisotoksen normaalijakaumasta* N(μ,σ^2) ja olkoon

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

havaintojen *aritmeettinen keskiarvo* ja

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

niiden *otosvarianssi*. Edellä on todettu, että

$$\bar{X} \quad \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{(n-1)s^2}{\sigma^2} \quad \chi^2(n-1)$$

ja lisäksi

$$\bar{X} \perp s^2$$

Aritmeettista keskiarvoa $\bar{X}$ koskevasta jakaumatuloksesta seuraa, että

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \quad \mathrm{N}(0,1)$$

Siten suoraan *t-jakauman* määritelmästä seuraa, että

$$t = \frac{\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{1}{n-1}\left(\dfrac{(n-1)s^2}{\sigma^2}\right)}} = \frac{\bar{X}-\mu}{s/\sqrt{n}} \quad t(n-1)$$

Ks. monisteen **Todennäköisyyslaskenta** lukua **Normaalijakaumasta johdettuja jakaumia**.

■

4.4. Suhteellisen frekvenssin otosjakauma

Frekvenssi ja suhteellinen frekvenssi

Olkoon P jokin otosvaruuden S alkioiden ominaisuu. Jos alkiolla x on ominaisuus P merkitään

$$P(x)$$

Olkoon

$$A = \{x \in S \,|\, P(x)\}$$

niiden otosavaruuden S alkioiden joukko, joilla on ominaísuus P. Oletetaan, että tapahtuman A todennäköisyys on

$$\Pr(A) = p$$

jolloin tapahtuman A komplementtitapahtuman A^c todennäköisyys on

$$\Pr(A^c) = 1 - \Pr(A) = 1 - p = q$$

Poimitaan otosavaruudesta S *satunnaisotos*, jonka koko on n. Olkoon A-tyyppisten alkioiden *frekvenssi* eli lukumäärä otoksessa f ja olkoon

$$\hat{p} = \frac{f}{n}$$

vastaava **suhteellinen frekvenssi** eli **osuus**. Sekä frekvenssi f että suhteellinen frekvenssi $\hat{p} = f/n$ ovat *satunnaismuuttujia*, joiden saamat arvot vaihtelevat satunnaisesti otoksesta toiseen.

Frekvenssin odotusarvo, varianssi ja otosjakauma

Frekvenssillä f on seuraava *odotusarvo* ja *varianssi*:

$$\mathrm{E}(f) = np$$

$$\mathrm{Var}(f) = \mathrm{D}^2(f) = npq$$

jossa $q = 1 - p$. Lisäksi *frekvenssi* f noudattaa otoksessa *binomijakaumaa* parametrein n ja $p = \Pr(A)$:

$$f \quad \mathrm{Bin}(n, p)$$

Suhteellisen frekvenssin odotusarvo ja varianssi

Suhteellisella frekvenssillä $\hat{p} = f/n$ on seuraava *odotusarvo* ja *varianssi*:

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

$$\mathrm{Var}(\hat{p}) = \mathrm{D}^2(\hat{p}) = \frac{pq}{n}$$

jossa $q = 1 - p$. Ks. monisteen **Todennäköisyyslaskenta** luvun **Diskreettejä jakaumia** kohtaa **Bernoulli-jakauma** ja kohtaa **Binomijakauma**.

Suhteellisen frekvenssin $\hat{p} = f/n$ *standardipoikkeamaa*

$$\mathrm{D}(\hat{p}) = \sqrt{\frac{pq}{n}}$$

kutsutaan tavallisesti **suhteellisen frekvenssin keskivirheeksi** ja se kuvaa suhteellisen frekvenssin otosvaihtelua oman odotusarvonsa p ympärillä.

Suhteellisen frekvenssin otosjakauma: Käyttäytyminen otoskoon kasvaessa

Koska suhteellisen frekvenssin $\hat{p} = f/n$ odotusarvo on

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

ja varianssi on

$$\mathrm{Var}(\hat{p}) = \mathrm{D}^2(\hat{p}) = \frac{pq}{n}$$

niin suhteellisen frekvenssin otosjakauma *keskittyy yhä voimakkaammin tapahtuman A todennäköisyyden* $\Pr(A) = p$ *ympärille, kun otoskoko n kasvaa*.

Suhteellisen frekvenssin otosjakauma: Asymptoottinen jakauma

Keskeisestä raja-arvolauseen seuraa että suhteellinen frekvenssi $\hat{p} = f/n$ noudattaa em. oletusten pätiessä *suurissa otoksissa approksimatiivisesti* (*asymptoottisesti*) *normaalijakaumaa* parametrein p ja pq/n:

$$\hat{p} \quad {}_{a} \, \mathrm{N}\left(p, \frac{pq}{n}\right)$$

Siten standardoitu satunnaismuuttuja

$$Z = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

noudattaa *suurissa otoksissa approksimatiivisesti* (*asymptoottisesti*) *standardoitua normaalijakaumaa* N(0,1):

$$Z \quad {}_{a} \, \mathrm{N}(0,1)$$

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

5. Estimointi

5.1. Todennäköisyysjakauman parametrit ja niiden estimointi

5.2. Hyvän estimaattorin ominaisuudet

Tilastollinen aineisto koostuu tutkimuksen kohteita kuvaavien muuttujien **havaituista arvoista**. **Tilastollisella mallilla** tarkoitetaan sitä *todennäköisyysjakaumaa, jonka ajatellaan generoineen tutkimuksen kohteena olevan aineiston*.

Koska tämän todennäköisyysjakauman **parametrit** ovat tavallisesti *tuntemattomia,* tilastollosen analyysin eräänä on osatehtävänä on pyrkiä **estimoimaan** eli **arvioimaan** parametrit tutkimuksen kohteena olevaa ilmiötä kuvavasta tilastollisesta aineistosta.

Kuvaamme tässä luvussa *parametrien estimoinnin ongelmia* yleisesti sekä esitelemme **hyvyys-kriteereitä** *estimoinnin onnistumiselle*.

Avainsanat:

Estimaatti, Estimaattori, Estimointi, Estimointimenetelmä, Harha, Harhattomuus, Havainto, Havaintoarvo, Hyvyyskriteeri, Keskineliövirhe, Luottamusväli, Minimivarianssisuus, Momentti-menetelmä, Odotusarvo, Otos, Otosjakauma, Parametri, Piste-estimointi, Satunnaisotos, Suurimman uskottavuuden menetelmä, Tarkentuvuus, Tehokkuus, Tilastollinen aineisto, Tilastollinen malli, Todennäköisyysjakauma, Tyhjentävyys, Täystehokkuus, Varianssi

5.1. Todennäköisyysjakauman parametrit ja niiden estimointi

Tilastolliset aineistot

Tilastollinen aineisto koostuu tutkimuksen kohteita kuvaavien muuttujien *havaituista arvoista*. Siitä, että tilastollisissa tutkimusasetelmissa havaintoarvoihin liittyy aina *epävarmuutta* tai *satunnaisuutta*, seuraa seuraavat kaksi seikkaa:

(i) Tilastollisissa tutkimusasetelmissa ajatellaan, että *havaintoarvot on generoinut ilmiö, joka on luonteeltaan satunnainen*.

(ii) Tilastollisen tutkimuksen kohteita kuvaavat muuttujat tulkitaan *satunnaismuuttujiksi* ja havaintoarvot tulkitaan näiden *satunnaismuuttujien realisoituneiksi arvoiksi*.

Tilastolliset mallit

Tilastollisen aineiston **tilastollisella mallilla** tarkoitetaan niiden satunnaismuuttujien *todennäköisyysjakaumaa, jonka ajatellaan generoineen havainnot*. Havaintoarvojen ajatellaan syntyneen *arpomalla* käyttäen arvontatodennäköisyyksinä aineiston mallina käytetystä todennäköisyysjakaumasta saatavin todennäköisyyksin.

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:

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

Satunnaismuuttujan X pistetodennäköisyys- tai tiheysfunktio $f(x\,;\,\theta)$ kuvaa satunnaismuuttujaan X liittyvien *todennäköisyyksien jakautumista* ja parametri θ kuvaa jotakin jakauman *karakteristista ominaisuutta*.

Kun tilastollisia malleja sovelletaan reaalimaailman ilmiöitä kuvaavien havaintoaineistojen analysointiin, kohdataan tavallisesti seuraavat mallin **parametreja** koskevat ongelmat:

(i) Parametrien arvoja *ei tunneta* ja ne on **estimoitava** eli *arvioitava* havainnoista; **käsittelemme tätä ongelmaa tässä luvussa**.

(ii) Parametrien arvoista on esitetty *oletuksia* tai *väitteitä*, joita halutaan **testata** eli asettaa koetteelle havaintoaineistosta saatua informaatiota vastaan; lisätietoja: ks. lukua **Tilastollinen testaus**.

Tilastollisten mallien parametrien *estimointi* ja *testaus* muodostavat keskeisen osan **tilastollista päättelyä**.

Satunnaisotanta

Satunnaisotos poimitaan *arpomalla* havaintoyksiköt perusjoukosta otokseen. Arpomisessa käytettävää menetelmää kutsutaan **satunnaisotannaksi**. Satunnaisotannassa *sattuma* määrää mitkä perusjoukon alkioista tulevat otokseen.

Jos havaintoyksiköt poimitaan perusjoukosta satunnaisotannalla, pätee seuraava:

(i) **Havaintoyksiköitä kuvaavien muuttujien *havaitut arvot ovat satunnaisia* siinä mielessä, että ne vaihtelevat satunnaisesti otoksesta toiseen**.

(ii) ***Kaikki* havaintoyksiköitä kuvaavien muuttujien *havaituista arvoista lasketut tunnusluvut ovat satunnaisia* siinä mielessä, että ne vaihtelevat satunnaisesti otoksesta toiseen**.

Satunnaisotos

Olkoon

$$X_i\ ,\ i = 1, 2, \ldots , n$$

satunnaisotos jakaumasta, jonka *pistetodennäköisyys*- tai *tiheysfunktio* $f(x;\theta)$ riippuu *parametrista* θ. Tällöin havainnot X_i , $i = 1, 2, \ldots , n$ ovat *riippumattomia*, *identtisesti jakautuneita satunnaismuuttujia*, joilla on *sama pistetodennäköisyys*- tai *tiheysfunktio* $f(x;\theta)$:

$$X_1, X_2, \ldots, X_n \perp$$
$$X_i \quad f(x;\theta)\ ,\ i = 1, 2, \ldots, n$$

Sanomme tällöin, että satunnaismuuttujat

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

muodostavat **satunnaisotoksen** jakaumasta, jonka *pistetodennäköisyys*- tai *tiheysfunktio* on $f(x;\theta)$ ja kutsumme satunnaismuuttujia $X_1, X_2, \ldots , X_n$ **havainnoiksi**. *Otoksen poimimisen jälkeen* satunnaismuuttujat $X_1, X_2, \ldots , X_n$ saavat havaituiksi arvoikseen **havaintoarvot**

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

Merkitsemme tätä seuraavasti:

$$X_1 = x_1, X_2 = x_2, \ldots , X_n = x_n$$

Havaintoarvot ovat kiinteitä lukuja, mutta ne vaihtelevat satunnaisesti otoksesta toiseen. Siten satunnaisuus liittyy satunnaisotannassa siihen, että havaintoarvot vaihtelevat toisistaan riippumatta ja satunnaisesti otoksesta toiseen. **Satunnaisuus ei siis liity otannan tuloksena saatuihin havaintoarvoihin**, **vaan otoksen poimintaan**.

Estimaattorit ja estimaatit

Oletetaan, että havainnot

$$X_i\ ,\ i = 1, 2, \ldots , n$$

muodostavat *satunnaisotoksen* jakaumasta $f(x;\theta)$. Oletetaan, että todennäköisyysjakauman $f(x;\theta)$ *parametri* θ *on tuntematon* ja sen *estimoimiseen* käytetään havaintojen X_i , $i = 1, 2, \ldots , n$ (mitallista) funktiota eli (*otos*-) *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**.

Huomaa, että estimaattorin $T = g(X_1, X_2, \dots , X_n)$ havaintoarvoista $x_1, x_2, \dots , x_n$ lasketut arvot eli estimaatit $t = g(x_1, x_2, \dots , x_n)$ *vaihtelevat satunnaisesti otoksesta toiseen*.

Estimaattorin otosjakauma

Oletetaan, että havainnot

$$X_i , i = 1, 2, \dots , n$$

muodostavat *satunnaisotoksen* jakaumasta $f(x;\theta)$ ja olkoon

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

jokin *parametrin* θ *estimaattori*.

Koska estimaattori T on *satunnaismuuttuja*, sillä on *todennäköisyysjakauma*, jota kutsutaan *estimaattorin T* **otosjakaumaksi**. Estimaattorin T otosjakauma muodostaa *tilastollisen mallin* eli *todennäköisyysmallin estimaattorin T arvojen satunnaiselle vaihtelulle otoksesta toiseen*.

Estimaattoreiden johtaminen

Hyvien estimaattoreiden johtaminen todennäköisyysjakaumien 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**. Tätä ongelmaa käsitellään luvussa **Estimointimenetelmät**.

Parametrin estimaattiin on syytä aina liittää **luottamusväliksi** kutsuttu *väli, joka sisältää estimoidun parametrin todellisen, mutta tuntemattoman arvon tietyllä, soveltajan valittavissa olevalla todennäköisyydellä*. Luottamusvälin määräämistä on tapana kutsua **väliestimoinniksi**.

Ks. lukua **Väliestimointi**.

5.2. Hyvän estimaattorin omininaisuuksia

Todennäköisyysjakauman parametreille on tavallisesti tarjolla useita *vaihtoehtoisia estimaattoreita*. Estimaattorin valintaa ohjaavat **hyvyyskriteerit**, joilla pyritään takamaan se, että valittu estimaattori tuottaa järkeviä arvoja estimoitavalle parametrille.

Oletetaan, että havainnot

$$X_i , i = 1, 2, \dots , n$$

muodostavat *satunnaisotoksen* jakaumasta $f(x;\theta)$ ja olkoon

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

jokin *parametrin* θ *estimaattori*.

Tyhjentävyys

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

Tässä annettu tyhjentävyyden määritelmä *ei ole matemaattisesti kelvollinen*, koska sen perusteella *ei pystytä käytännössä toteamaan onko estimaattori tyhjentävä vai ei*. Määritelmä antaa kuitenkin tyhjentävyyden käsitteen takana olevasta idesta riittävän käsityksen tämän esityksen tarpeisiin. Emme määrittele tyhjentävyyden käsitettä täsmällisesti tässä esityksessä.

Harhattomuus

Estimaattori T on **harhaton** parametrille θ, jos sen odostusarvo yhtyy estimoitavan parametrin θ arvoon:

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

Estimaattorin *harhattomuus* merkitsee sitä, että estimaattori tuottaa *keskimäärin* oikean kokoisia arvoja (estimaatteja) estimoitavalle parametrille. Estimaattorin tuottama arvo parametrille saattaa yksittäisessä tilanteessa (tietylle otokselle) poiketa paljonkin parametrin todellisesta arvosta, mutta odotusarvon *frekvenssitulkinnan* mukaan estimaattorin tuottamat otoskohtaiset arvot parametrille kasautuvat kuitenkin otantaa toistettaessa parametrin todellisen arvon ympärille.

On ilmeistä, että hyvän estimaattorin tuottamat arvot vaihtelevat otoksesta toiseen vain *vähän* parametrin todelllisen arvon ympärillä eli *hyvän estimaattorin varianssi on pieni*. Tätä estimaattorin ominaisuutta kuvataan käsitteellä **tehokkuus**; ks. tehokkuuden määritelmää alla.

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

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}[(\hat{\theta} - \theta)^2] = \mathrm{Var}(\hat{\theta}) + [\mathrm{Bias}(\hat{\theta})]^2$$

Jos $\hat{\theta}$ on parametrin θ *harhaton* estimaattori eli $\mathrm{E}(\hat{\theta}) = \theta$, niin

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

ja siten

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

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

Tehokkuus

Olkoot T_1 ja T_2 kaksi parametrin θ estimaattoria. Estimaattori T_1 on **tehokkaampi** kuin estimaattori T_2, jos *estimaattorin* T_1 *varianssi on pienempi kuin estimaattorin* T_2 *varianssi*:

$$\mathrm{Var}(T_1) < \mathrm{Var}(T_2)$$

Esimerkki 1: Normaalijakautuneen otoksen aritmeettisen keskiarvon ja mediaanin tehokkuus.

Olkoon

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

satunnaisotos normaalijakaumasta $\mathrm{N}(\mu, \sigma^2)$. Estimoidaan jakauman odotusarvoparametri μ havaintojen $X_1, X_2, \ldots , X_n$ *aritmeettisella keskiarvolla*

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

Olemme todenneet luvussa **Otokset ja otosjakaumat**, että estimaattori $\bar{X}$ on *harhaton* odotusarvoparametrille μ:

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

ja estimaattorin $\bar{X}$ varianssi on

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

Voidaan osoittaa, että myös havaintojen $X_1, X_2, \ldots , X_n$ *mediaani* Me on *harhaton* odotusarvoparametrille μ:

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

Sen sijaan estimaattorin Me varianssi on

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

Koska siis

$$\mathrm{Var}(\bar{X}) < \mathrm{Var}(Me)$$

havaintojen aritmeettinen keskiarvo $\bar{X}$ on normaalijakauman odotusarvoparemetrin estimaattorina *tehokkaampi* kuin havaintojen mediaani Me.

Täystehokkuus eli minimivarianssisuus

Estimaattori T on **täystehokas** eli **minimivarianssinen** parametrille θ, jos sen varianssi

$$\mathrm{Var}(T)$$

on *pienempi kuin minkä tahansa muun estimaattorin*.

Minimivarianssisuus on ominaisuus, jota on harvoin mahdollista saavuttaa parametrin kaikkien mahdollisten estimaattoreiden joukossa. Sen sijaan sopivasti rajoitetussa estimaattoreiden luokassa tämä saattaa hyvin olla mahdollista.

Esimerkki 2: Normaalijakautuneen otoksen aritmeettisen keskiarvon minimivarianssisuus.

Olkoon

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

satunnaisotos normaalijakaumasta N(μ, σ^2). Voidaan osoittaa, että havaintojen $X_1, X_2, \ldots , X_n$ *aritmeettisen keskiarvon*

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

varianssi on *pienin kaikkien odostusparametrin* μ *harhattomien estimaattoreiden joukossa*. Siten estimaattori $\overline{X}$ on *minimivarianssinen hahattomien estimaattoreiden luokassa*.

Tarkentuvuus

Estimaattori T on (vahvasti) **tarkentuva** parametrille θ, jos se *konvergoi* melkein varmasti kohti *parametrin* θ *oikeata arvoa*, kun otoskoon n annetaan kasvaa rajatta:

$$\lim_{n\to\infty} \Pr(T_n \to \theta) = 1$$

Lisätietoja stokastiikan konvergenssikäsitteistä: Ks. monisteen **Todennäköisyyslaskenta** lukua **Konvergenssikäsitteet ja raja-arvolauseet**.

6. Estimointimenetelmät

Tilastollinen aineisto koostuu tutkimuksen kohteita kuvaavien muuttujien **havaituista arvoista**.

Tilastollisen aineiston **tilastollisella mallilla** tarkoitetaan niiden satunnaismuuttujien *todennäköisyysjakaumaa, jonka ajatellaan generoineen havainnot*. Koska tämän jakauman **parametrit** ovat tavallisesti *tuntemattomia* ne pyritään on **estimoimaan** eli **arvioimaan** kerättyjen havaintojen perusteella.

Kutsumme parametrin tuntemattoman arvon estimoimiseen käytettävää havaintojen funktiota ko. parametrin **estimaattoriksi** ja sen havaintoarvoista laskettua arvoa ko. parametrin **estimaatiksi**.

Hyvien estimaattoreiden johtaminen tilastollisten mallien parametreille on teoreettisen tilastotieteen keskeisiä ongelmia. Kutsumme estimaattoreiden johtamiseen käytettyjä menetelmiä **estimointimenetelmiksi**.

Tässä luvussa käsitellään kahta keskeistä estimointimenetelmää: **suurimman uskottavuuden menetelmä** ja **momenttimenetelmä**. Lisäksi johdamme kummallakin menetelmällä **normaalijakauman**, **eksponenttijakauman** ja **Bernoulli-jakauman** parametrien estimaattorit.

Avainsanat:

Artimeettinen keskiarvo, Bernoulli-jakauma, Eksponenttijakauma, Estimaatti, Estimaattori, Estimointi, Estimointimenetelmä, Frekvenssi, Harha, Harhattomuus, Havainto, Havaintoarvo, Hyvyyskriteeri, Keskineliövirhe, Logaritminen uskottavuusfunktio, Luottamusväli, Maksimi, Maksimointi, Momentti, Momenttimenetelmä, Normaalijakauma, Odotusarvo, Otos, Otosjakauma, Otosmomentti, Otosvarianssi, Parametri, Piste-estimointi, Satunnaisotos, Suhteellinen frekvenssi, Suurimman uskottavuuden menetelmä, Tarkentuvuus, Tehokkuus, Tilastollinen aineisto, Tilastollinen malli, Todennäköisyys, Todennäköisyysjakauma, Tyhjentävyys, Uskottavuusfunktio, Varianssi

6.1. Estimointi

Satunnaisotos

Olkoon

$$X_i\ ,\ i = 1, 2, \ldots , n$$

satunnaisotos jakaumasta, jonka *pistetodennäköisyys*- tai *tiheysfunktio* $f(x;\theta)$ riippuu *parametrista* θ. Tällöin havainnot X_i , $i = 1, 2, \ldots , n$ ovat *riippumattomia, identtisesti jakautuneita satunnais-muuttujia*, joilla on *sama pistetodennäköisyys*- tai *tiheysfunktio* $f(x;\theta)$:

$$X_1, X_2, \ldots, X_n \perp$$
$$X_i \quad f(x;\theta)\,,\ i = 1, 2, \ldots, n$$

Estimaattori ja estimaatti

Oletetaan, että havainnot

$$X_i\ ,\ i = 1, 2, \ldots , n$$

muodostavat *satunnaisotoksen* jakaumasta $f(x;\theta)$. Oletetaan, että todennäköisyysjakauman $f(x;\theta)$ *parametri* θ *on tuntematon* ja sen *estimoimiseen* käytetään havaintojen X_i , $i = 1, 2, \ldots , n$ (mitallista) funktiota eli (*otos*-) *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**.

Estimaattoreiden johtaminen

Hyvien estimaattoreiden johtaminen todennäköisyysjakaumien tuntemattomille parametreille on teoreettisen tilastotieteen keskeisiä ongelmia.

Tärkeimmät estimaattoreiden johtamiseen käytettävät menetelmät:

- **Suurimman uskottavuuden menetelmä**
- **Momenttimenetelmä**

6.2. Suurimman uskottavuuden menetelmä

Uskottavuusfunktio

Olkoon X_i , $i = 1, 2, \ldots , n$ satunnaisotos jakaumasta $f(x;\theta)$, jonka parametrina on θ. Tällöin

$$X_1, X_2, \ldots, X_n \perp$$
$$X_i \quad f(x;\theta)\,,\ i = 1, 2, \ldots, n$$

Koska olemme olettaneet, että havainnot X_i , $i = 1, 2, \ldots , n$ ovat *riippumattomia* ja *noudattavat samaa jakaumaa* $f(x;\theta)$, *otoksen yhteisjakauman tiheysfunktio* on

$$f(x_1, x_2, \ldots, x_n; \theta) = f(x_1;\theta) \times f(x_2;\theta) \times \cdots \times f(x_n;\theta)$$

jossa

$$f(x_i;\theta),\ i = 1, 2, \ldots, n$$

on yksittäiseen havaintoon X_i liittyvä pistetodennäköisyys- tai tiheysfunktio.

Otoksen X_i , $i = 1, 2, \ldots , n$ **uskottavuusfunktio**

$$L(\theta; x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n; \theta)$$

on havaintojen X_i , $i = 1, 2, \ldots , n$ *yhteisjakauman pistetodennäköisyys*- tai *tiheysfunktion f* arvo pisteessä

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

tulkittuna *parametrin* θ *arvojen funktioksi*.

Huomautus:

- Voimme olettaa, että uskottavuusfunktio L sisältää kaiken *stokastisen informaation* otoksesta.

Suurimman uskottavuuden estimaattori

Olkoon

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

parametrin θ arvo, joka *maksimoi otoksen* X_i , $i = 1, 2, \ldots , n$ *uskottavuusfunktion*

$$L(\theta; x_1, x_2, \ldots, x_n)$$

parametrin θ suhteen.

Huomautus:

- Uskottavuusfunktion L maksimin antava parametrin θ arvo t on muuttujien (= havaintoarvojen) $x_1, x_2, \ldots , x_n$ funktio.

Sijoittamalla uskottavuusfunktion L maksimin parametrin θ suhteen antavassa lausekkeessa

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

muuttujien

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

paikalle havainnot (= satunnaismuuttujat)

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

saadaan parametrin θ **suurimman uskottavuuden estimaattori** eli **SU-estimaattori**

$$\hat{\theta} = g(X_1, X_2, \ldots, X_n)$$

Parametrin θ suurimman uskottavuuden estimaattori $\hat{\theta}$ tuottaa parametrille θ arvon, joka *maksimoi juuri sen otoksen uskottavuuden*, *joka saatiin* eli *juuri niiden havaintoarvojen uskottavuuen*, *jotka saatiin*. Tämä ilmaistaan usein seuraavalla (epätäsmällisellä) tavalla: Parametrin θ suurimman uskottavuuden estimaattorin $\hat{\theta}$ otoskohtainen arvo *maksimoi todennäköisyyden saada juuri se otos, joka saatiin*.

Suurimman uskottavuuden estimaattorin määrääminen

Parametrin θ suurimman uskottavuuden estimaattori määrätään *maksimoimalla uskottavuusfunktio*

$$L(\theta; x_1, x_2, \ldots, x_n)$$

parametrin θ suhteen. Kaikissa säännöllisissä tapauksissa maksimi löydetään merkitsemällä uskottavuusfunktion $L(\theta)$ *derivaatta*

$$L'(\theta)$$

nollaksi ja ratkaisemalla θ saadusta *normaaliyhtälöstä*

$$L'(\theta) = 0$$

Määräämme alla seuraavien jakaumien parametrien suurimman uskottavuuden estimaattorit:

- **Normaalijakauma**
- **Eksponenttijakauma**
- **Bernoulli-jakauma**

Logaritminen uskottavuusfunktio

Uskottavuusfunktion maksimi kannattaa tavallisesti etsiä maksimoimalla uskottavuusfunktion sijasta *uskottavuusfunktion logaritmi* eli **logaritminen uskottavuusfunktio**

$$l(\theta; x_1, x_2, \ldots, x_n) = \log L(\theta; x_1, x_2, \ldots, x_n)$$

Tämä johtuu seuraavista seikoista:

(i) Koska logaritmi on *aidosti monotoninen funktio*, logaritminen uskottavuusfunktio ja uskottavuusfunktio saavuttavat ääriarvonsa *samassa pisteessä*.

(ii) Logaritminen uskottavuusfunktio on monien todennäköisyysjakaumien tapauksessa muodoltaan *yksinkertaisempi* kuin uskottavuusfunktio.

Koska olemme olettaneet, että havainnot X_i , $i = 1, 2, \ldots , n$ ovat riippumattomia ja noudattavat *samaa* jakaumaa $f(x;\theta)$, *logaritminen uskottavuusfunktio* voidaan kirjoittaa seuraavaan muotoon:

$$\begin{aligned} l(\theta) &= \log L(\theta) \\ &= \log\left(f(x_1;\theta)\times f(x_2;\theta)\times \cdots \times f(x_n;\theta)\right) \\ &= \log f(x_1;\theta) + \log f(x_2;\theta) + \cdots + \log f(x_n;\theta) \\ &= l(\theta;x_1) + l(\theta;x_2) + \cdots + l(\theta;x_n) \end{aligned}$$

jossa

$$l(\theta\,;x_i) = \log f(x_i\,;\theta)\ , \ i = 1, 2, \ldots , n$$

on havaintoarvoon x_i liittyvä logaritminen uskottavuusfunktio. Logaritmisen uskottavuusfunktion summaesityksen

$$l(\theta) = l(\theta\,;x_1) + l(\theta\,;x_2) + \cdots + l(\theta\,;x_n)$$

maksimointi on tavallisesti paljon helpompaa kuin uskottavuusfunktion itsensä maksimointi.

Suurimman uskottavuuden estimaattorin asymptoottiset ominaisuudet

Parametrin θ SU-estimaattori $\hat{\theta}$ ei välttämättä täytä hyvän estimaattorin kriteereitä *äärellisillä havaintojen lukumäärillä*. Onneksi SU-estimaattori $\hat{\theta}$ käyttöä parametrin θ estimaattorina voidaan kuitenkin perustella SU-estimaattorin *yleisillä* **asymptoottisilla ominaisuuksilla**:

Hyvin yleisin ehdoin pätee:

(i) SU-estimaattori $\hat{\theta}$ on **tarkentuva** eli

$$\lim_{n\to\infty} \Pr(\hat{\theta} \to \theta) = 1$$

Siten SU-estimaattorin arvo *lähestyy stokastisesti parametrin oikeata arvoa*, kun otoskoon annetaan kasvaa rajatta. Tämä merkistee sitä, että SU-estimaattori toteuttaa **suurten lukujen lain**.

(ii) SU-estimaattori $\hat{\theta}$ on **asymptoottisesti normaalinen**.

Siten *SU-estimaattorin jakaumaa voidaan suurissa otoksissa approksimoida normaalijakaumalla*. Tämä merkitsee sitä, että SU-estimaattori toteuttaa **keskeisen raja-arvolauseen**.

Lisätietoja stokastiikan konvergenssikäsitteistä: ks. monisteen **Todennäköisyyslaskenta** lukua **Stokastiikan konvergenssikäsitteet**.

Huomautus:

- SU-estimaattorin asymptoottinen normaalisuus on tärkeä lisäperuste *normaalijakauman keskeiselle asemalle* tilastotieteessä.

6.3. Normaalijakauman parametrien suurimman uskottavuuden estimointi

Satunnaismuuttuja X noudattaa **normaalijakaumaa** N(μ, σ^2), jos sen *tiheysfunktio* on muotoa

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, -\infty < \mu < +\infty\,, \sigma > 0$$

Normaalijakauman *parametreina* ovat jakauman *odotusarvo*

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

ja *varianssi*

$$\mathrm{Var}(X) = \sigma^2$$

Lisätietoja normaalijakaumasta: Ks. monisteen **Todennäköisyyslaskenta** lukua **Jatkuvia jakaumia**.

SU-estimaattoreiden johto

Olkoon

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

satunnaisotos normaalijakaumasta N(μ, σ^2). Tällöin havainnot $X_1, X_2, \ldots, X_n$ ovat *riippumattomia, samaa normaalijakaumaa* N(μ, σ^2) noudattavia satunnaismuuttujia.

Siten otoksen $X_1, X_2, \ldots, X_n$ *uskottavuusfunktio* on

$$\begin{aligned} &L(\mu,\sigma^2 ; x_1, x_2, \ldots, x_n) \\ &= f(x_1 ; \mu,\sigma^2)\times f(x_2 ; \mu,\sigma^2)\times \cdots \times f(x_n ; \mu,\sigma^2) \\ &= \sigma^{-n}(2\pi)^{-\frac{1}{2}n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \end{aligned}$$

ja sen *logaritminen uskottavuusfunktio* on

$$\begin{aligned} &l(\mu,\sigma^2 ; x_1, x_2, \ldots, x_n) \\ &= \log L(\mu,\sigma^2 ; x_1, x_2, \ldots, x_n) \\ &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2}n\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \end{aligned}$$

Normaalijakauman N(μ,σ^2) odotusarvon μ ja varianssin σ^2 **SU-estimaattorit** ovat havaintojen $X_1, X_2, \ldots, X_n$ *aritmeettinen keskiarvo*

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

ja *otosvarianssi* laskettuna kaavalla, jossa jakajana on havaintojen lukumäärä n:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

Perustelu:

Derivoidaan logaritminen uskottavuusfunktio

$$l(\mu,\sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2}n\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

ensin parametrin μ suhteen ja merkitään derivaatta nollaksi:

$$\frac{\partial l(\mu,\sigma^2)}{\partial\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$$

Derivaatan ainoa *nollakohta*

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$$

antaa logaritimisen uskottavuusfunktion *maksimin* parametrin μ suhteen.

Sijoitetaan ratkaisu $\hat{\mu} = \overline{x}$ logaritmiseen uskottavuusfunktioon:

$$l(\overline{x},\sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2}n\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

Derivoidaan funktio $l(\overline{x},\sigma^2)$ parametrin σ^2 suhteen ja merkitään derivaatta nollaksi:

$$\frac{\partial l(\bar{x},\sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

Derivaatan ainoa *nollakohta*

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

antaa log-uskottavuusfunktion *maksimin* parametrin σ^2 suhteen.

■

Huomaa, että parametrien μ ja σ^2 *suurimman uskottavuuden estimaattorit yhtyvät niiden momentti-estimaattoreihin*; lisätietoja momenttimenetelmästä: ks. kappaletta **Momenttimenetelmä**.

SU-estimaattoreiden ominaisuudet

Normaalijakauman N(μ,σ^2) odotusarvon μ SU-estimaattorilla $\bar{X}$ on seuraavat ominaisuudet:

(i) $\bar{X}$ on *harhaton*.

(ii) $\bar{X}$ ja $\hat{\sigma}^2$ ovat yhdessä *tyhjentäviä* parametreille μ ja σ^2 .

(iii) $\bar{X}$ on *tehokas* eli minimivarianssinen estimaattori.

(iv) $\bar{X}$ on *tarkentuva*.

(v) $\bar{X}$ noudattaa *normaalijakaumaa*:

$$\bar{X} \quad \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Normaalijakauman N(μ,σ^2) varianssin σ^2 SU-estimaattorilla $\hat{\sigma}^2$ on seuraavat ominaisuudet:

(i) $\hat{\sigma}^2$ on *harhainen*, mutta estimaattori

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n}{n-1}\hat{\sigma}^2$$

on *harhaton*.

(ii) $\bar{X}$ ja $\hat{\sigma}^2$ ovat yhdessä *tyhjentäviä* parametreille μ ja σ^2.

(iii) $\hat{\sigma}^2$ ei ole *tehokas* eli minimivarianssinen estimaattori.

(iv) $\hat{\sigma}^2$ on *tarkentuva*.

(v) $(n – 1)\, s^2 /\sigma^2$ noudattaa χ^2 *-jakaumaa*:

$$\frac{(n-1)s^2}{\sigma^2} \quad \chi^2(n-1)$$

6.4. Eksponenttijakauman parametrien suurimman uskottavuuden estimointi

Satunnaismuuttuja X noudattaa **eksponenttijakaumaa** Exp(λ), jos sen *tiheysfunktio* on

$$f(x) = \lambda e^{-\lambda x}\ ,\ x \geq 0\ ,\ \lambda > 0$$

Eksponenttijakauman ainoana *parametrina* on

$$\lambda = \frac{1}{\mathrm{E}(X)}$$

Lisätietoja eksponenttijakaumasta: Ks. monisteen **Todennäköisyyslaskenta** lukua **Jatkuvia jakaumia**.

Su-estimaattorin johto

Olkoon

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

satunnaisotos eksponenttijakaumasta Exp(λ). Tällöin havainnot $X_1, X_2, \ldots , X_n$ ovat *riippumattomia*, *samaa eksponenttijakaumaa* Exp(λ) noudattavia satunnaismuuttujia.

Otoksen $X_1, X_2, \ldots , X_n$ *uskottavuusfunktio* on

$$\begin{aligned} &L(\lambda\,; x_1, x_2, \ldots , x_n) \\ &\quad = f(x_1\,;\lambda) \times f(x_2\,;\lambda) \times \cdots \times f(x_n\,;\lambda) \\ &\quad = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right) \end{aligned}$$

ja sen *logaritminen uskottavuusfunktio* on

$$\begin{aligned} &l(\lambda\,; x_1, x_2, \ldots , x_n) \\ &\quad = \log L(\lambda\,; x_1, x_2, \ldots , x_n) \\ &\quad = n\log(\lambda) - \lambda \sum_{i=1}^{n} x_i \end{aligned}$$

Eksponenttijakauman Exp(λ) parametrin λ **SU-estimaattori** on

$$\hat{\lambda} = \frac{1}{\bar{X}}$$

jossa

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

on havaintojen $X_1, X_2, \ldots , X_n$ *aritmeettinen keskiarvo*.

Perustelu:

Derivoidaan logaritminen uskottavuusfunktio

$$l(\lambda) = n\log(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

parametrin λ suhteen ja merkitään derivaatta nollaksi:

$$\frac{\partial l(\lambda)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

Derivaatan ainoa *nollakohta*

$$\hat{\lambda} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$$

antaa logaritmisen uskottavuusfunktion *maksimin* parametrin λ suhteen.

■

Huomaa, että parametrin λ *suurimman uksottavuuden estimaattori yhtyy sen momentti-estimaattoriin*; lisätietoja momenttimenetelmästä: ks. kappaletta **Momenttimenetelmä**.

Sivuutamme tässä parametrin λsuurimman uskottavuuden estimaattorin stokastiset ominaisuudet.

6.5. Bernoulli-jakauman parametrien suurimman uskottavuuden estimointi

Olkoon A *tapahtuma*, jonka todennäköisyys on p:

$$\Pr(A) = p$$

Määritellään satunnaismuuttuja X seuraavasti:

$$X = \begin{cases} 1, \text{ jos } A \text{ tapahtuu} \\ 0, \text{ jos } A \text{ ei tapahdu} \end{cases}$$

Tällöin satunnaismuuttuja X noudattaa **Bernoulli-jakaumaa** parametrilla p:

$$X \quad \mathrm{Ber}(p)$$

jossa

$$\Pr(A) = p = \mathrm{E}(X)$$

Satunnaismuuttujan X *pistetodennäköisyysfunktio* on

$$f(x) = p^x (1-p)^{1-x}\ ,\ x = 0,1$$

Lisätietoja Bernoulli-jakaumasta: Ks. monisteen **Todennäköisyyslaskenta** lukua **Diskreettejä jakaumia**.

SU-estimaattorin johto

Olkoon

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

satunnaisotos Bernoulli-jakaumasta Ber(p). Tällöin havainnot $X_1, X_2, \ldots , X_n$ ovat *riippumattomia, samaa Bernoulli-jakaumaa* Ber(p) noudattavia satunnaismuuttujia.

Otoksen $X_1, X_2, \ldots , X_n$ *uskottavuusfunktio* on

$$\begin{aligned} &L(p\,;x_1, x_2, \ldots , x_n) \\ &\quad = f(x_1\,;p)\times f(x_2\,;p)\times \cdots \times f(x_n\,;p) \\ &\quad = p^{\Sigma x_i}(1-p)^{n-\Sigma x_i} \end{aligned}$$

ja sen *logaritminen uskottavuusfunktio* on

$$\begin{aligned} l(p\,;x_1,x_2,\ldots,x_n) &= \log L(p\,;x_1,x_2,\ldots,x_n) \\ &= \sum_{i=1}^{n} x_i \log(p) + (n-\sum_{i=1}^{n} x_i)\log(1-p) \end{aligned}$$

Bernoulli-jakauman Ber(p) odotusarvoparametrin p **SU-estimaattori** on havaintojen $X_1, X_2, \ldots, X_n$ *aritmeettinen keskiarvo*

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

Perustelu:

Derivoidaan logaritminen uskottavuusfunktio

$$l(p) = \sum_{i=1}^{n} x_i \log(p) + (n-\sum_{i=1}^{n} x_i)\log(1-p)$$

parametrin p suhteen ja merkitään derivaatta nollaksi:

$$\frac{\partial l(p)}{\partial p} = \frac{\Sigma x_i}{p} - \frac{n-\Sigma x_i}{1-p} = 0$$

Derivaatan ainoa *nollakohta*

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

antaa logaritmisen uskottavuusfunktion *maksimin*.

■

Parametrin p SU-estimaattori

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

on kiinnostuksen kohteena olevan tapahtuman A *suhteellinen frekvenssi* otoksessa, koska

$$\sum_{i=1}^{n} X_i = f$$

on tapahtuman A *frekvenssi* otoksessa, sillä summa $\sum X_i$ koostuu ykkösistä ja nollista ja ykkösten lukumäärä summassa on sama kuin tapahtuman A esiintymisten lukumäärä.

Huomaa, että parametrin p *suurimman uskottavuuden estimaattori yhtyy sen momentti-estimaattoriin*; lisätietoja momenttimenetelmästä: ks. kappaletta **Momenttimenetelmä**.

SU-estimaattorin ominaisuudet

Bernoulli-jakauman Ber(p) odotusarvoparametrin p SU-estimaattorilla $\hat{p}$ on seuraavat ominaisuudet:

(i) $\hat{p}$ on *harhaton*.

(ii) $\hat{p}$ on *tyhjentävä*.

(iii) $\hat{p}$ on (asymptoottisesti) *tehokas* eli minimivarianssinen estimaattori.

(iv) $\hat{p}$ on *tarkentuva*.

(v) $\hat{p}$ noudattaa *asymptoottisesti normaalijakaumaa*:

$$\hat{p} \underset{a}{\sim} \mathrm{N}\left(p, \frac{pq}{n}\right)$$

6.6. Momenttimenetelmä

Satunnaisotos

Olkoon

$$X_i\,,\ i = 1, 2, \ldots , n$$

satunnaisotos jakaumasta, jonka *pistetodennäköisyys*- tai *tiheysfunktio* $f(x;\theta)$, jonka *parametrina* on p-vektori

$$\theta = (\theta_1, \theta_2, \ldots , \theta_p)$$

Tällöin havainnot X_i , $i = 1, 2, \ldots , n$ ovat *riippumattomia, identtisesti jakautuneita satunnaismuuttujia*, joilla on *sama pistetodennäköisyys*- tai *tiheysfunktio* $f(x;\theta)$:

$$\begin{aligned} &X_1, X_2, \ldots, X_n \perp \\ &X_i \sim f(x;\theta)\,,\ i = 1, 2, \ldots, n \end{aligned}$$

Momentit

Oletetaan, että jakaumalla $f(x;\theta)$ on kaikki (*origo*-) *momentit* kertalukuun p saakka:

$$\mathrm{E}(X_i^k) = \alpha_k\,,\ k = 1, 2, \ldots, p\,,\ i = 1, 2, \ldots, n$$

Oletetaan, että momenttien

$$\alpha_1, \alpha_2, \ldots , \alpha_p$$

ja parametrien

$$\theta_1, \theta_2, \ldots , \theta_p$$

välillä on jatkuva *bijektio* eli kääntäen yksikäsitteinen kuvaus:

$$(1) \qquad \begin{cases} \alpha_1 = g_1(\theta_1, \theta_2, \ldots, \theta_p) \\ \alpha_2 = g_2(\theta_1, \theta_2, \ldots, \theta_p) \\ \vdots \\ \alpha_p = g_p(\theta_1, \theta_2, \ldots, \theta_p) \end{cases}$$

Tällöin parametrit

$$\theta_1, \theta_2, \ldots , \theta_p$$

voidaan esittää momenttien

$$\alpha_1, \alpha_2, \ldots , \alpha_p$$

funktioina:

(2) $$\begin{cases} \theta_1 = h_1(\alpha_1, \alpha_2, \ldots, \alpha_p) \\ \theta_2 = h_2(\alpha_1, \alpha_2, \ldots, \alpha_p) \\ \vdots \\ \theta_p = h_p(\alpha_1, \alpha_2, \ldots, \alpha_p) \end{cases}$$

Momenttiestimaattoreiden määrääminen

Estimoidaan momentit α_1, α_2 , … , α_p vastaavilla *otosmomenteilla*:

$$a_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \ , \ k = 1, 2, \ldots, p$$

Sijoittamalla estimaattorit a_1, a_2 , … , a_p momenttien α_1, α_2 , … , α_p paikalle yo. yhtälöihin (2), saadaan parametrien θ_1, θ_2, … , θ_p **momenttiestimaattorit** eli **MM-estimaattorit**

$$\begin{cases} \hat{\theta}_1 = h_1(a_1, a_2, \ldots, a_p) \\ \hat{\theta}_2 = h_2(a_1, a_2, \ldots, a_p) \\ \vdots \\ \hat{\theta}_p = h_p(a_1, a_2, \ldots, a_p) \end{cases}$$

Määräämme alla seuraavien jakaumien parametrien momenttiestimaattorit:

- **Normaalijakauma**
- **Eksponenttijakauma**
- **Bernoulli-jakauma**

Momenttimenetelmä vs suurimman uskottavuuden menetelmä

Momenttimenetelmä ja suurimman uskottavuuden menetelmä tuottavat monissa tapauksissa *samat estimaattorit* todennäköisyysjakauman parametreille. *Tämä ei ole kuitenkaan yleisesti totta*.

Momenttimenetelmä on näistä kahdesta estimointimenetelmästä *vanhempi*. *Suurimman uskottavuuden menetelmällä katsotaan hyvin yleisesti olevan paremmat teoreettiset perustelut kuin momenttimenetelmällä* ja siksi *suurimman uskottavuuden menetelmä onkin hyvin pitkälti syrjäyttänyt momenttimenetelmän* todennäköisyysjakaumien parametrien estimaattoreita johdettaessa.

6.7. Normaalijakauman parametrien momenttiestimointi

Satunnaismuuttuja X noudattaa **normaalijakaumaa** N(μ, σ^2), jos sen *tiheysfunktio* on muotoa

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, -\infty < \mu < +\infty \, , \sigma > 0$$

Normaalijakauman *parametreina* ovat jakauman *odotusarvo*

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

ja *varianssi*

$$\mathrm{Var}(X) = \mathrm{D}^2(X) = \sigma^2$$

Lisätietoja normaalijakaumasta: Ks. monisteen **Todennäköisyyslaskenta** lukua **Jatkuvia jakaumia**.

MM-estimaattoreiden johto

Määritellään satunnaismuuttujan X 1. ja 2. *momentti* kaavoilla

$$\alpha_k = \mathrm{E}(X^k)\,,\; k = 1, 2$$

Normaalijakauman parametrien μ ja σ^2 sekä momenttien α_1 ja α_2 välillä on seuraava *bijektio*:

(i) Parametrit lausuttuina momenttien funktioina:

$$\begin{cases} \mu = \mathrm{E}(X) = \alpha_1 \\ \sigma^2 = \mathrm{Var}(X) = \mathrm{E}[(X-\mu)^2] = \mathrm{E}(X^2) - \mu^2 = \alpha_2 - \alpha_1^2 \end{cases}$$

(ii) Momentit lausuttuina parametrien funktioina:

$$\begin{cases} \alpha_1 = \mathrm{E}(X) = \mu \\ \alpha_2 = \mathrm{E}(X^2) = \sigma^2 + \mu^2 \end{cases}$$

Olkoon

$$X_1, X_2, \dots, X_n$$

satunnaisotos normaalijakaumasta $\mathrm{N}(\mu, \sigma^2)$. Tällöin havainnot $X_1, X_2, \dots, X_n$ ovat *riippumattomia, samaa normaalijakaumaa* $\mathrm{N}(\mu, \sigma^2)$ noudattavia satunnaismuuttujia.

Määritellään havaintojen $X_1, X_2, \dots, X_n$ 1. ja 2. *otosmomentti* kaavoilla

$$a_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k\,,\; k = 1, 2$$

Siten normaalijakauman $\mathrm{N}(\mu, \sigma^2)$ parametrien μ ja σ^2 **MM-estimaattorit** eli **momentti-estimaattorit** ovat

$$\begin{cases} \hat{\mu} = a_1 = \dfrac{1}{n}\sum_{i=1}^{n} X_i \\ \hat{\sigma}^2 = a_2 - a_1^2 = \dfrac{1}{n}\sum_{i=1}^{n} X_i^2 - a_1^2 \end{cases}$$

Siten odotusarvoparametrin μ MM-estimaattori

$$\hat{\mu} = a_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

on havaintojen $X_1, X_2, \dots, X_n$ *aritmeettinen keskiarvo* ja varianssiparametrin σ^2 MM-estimaattori

$$\hat{\sigma}^2 = a_2 - a_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = m_2$$

on havaintojen $X_1, X_2, \dots, X_n$ *otosvarianssi* laskettuna kaavalla, jossa jakajana on havaintojen lukumäärä n. Huomaa, että $\hat{\sigma}^2$ on sama kuin havaintojen 2. *keskusmomentti* m_2.

Normaalijakauman parametrien μ ja σ^2 *momenttiestimaattorit yhtyvät niiden suurimman uskottavuuden estimaattoreihin*; lisätietoja suurimman uskottavuuden menetelmästä: ks. kappaletta **Suurimman uskottavuuden menetelmä**.

6.8. Eksponenttijakauman parametrien momenttiestimointi

Satunnaismuuttuja X noudattaa **eksponenttijakaumaa** Exp(λ), jos sen *tiheysfunktio* on

$$f(x) = \lambda e^{-\lambda x}\ ,\ x \geq 0\ ,\ \lambda > 0$$

Eksponenttijakauman ainoana *parametrina* on

$$\lambda = \frac{1}{\mathrm{E}(X)}$$

Lisätietoja eksponenttijakaumasta: Ks. monisteen **Todennäköisyyslaskenta** lukua **Jatkuvia jakaumia**.

MM-estimaattorin johto

Määritellään satunnaismuuttujan X 1. *momentti* kaavalla

$$\alpha_1 = \mathrm{E}(X)$$

Eksponenttijakauman parametrin λ ja 1. momentin α_1 välillä on seuraava *bijektio*:

(i) Parametri λ lausuttuna momentin α_1 funktiona:

$$\lambda = \frac{1}{\mathrm{E}(X)} = \frac{1}{\alpha_1}$$

(ii) Momentti α_1 lausuttuna parametrin λ funktiona:

$$\alpha_1 = \mathrm{E}(X) = \frac{1}{\lambda}$$

Olkoon

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

satunnaisotos eksponenttijakaumasta Exp(λ). Tällöin havainnot $X_1, X_2, \ldots, X_n$ ovat *riippumattomia, samaa eksponenttijakaumaa* Exp(λ) noudattavia satunnaismuuttujia.

Määritellään havaintojen $X_1, X_2, \ldots, X_n$ 1. *otosmomentti* kaavalla

$$a_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Siten eksponenttijakauman Exp(λ) parametrin λ **MM-estimaattori** eli **momenttiestimaattori** on

$$\hat{\lambda} = \frac{1}{\bar{X}}$$

jossa

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

on havaintojen $X_1, X_2, \ldots, X_n$ *aritmeettinen keskiarvo*.

Eksponenttijakauman parametrin λ *momenttiestimaattori yhtyy sen suurimman uskottavuuden estimaattoriin*; lisätietoja suurimman uskottavuuden menetelmästä: ks. kappaletta **Suurimman uskottavuuden menetelmä**.

6.9. Bernoulli-jakauman parametrien momenttiestimointi

Olkoon A *tapahtuma*, jonka todennäköisyys on p:

$$\Pr(A) = p$$

Määritellään satunnaismuuttuja X seuraavasti:

$$X = \begin{cases} 1, \text{ jos } A \text{ tapahtuu} \\ 0, \text{ jos } A \text{ ei tapahdu} \end{cases}$$

Tällöin satunnaismuuttuja X noudattaa **Bernoulli-jakaumaa** parametrilla p:

$$X \quad \mathrm{Ber}(p)$$

jossa

$$\Pr(A) = p = \mathrm{E}(X)$$

Satunnaismuuttujan X *pistetodennäköisyysfunktio* on

$$f(x) = p^x(1-p)^{1-x}\ ,\ x = 0,1$$

Lisätietoja Bernoulli-jakaumasta: Ks. monisteen **Todennäköisyyslaskenta** lukua **Diskreettejä jakaumia**.

MM-estimaattorin johto

Määritellään satunnaismuuttujan X 1. *momentti* kaavalla

$$\alpha_1 = \mathrm{E}(X)$$

Bernoulli-jakauman odotusarvoparametrin p ja 1. momentin α_1 välillä on seuraava *bijektio*:

(i) Parametri p lausuttuna momentin α_1 funktiona:

$$p = \mathrm{E}(X) = \alpha_1$$

(ii) Momentti α_1 lausuttuna parametrin p funktiona:

$$\alpha_1 = \mathrm{E}(X) = p$$

Olkoon

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

satunnaisotos Bernoulli-jakaumasta Ber(p). Tällöin havainnot $X_1, X_2, \ldots, X_n$ ovat *riippumattomia*, *samaa Bernoulli-jakaumaa* Ber(p) noudattavia satunnaismuuttujia.

Määritellään havaintojen $X_1, X_2, \ldots, X_n$ 1. *otosmomentti* kaavalla

$$a_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Siten Bernoulli-jakauman Ber(p) parametrin p **MM-estimaattori** eli **momenttiestimaattori** on

$$\hat{p} = a_1 = \bar{X}$$

jossa

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

on havaintojen $X_1, X_2, \ldots, X_n$ *aritmeettinen keskiarvo*. Parametrin p SU-estimaattori

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

on kiinnostuksen kohteena olevan tapahtuman A *suhteellinen frekvenssi* otoksessa, koska

$$\sum_{i=1}^{n} X_i = f$$

on tapahtuman A *frekvenssi* otoksessa, sillä summa $\sum X_i$ koostuu ykkösistä ja nollista ja ykkösten lukumäärä summassa on sama kuin tapahtuman A esiintymisten lukumäärä.

Bernoulli-jakauman parametrin p *momenttiestimaattori yhtyy sen suurimman uskottavuuden estimaattoriin*; lisätietoja suurimman uskottavuuden menetelmästä: ks. kappaletta **Suurimman uskottavuuden menetelmä**.

7. Väliestimointi

Tilastollinen aineisto koostuu tutkimuksen kohteita kuvaavien muuttujien **havaituista arvoista**.

Tilastollisen aineiston **tilastollisella mallilla** tarkoitetaan niiden satunnaismuuttujien *todennäköisyysjakaumaa, jonka ajatellaan generoineen havainnot*. Koska tämän jakauman **parametrit** ovat tavallisesti *tuntemattomia* ne pyritään on **estimoimaan** eli **arvioimaan** kerättyjen havaintojen perusteella.

Kutsumme parametrin tuntemattoman arvon estimoimiseen käytettävää havaintojen funktiota ko. parametrin **estimaattoriksi** ja sen havaintoarvoista laskettua arvoa ko. parametrin **estimaatiksi**.

Koska estimaattori on **otostunnusluku** ja siten sen saamat arvot *vaihtelevat satunnaisesti otoksesta toiseen*, on järkevää pyrkiä antamaan käsitys siitä, *mikä on parametrin todellinen, mutta tuntematon arvo*. Kutsumme **luottamusväliksi** väliä, joka sisältää parametrin todellisen, mutta tuntemattoman arvon etukäteen valitulla todennäköisyydellä, jota kutsumme **luottamustasoksi**.

Tässä luvussa tarkastelemme **luottamusvälejä** ja *niiden tulkintaa* sekä konstruoimme luottamusvälit **normaalijakauman odotusarvo-** ja **varianssiparametreille** sekä **Bernoulli-jakauman odotusarvoparametrille**.

Avainsanat:

Aritmeettinen keskiarvo, Bernoulli-jakauma, Estimaatti, Estimaattori, Estimointi, Frekvenssi, Frekvenssitulkinta, Harha, Harhattomuus, Havainto, Havaintoarvo, χ^2-jakauma, Keskeinen raja-arvolause, Keskineliövirhe, Luottamustaso, Luottamusväli, Normaalijakauma, Odotusarvo, Otos, Otosjakauma, Otoskoko, Otosvarianssi, Parametri, Piste-estimointi, Satunnaisotos, Suhteellinen frekvenssi, *t*-jakauma, Tarkentuvuus, Tilastollinen aineisto, Tilastollinen malli, Todennäköisyys, Todennäköisyysjakauma, Varianssi, Väliestimointi

7.1. Todennäköisyysjakauman parametrit ja niiden estimointi

Satunnaisotos

Olkoon

$$X_i\,,\ i = 1, 2, \ldots , n$$

satunnaisotos jakaumasta, jonka *pistetodennäköisyys*- tai *tiheysfunktio* $f(x;\theta)$ riippuu *parametrista* θ. Tällöin havainnot X_i , $i = 1, 2, \ldots , n$ ovat *riippumattomia, identtisesti jakautuneita satunnaismuuttujia*, joilla on *sama pistetodennäköisyys*- tai *tiheysfunktio* $f(x;\theta)$:

$$X_1, X_2, \ldots, X_n \perp$$
$$X_i \quad f(x;\theta)\,,\ i = 1, 2, \ldots, n$$

Estimaattori ja estimaatti

Oletetaan, että havainnot

$$X_i\,,\ i = 1, 2, \ldots , n$$

muodostavat *satunnaisotoksen* jakaumasta $f(x;\theta)$. Oletetaan, että todennäköisyysjakauman $f(x;\theta)$ *parametri* θ *on tuntematon* ja sen *estimoimiseen* käytetään havaintojen X_i , $i = 1, 2, \ldots , n$ (mitallista) funktiota eli (*otos*-) *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**.

Estimaattoreiden johtaminen

Hyvien estimaattoreiden johtaminen todennäköisyysjakaumien 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**. Tätä ongelmaa käsitellään luvussa **Estimointimenetelmät**.

Parametrin estimaattiin on syytä aina liittää **luottamusväliksi** kutsuttu *väli, joka sisältää estimoidun parametrin todellisen, mutta tuntemattoman arvon tietyllä, soveltajan valittavissa olevalla todennäköisyydellä*. Luottamusvälin määräämistä on tapana kutsua **väliestimoinniksi**.

7.2. Luottamusvälit

Väliestimoinnissa todennäköisyysjakauman $f(x;\theta)$ tuntemattomalle parametrille θ pyritään määräämään *havainnoista riippuva väli, joka tietyllä, tutkijan valittavissa olevalla toden näköisyydellä, peittää parametrin todellisen arvon*. Konstruoitua väliä kutsutaan **luottamusväliksi** ja valittua todennäköisyyttä kutsutaan **luottamustasoksi**.

Huomautus:

- Luottamustasolle voidaan antaa *frekvenssitulkinta* samassa hengessä kuin todennäköisyydelle.

Luottamusvälin määrääminen

Tehdään seuraavat oletukset:

(i) Olkoon

$$f(x;\theta)$$

satunnaismuuttujan X *todennäköisyysjakauma*, jonka määrää *tuntematon parametri* θ.

(ii) Olkoon

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

satunnaisotos jakaumasta $f(x;\theta)$.

(iii) Olkoon

$$\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$$

parametrin θ *estimaattori*.

Valitaan **luottamustaso**

$$1-\alpha$$

ja määrätään sen jälkeen *satunnaismuuttujat*

$$L = L(X_1, X_2, \ldots, X_n)$$
$$U = U(X_1, X_2, \ldots, X_n)$$

siten, että

$$\Pr(\hat{\theta} - L \le \theta) = \alpha/2$$
$$\Pr(\hat{\theta} + U \ge \theta) = \alpha/2$$

Huomautus:

- Satunnaismuuttujat L ja U riippuvat normaalisti sekä havainnoista $X_1, X_2, \ldots, X_n$ että valitusta luottamustasosta $(1-\alpha)$.

Tällöin väli

$$(\hat{\theta}-L,\hat{\theta}+U)$$

on **parametrin θ luottamusväli luottamustasolla (1 − α)**. Välin konstruktiosta seuraa, että väli peittää tuntemattoman parametrin θ todellisen arvon todennäköisyydellä (1 − α):

$$\Pr(\hat{\theta}-L\le\theta\le\hat{\theta}+U)=1-\alpha$$

Jos estimaattorin $\hat{\theta}$ jakauma on symmetrinen, parametrin θ luottamusväli luottamustasolla (1 − α) on muotoa

$$\hat{\theta}\pm A$$

jossa satunnaismuuttuja

$$A=A(X_1,X_2,\ldots,X_n)$$

valitaan siten, että

$$\Pr(\hat{\theta}-A\le\theta\le\hat{\theta}+A)=1-\alpha$$

Huomautus:

- Satunnaismuuttuja A riippuu normaalisti sekä havainnoista X_1 , X_2 , … , X_n että valitusta luottamustasosta (1 − α).

Luottamustason ja -välin frekvenssitulkinta

Oletetaan, että luottamustasoksi on valittu (1 − α). Luottamustasolle ja siihen liittyvälle luottamusvälille voidaan antaa seuraava frekvenssitulkinta:

(i) Jos otantaa jakaumasta $f(x;\theta)$ toistetaan, niin keskimäärin

$$100\times(1-\alpha)\ \%$$

otoksista konstruoiduista luottamusväleistä *peittää* parametrin θ todellisen arvon.

(ii) Jos otantaa jakaumasta $f(x;\theta)$ toistetaan, niin keskimäärin

$$100\times\alpha\ \%$$

otoksista konstruoiduista luottamusväleistä *ei peitä* parametrin θ todellista arvoa.

Johtopäätökset luottamusvälistä

Oletetaan, että olemme tehneet johtopäätöksen, että konstruoitu luottamusväli peittää parametrin θ tuntemattoman todellisen arvon:

(i) Luottamusvälin konstruktiosta seuraa, että tehty johtopäätös on *oikea* keskimäärin

$$100\times(1-\alpha)\ \%\text{:ssa}$$

tapauksia

(ii) Luottamusvälin konstruktiosta seuraa, että tehty johtopäätös on *väärä* keskimäärin

$$100\times\alpha\ \%\text{:ssa}$$

tapauksia.

Huomautus:

- *Virheellisen johtopäätöksen mahdollisuutta ei saada häviämään, ellei luottamusväliä tehdä äärettömän leveäksi*, jolloin väli *ei sisällä informaatiota* parametrin θ oikeasta arvosta.

Luottamusvälit: Esimerkkejä

Määräämme alla seuraavat luottausvälit:

- **Normaalijakauman odotusarvon luottamusväli**
- **Normaalijakauman varianssin luottamusväli**
- **Bernoulli-jakauman odotusarvoparametrin luottamusväli**

7.3. Normaalijakauman odotusarvon luottamusväli, kun jakauman varianssi on tunnettu

Otos normaalijakaumasta

Olkoon

$$X_i\,,\ i = 1, 2, \ldots , n$$

satunnaisotos normaalijakaumasta $N(\mu, \sigma^2)$, jossa jakauman varianssi σ^2 on *tunnettu*. Satunnaismuuttujat X_i , $i = 1, 2, \ldots , n$ ovat *riippumattomia* ja noudattavat *samaa normaalijakaumaa* $N(\mu, \sigma^2)$:

$$X_1, X_2, \ldots, X_n \perp$$
$$X_i \sim N(\mu, \sigma^2)\,,\ i = 1, 2, \ldots, n$$

Normaalijakauman parametrien estimointi

Koska olemme olettaneet, että normaalijakauman $N(\mu, \sigma^2)$ varianssi σ^2 on *tunnettu*, *estimoimme* vain jakauman odotusarvoparametrin $E(X) = \mu$ sen *harhattomalla estimaattorilla*

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

$\bar{X}$ on havaintojen X_i , $i = 1, 2, \ldots , n$ *aritmeettinen keskiarvo*.

Odotusarvon luottamusvälin konstruointi

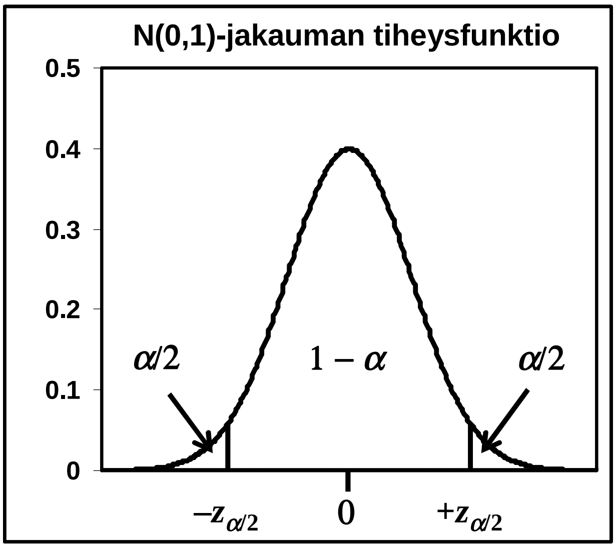
Valitaan *luottamustasoksi*

$$1 - \alpha$$

Luottamustaso kiinnittää todennäköisyyden, jolla konstruoitava luottamusväli peittää normaalijakauman odotusarvon μ todellisen arvon.

Määrätään *luottamuskertoimet* $-z_{\alpha/2}$ ja $+z_{\alpha/2}$ siten, että

$$\Pr(Z \le -z_{\alpha/2}) = \frac{\alpha}{2}$$



ja

$$\Pr(Z \geq +z_{\alpha/2}) = \frac{\alpha}{2}$$

jossa satunnaismuuttuja Z noudattaa *standardoitua normaalijakaumaa* N(0,1):

$$Z \quad \mathrm{N}(0,1)$$

Luottamuskertoimet $-z_{\alpha/2}$ ja $+z_{\alpha/2}$ toteuttavat ehdon

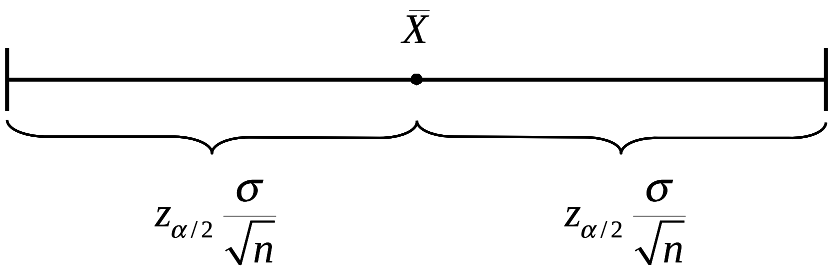
$$\Pr(-z_{\alpha/2} \leq Z \leq +z_{\alpha/2}) = 1-\alpha$$

Normaalijakauman **odotusarvoparametrin μ luottamusväli luottamustasolla (1 − α)** on *tunnetun varianssin σ^2 tapauksessa* muotoa

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \,,\, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right)$$

jossa

$\bar{X}$	=	havaintojen *aritmeettinen keskiarvo*
σ^2	=	jakauman *varianssi*
n	=	havaintojen *lukumäärä*
$-z_{\alpha/2}$ ja $+z_{\alpha/2}$	=	luottamustasoon (1 − α) liittyvät *luottamuskertoimet standardoisusta normaalijakaumasta* N(0,1)



Perustelu:

Olkoon

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

satunnaisotos normaalijakaumasta N(μ, σ^2), jossa jakauman varianssi σ^2 on tunnettu ja olkoon

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

havaintojen $X_1, X_2, \ldots , X_n$ *aritmeettinen keskiarvo*.

Määritellään satunnaismuuttuja

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Satunnaismuuttujan Z jakauma johdettiin luvun **Otokset ja otosjakaumat** kappaleessa **Aritmeettisen keskiarvon ja otosvarianssin otosjakaumat**. Tällöin todettiin, että aritmeettinen keskiarvo $\bar{X}$ noudattaa normaalijakautuneen otoksen tapauksessa *normaalijakaumaa* parametrein μ ja σ^2/n:

$$\bar{X} \quad \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Siten *standardoitu satunnaismuuttuja*

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

noudattaa *standardoitua normaalijakaumaa* N(0,1):

$$Z \quad \mathrm{N}(0,1)$$

Määrätään standardoidusta normaalijakaumasta piste + $z_{\alpha/2}$ siten, että

$$\Pr(Z \geq +z_{\alpha/2}) = \frac{\alpha}{2}$$

jolloin standardoidun normaalijakauman *symmetrian* perusteella

$$\Pr(Z \leq -z_{\alpha/2}) = \frac{\alpha}{2}$$

ja edelleen

$$\Pr(-z_{\alpha/2} \leq Z \leq +z_{\alpha/2}) = 1 - \alpha$$

Tarkastellaan epäyhtälöketjua

$$-z_{\alpha/2} \leq Z \leq +z_{\alpha/2}$$

Sijoittamalla tähän epäyhtälöketjuun satunnaismuuttujan Z lauseke, saadaan epäyhtälöketju

$$-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq +z_{\alpha/2}$$

Tästä epäyhtälöketjusta saadaan sen kanssa *yhtäpitävä* epäyhtälöketju

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Yhdistämällä saatu epäyhtälö siihen, että

$$\Pr(-z_{\alpha/2} \leq Z \leq +z_{\alpha/2}) = 1 - \alpha$$

saadaan vihdoin

$$\Pr\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

■

Koska normaalijakauman odotusarvon luottamusväli

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\,,\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

on *symmetrinen* keskipisteensä $\bar{X}$ suhteen, luottamusväli esitetään usein muodossa

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Luottamusvälin *pituus* on

$$2 \times z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Luottamusvälin konstruktiosta seuraa, että

$$\Pr\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha$$

Siten luottamusväli *peittää* parametrin μ todellisen arvon todennäköisyydellä $(1 - \alpha)$ ja se *ei peitä* parametrin μ todellista arvoa todennäköisyydellä α.

Luottamusvälin ominaisuudet

(i) Normaalijakauman odotusarvon μ luottamusvälin *keskipiste* $\bar{X}$ vaihtelee otoksesta toiseen.

(ii) Luottamusvälin *pituus ei vaihtele* otoksesta toiseen.

(iii) Luottamusvälin *pituus* riippuu valitusta luottamustasosta $(1 - \alpha)$, havaintojen lukumäärästä n ja jakauman varianssista σ^2.

(iv) Luottamusväli *lyhenee* (*pitenee*), jos *luottamustasoa* $(1 - \alpha)$ *pienennetään* (*kasvatetaan*).

(v) Luottamusväli *lyhenee* (*pitenee*), jos *havaintojen lukumäärää n kasvatetaan* (*pienennetään*).

(vi) Luottamusväli *lyhenee* (*pitenee*), jos jakauman *varianssi* σ^2 *pienenee* (*kasvaa*).

Luottamusvälin frekvenssitulkinta

Normaalijakauman odotusarvon μ luottamusvälillä on seuraava *frekvenssitulkinta*:

(i) Jos otantaa jakaumasta $N(\mu, \sigma^2)$ toistetaan, niin keskimäärin

$$100 \times (1 - \alpha)\ \%$$

otoksista konstruoiduista luottamusväleistä *peittää* parametrin μ todellisen arvon.

(ii) Jos otantaa jakaumasta $N(\mu, \sigma^2)$ toistetaan, niin keskimäärin

$$100 \times \alpha\ \%$$

otoksista konstruoiduista luottamusväleistä *ei peitä* parametrin μ todellista arvoa.

Johtopäätökset luottamusvälistä

Oletetaan, että olemme tehneet *johtopäätöksen*, että konstruoitu luottamusväli peittää odotusarvoparametrin μ todellisen arvon:

(i) Luottamusvälin konstruktiosta seuraa, että tehty johtopäätös on *oikea* keskimäärin

$$100 \times (1 - \alpha)\ \%\text{:ssa}$$

tapauksia.

(ii) Luottamusvälin konstruktiosta seuraa, että tehty johtopäätös on *väärä* keskimäärin

$$100\times\alpha\,\%\text{:ssa}$$

tapauksia.

Virheellisen johtopäätöksen mahdollisuutta ei saada häviämään, ellei luottamusväliä tehdä äärettömän leveäksi, jolloin väli *ei enää sisällä informaatiota* odotusarvoparametrin μ todellisesta arvosta.

Vaatimukset luottamusvälille

Olisi toivottavaa pystyä konstruoimaan parametrille μ mahdollisimman *lyhyt* luottamusväli, johon liittyvä luottamustaso olisi samanaikaisesti mahdollisimman *korkea*.

Molempien vaatimusten samanaikainen täyttäminen *ei ole* kuitenkaan mahdollista, *jos otoskoko pidetään kiinteänä*:

(i) *Luottamustason kasvattaminen pidentää luottamusväliä*, jolloin tieto parametrin μ todellisen arvon sijainnista tulee *epätarkemmaksi*.

(ii) *Luottamusvälin lyhentäminen pienentää luottamustasoa*, jolloin tieto parametrin μ todellisen arvon sijainnista tulee *epävarmemmaksi*.

Otoskoon määrääminen

Oletetaan, että normaalijakauman odotusarvoparametrille μ halutaan konstruoida luottamusväli, jonka *toivottu pituus* on $2A$. Approksimatiivinen *otoskoko* saadaan kaavasta

$$n=\left(\frac{z_{\alpha/2}\sigma}{A}\right)^2$$

7.4. Normaalijakauman odotusarvon luottamusväli, kun jakauman varianssi on tuntematon

Otos normaalijakaumasta

Olkoon

$$X_i\,,\ i = 1, 2, \ldots , n$$

satunnaisotos normaalijakaumasta $\mathrm{N}(\mu, \sigma^2)$, jossa molemmat parametrit μ ja σ^2 ovat *tuntemattomia*. Satunnaismuuttujat X_i , $i = 1, 2, \ldots , n$ ovat *riippumattomia* ja noudattavat *samaa normaalijakaumaa* $\mathrm{N}(\mu, \sigma^2)$:

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

$$X_i \sim \mathrm{N}(\mu,\sigma^2)\,,\ i=1,2,\ldots,n$$

Normaalijakauman parametrien estimointi

Estimoidaan normaalijakauman $\mathrm{N}(\mu, \sigma^2)$ parametrit μ ja σ^2 niiden *harhattomilla estimaattoreilla*:

Odotusarvoparametrin $\mathrm{E}(X) = \mu$ harhaton estimaattori on havaintojen *aritmeettinen keskiarvo*

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

ja *varianssiparametrin* Var(X) = σ^2 harhaton estimaattori on havaintojen *otosvarianssi*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Odotusarvon luottamusvälin konstruointi

Valitaan *luottamustasoksi*

$$1-\alpha$$

Luottamustaso kiinnittää todennäköisyyden, jolla konstruoitava luottamusväli peittää normaalijakauman odotusarvon μ todellisen arvon.

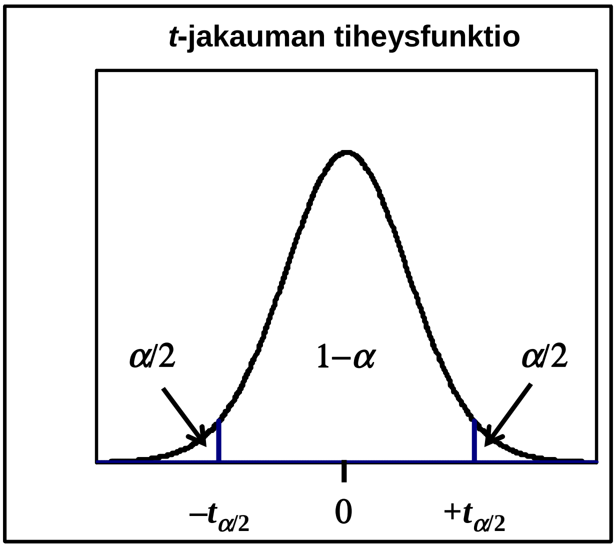
Määrätään *luottamuskertoimet* $-t_{\alpha/2}$ ja $+t_{\alpha/2}$ siten, että

$$\Pr(t \le -t_{\alpha/2}) = \frac{\alpha}{2}$$

$$\Pr(t \ge +t_{\alpha/2}) = \frac{\alpha}{2}$$

jossa satunnaismuuttuja t noudattaa *t-jakaumaa* vapausastein (n – 1):

$$t \quad t(n-1)$$



Siten luottamuskertoimet $-t_{\alpha/2}$ ja $+t_{\alpha/2}$ toteuttavat ehdon

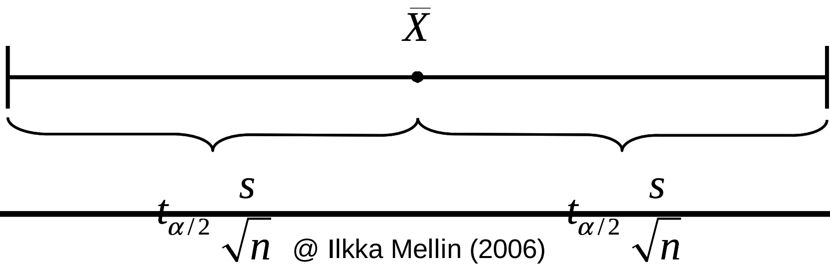
$$\Pr(-t_{\alpha/2} \le t \le +t_{\alpha/2}) = 1-\alpha$$

Normaalijakauman **odotusarvoparametrin μ luottamusväli luottamustasolla (1 – α)** on *tuntemattoman varianssin σ^2 tapauksessa* muotoa

$$\left(\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}}\,,\ \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$

jossa

$\bar{X}$	=	havaintojen *aritmeettinen keskiarvo*
s^2	=	*otosvarianssi*
n	=	*havaintojen lukumäärä*
$-t_{\alpha/2}$ ja $+t_{\alpha/2}$	=	luottamustasoon (1 – α) liittyvät *luottamuskertoimet* *t-jakaumasta* $t(n-1)$



Perustelu:

Olkoon

$$X_1, X_2, \dots , X_n$$

satunnaisotos normaalijakaumasta N(μ, σ^2) ja olkoon

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

havaintojen $X_1, X_2, \dots , X_n$ *aritmeettinen keskiarvo* ja

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

havaintojen $X_1, X_2, \dots , X_n$ (harhaton) *otosvarianssi*.

Määritellään satunnaismuuttuja

$$t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$$

Satunnaismuuttuja t voidaan kirjoittaa muotoon

$$t = \frac{\bar{X}-\mu}{s/\sqrt{n}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \cdot \frac{\sigma}{\sqrt{s^2}} = \frac{\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)s^2/\sigma^2}{n-1}}}$$

Satunnaismuuttujan t jakauma johdettiin luvun **Otokset ja otosjakaumat** kappaleessa **Aritmeettisen keskiarvon ja otosvarianssin otosjakaumat**. Tällöin todettiin että satunnaismuuttujan t osoittaja määrittelee satunnaismuuttujan

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

joka noudattaa *standardoitua normaalijakaumaa* N(0,1):

$$Z \quad \mathrm{N}(0,1)$$

Edelleen totesimme, että satunnaismuuttuja t nimittäjä määrittelee satunnaismuuttujan

$$V = (n-1)\frac{s^2}{\sigma^2}$$

joka noudattaa χ^2-*jakaumaa vapausastein* (n – 1):

$$V \quad \chi^2(n-1)$$

Lisäksi todettiin, että satunnaismuuttujat Z ja V ovat *riippumattomia*. Siten satunnaismuuttuja

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{Z}{\sqrt{\dfrac{V}{n-1}}}$$

noudattaa *t-jakaumaa* vapausastein (n – 1) suoraan *t*-jakauman *määritelmän mukaan*:

$$t \quad t(n-1)$$

Määrätään *t*-jakaumasta vapausastein (n – 1) piste + $t_{\alpha/2}$ siten, että

$$\Pr(t \geq +t_{\alpha/2}) = \frac{\alpha}{2}$$

jolloin *t*-jakauman *symmetrian* perusteella

$$\Pr(t \leq -t_{\alpha/2}) = \frac{\alpha}{2}$$

ja edelleen

$$\Pr(-t_{\alpha/2} \leq t \leq +t_{\alpha/2}) = 1 - \alpha$$

Tarkastellaan epäyhtälöketjua

$$-t_{\alpha/2} \leq t \leq +t_{\alpha/2}$$

Sijoittamalla tähän epäyhtälöketjuun satunnaismuuttujan t lauseke, saadaan epäyhtälöketju

$$-t_{\alpha/2} \leq \frac{\bar{X} - \mu}{s/\sqrt{n}} \leq +t_{\alpha/2}$$

Tästä epäyhtälöketjusta saadaan sen kanssa *yhtäpitävä* epäyhtälöketju

$$\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

Yhdistämällä saatu epäyhtälö siihen, että

$$\Pr(-t_{\alpha/2} \leq t \leq +t_{\alpha/2}) = 1 - \alpha$$

saadaan vihdoin

$$\Pr\left(\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

■

Koska luottamusväli

$$\left(\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}}\,,\ \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$

on *symmetrinen* keskipisteensä $\bar{X}$ suhteen, luottamusväli esitetään usein muodossa

$$\bar{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

Luottamusvälin *pituus* on

$$2 \times t_{\alpha/2} \frac{s}{\sqrt{n}}$$

Luottamusvälin konstruktiosta seuraa, että

$$\Pr\left(\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

Siten luottamusväli *peittää* parametrin μ todellisen arvon todennäköisyydellä $(1 - \alpha)$ ja se *ei peitä* parametrin μ todellista arvoa todennäköisyydellä α.

Luottamusvälin ominaisuudet

(i) Normaalijakauman odotusarvon μ luottamusvälin *keskipiste* $\bar{X}$ vaihtelee otoksesta toiseen.

(ii) Luottamusvälin *pituus vaihtelee* otoksesta toiseen.

(iii) Luottamusvälin *pituus* riippuu valitusta luottamustasosta $(1 - \alpha)$, havaintojen lukumäärästä n ja otosvarianssista s^2.

(iv) Luottamusväli *lyhenee* (*pitenee*), jos *luottamustasoa* $(1 - \alpha)$ *pienennetään* (*kasvatetaan*).

(v) Luottamusväli *lyhenee* (*pitenee*), jos *havaintojen lukumäärää n kasvatetaan* (*pienennetään*).

(vi) Luottamusväli *lyhenee* (*pitenee*), jos *otosvarianssi s^2 pienenee* (*kasvaa*).

Luottamusvälin frekvenssitulkinta

Normaalijakauman odotusarvon μ luottamusvälillä on seuraava *frekvenssitulkinta*:

(i) Jos otantaa jakaumasta $N(\mu, \sigma^2)$ toistetaan, niin keskimäärin

$$100 \times (1 - \alpha)\ \%$$

otoksista konstruoiduista luottamusväleistä *peittää* parametrin μ todellisen arvon.

(ii) Jos otantaa jakaumasta $N(\mu, \sigma^2)$ toistetaan, niin keskimäärin

$$100 \times \alpha\ \%$$

otoksista konstruoiduista luottamusväleistä *ei peitä* parametrin μ todellista arvoa.

Johtopäätökset luottamusvälistä

Oletetaan, että olemme tehneet *johtopäätöksen*, että konstruoitu luottamusväli peittää odotusarvoparametrin μ todellisen arvon:

(i) Luottamusvälin konstruktiosta seuraa, että tehty johtopäätös on *oikea* keskimäärin

$$100 \times (1 - \alpha)\ \%\text{:ssa}$$

tapauksia.

(ii) Luottamusvälin konstruktiosta seuraa, että tehty johtopäätös on *väärä* keskimäärin

$$100 \times \alpha\ \%\text{:ssa}$$

tapauksia.

Virheellisen johtopäätöksen mahdollisuutta ei saada häviämään, ellei luottamusväliä tehdä äärettömän leveäksi, jolloin väli *ei enää sisällä informaatiota* odotusarvoparametrin μ todellisesta arvosta.

Vaatimukset luottamusvälille

Olisi toivottavaa pystyä konstruoimaan odotusarvoparametrille μ mahdollisimman *lyhyt* luottamusväli, johon liittyvä luottamustaso olisi samanaikaisesti mahdollisimman *korkea*.

Molempien vaatimusten samanaikainen täyttäminen *ei ole* kuitenkaan mahdollista, *jos otoskoko pidetään kiinteänä*:

(i) *Luottamustason kasvattaminen pidentää luottamusväliä*, jolloin tieto parametrin μ todellisen arvon sijainnista tulee *epätarkemmaksi*.

(ii) *Luottamusvälin lyhentäminen pienentää luottamustasoa*, jolloin tieto parametrin μ todellisen arvon sijainnista tulee *epävarmemmaksi*.

Otoskoon määrääminen

Oletetaan, että normaalijakauman odotusarvoparametrille μ halutaan konstruoida luottamusväli, jonka *toivottu pituus* on $2A$. Approksimatiivinen *otoskoko* saadaan kaavasta

$$n = \left(\frac{z_{\alpha/2}\sigma}{A} \right)^2$$

jossa

$z_{\alpha/2}$ = luottamustasoon $(1 - \alpha)$ liittyvä *luottamuskerroin normaalijakaumasta*

Tämä merkitsee sitä, että otoskoon määräämiseksi käytettävissä on oltava edes karkea *arvio normaalijkauman varianssin* σ^2 *suuruudesta*.

Normaalijakauman odotusarvon luottamusvälin määrääminen: Esimerkki

Kone tekee ruuveja, joiden pituudet vaihtelevat satunnaisesti noudattaen normaalijakaumaa; ks. esimerkkiä luvussa **Tilastollisten aineistojen kuvaaminen**.

Ruuvien joukosta poimitaan *satunnaisotos*, jonka koko on

$$n = 30$$

ja otokseen poimittujen ruuvien pituudet mitataan.

Taulukko oikealla esittää pituuksien *luokiteltua frekvenssijakaumaa*.

Luokkavälit	Luokkafrekvenssit
(9.85,9.90]	1
(9.90,9.95]	2
(9.95,10.00]	6
(10.00,10.05]	3
(10.05,10.10]	5
(10.10,10.15]	4
(10.15,10.20]	5
(10.20,10.25]	3
(10.25,10.30]	1

Otosta kuvaavat seuraavat *tunnusluvut*:

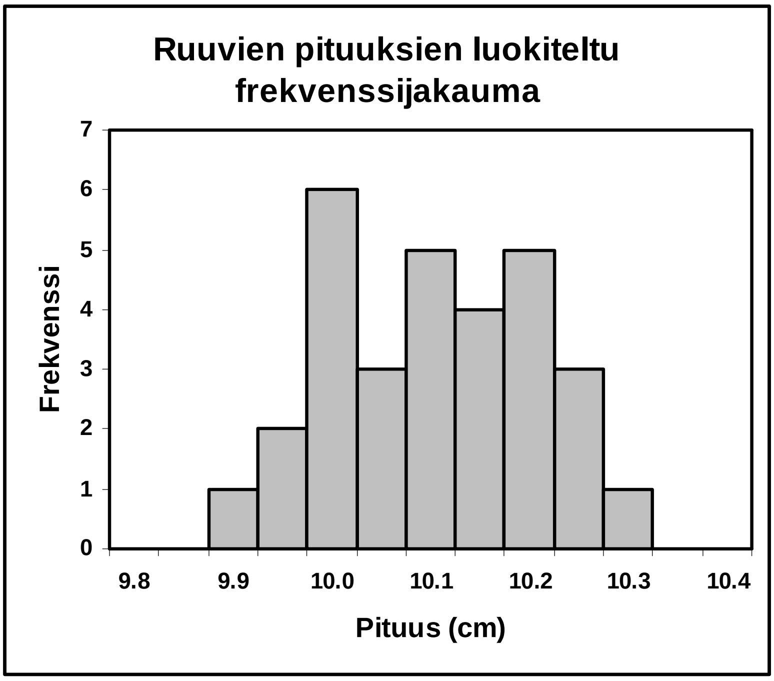
Pituuksien *aritmeettinen keskiarvo* on

$$\bar{X} = 10.09 \text{ cm}$$

ja pituuksien *otoskeskihajonta* on

$$s = 0.1038 \text{ cm}$$

Kuva alla esittää otokseen poimittujen ruuvien pituuksien luokiteltua frekvenssijakaumaa vastaavaa *histogrammia*. Luokkavälit määräävät kuvion suorakaiteiden kannat ja suorakaiteiden korkeudet on valittu niin, että suorakaiteiden pinta-alat suhtautuvat toisiinsa kuten vastaavat luokkafrekvenssit.



Huomautus:

- Jos otantaa toistetaan, kaikki otosta koskevat tiedot (sekä havaintoarvot että niistä määrätyt otossuureet kuten aritmeettiset keskiarvot ja otoskeskihajonnat sekä havainto-arvojen jakaumaa kuvaavat graafiset esitykset kuten hisrogrammit) vaihtelevat satunnaisesti otoksesta toiseen.

Ongelma: Mitä koneen tekemien ruuvien todellisesta keskipituudesta voidaan tietää yhdestä otoksesta saatujen tietojen perusteella?

Ratkaisu: Konstruoidaan ruuvien todelliselle keskipituudelle **luottamusväli**. Väli sisältää todellisen keskipituuden valitulla todennäköisyydellä.

Olkoot siis otokseen poimittujen ruuvien pituudet

$$X_1, X_2, \ldots, X_{30}$$

Oletetaan, että havainnot $X_1, X_2, \ldots, X_{30}$ ovat riippumattomia ja noudattavat samaa normaali-jakaumaa $\mathrm{N}(\mu, \sigma^2)$ ja olkoon

$$\bar{X} = \frac{1}{30}\sum_{i=1}^{30} X_i = 10.09$$

havaintojen $X_1, X_2, \ldots, X_{30}$ *aritmeettinen keskiarvo* ja

$$s^2 = \frac{1}{30-1}\sum_{i=1}^{30}(X_i - \bar{X})^2 = 0.1038^2$$

havaintojen $X_1, X_2, \ldots, X_{30}$ *otosvarianssi*.

Valitaan *luottamustasoksi*

$$1 - \alpha = 0.95$$

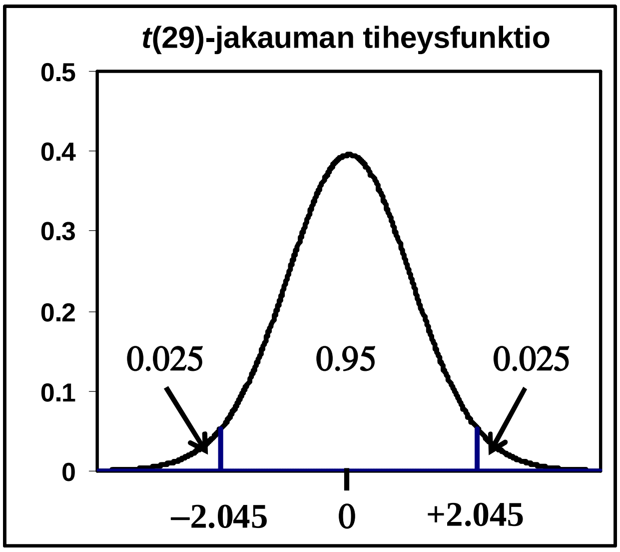
jolloin

$$\alpha/2 = 0.025$$

Luottamuskertoimet $-t_{0.025}$ ja $+ t_{0.025}$ on siis valittava siten, että

$$\Pr(t \le -t_{0.025}) = \Pr(t \ge +t_{0.025}) = 0.025$$

jossa satunnaismuuttuja t noudattaa *t-jakaumaa* vapausastein $n - 1 = 29$.



Luottamuskertoimet $-t_{0.025}$ ja $+ t_{0.025}$ toteuttavat ehdon

$$\Pr(-t_{0.025} \le t \le +t_{0.025}) = 0.95$$

t-jakauman taulukoista nähdään, että

$$\Pr(t \ge +2.045) = 0.025$$

t-jakauman *symmetrian* takia

$$\Pr(t \le -2.045) = 0.025$$

Siten luottamustasoa 0.95 vastaavat luottamuskertoimet ovat:

$$-\, t_{0.025} = -2.045$$

$$+\, t_{0.025} = +2.045$$

Kuvio oikealla havainnollistaa luottamuskertoimien valintaa.

Siten ruuvien todellisen keskipituuden μ *luottamusväliksi* saadaan:

$$\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} = 10.09 \pm 2.045 \times \frac{0.1038}{\sqrt{30}} = 10.09 \pm 0.04 = (10.05, 10.13)$$

Siten tiedämme, että ruuvien todellinen keskipituus on todennäköisyydellä 0.95 välillä

(10.05, 10.13)

Tarkastellaan vielä luottamustason 0.95 ja sitä vastaavan luottamusvälin *frekvenssitulkintaa*:

Oletetaan, että poimimme koneen tekemien ruuvien joukosta toistuvasti satunnaisotoksia, joiden koko on 30 ja konstruoimme jokaisesta otoksesta 95 %:n luottamusvälin edellä esitetyllä tavalla. Tällöin pätee seuraava:

(i) Konstruoidusta väleistä keskimäärin 95 % *peittää* ruuvien todellisen, mutta tuntemattoman keskipituuden.

(ii) Konstruoidusta väleistä keskimäärin 5 % *ei peitä* ruuvien todellista, mutta tuntematonta keskipituutta.

7.5. Normaalijakauman varianssin luottamusväli

Otos normaalijakaumasta

Olkoon

$$X_i\,,\ i = 1, 2, \ldots , n$$

satunnaisotos normaalijakaumasta N(μ, σ^2), jossa molemmat parametrit μ ja σ^2 ovat *tuntemattomia*. Satunnaismuuttujat X_i , i = 1, 2, … , n ovat *riippumattomia* ja noudattavat *samaa normaalijakaumaa* N(μ, σ^2):

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

$$X_i \sim \mathrm{N}(\mu, \sigma^2)\,,\ i = 1, 2, \ldots , n$$

Normaalijakauman parametrien estimointi

Estimoidaan normaalijakauman N(μ, σ^2) parametrit μ ja σ^2 niiden *harhattomilla estimaattoreilla*:

Odotusarvoparametrin E(X) = μ harhaton estimaattori on havaintojen *aritmeettinen keskiarvo*

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

ja *varianssiparametrin* Var(X) = σ^2 harhaton estimaattori on havaintojen *otosvarianssi*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Varianssin luottamusvälin konstruointi

Valitaan *luottamustasoksi*

$$1 - \alpha$$

Luottamustaso kiinnittää todennäköisyyden, jolla konstruoitava luottamusväli peittää normaalijakauman varianssin σ^2 todellisen arvon.

Määrätään *luottamuskertoimet* $\chi^2_{1-\alpha/2}$ ja $\chi^2_{\alpha/2}$ siten, että

$$\Pr(\chi^2 \le \chi^2_{1-\alpha/2}) = \frac{\alpha}{2}$$

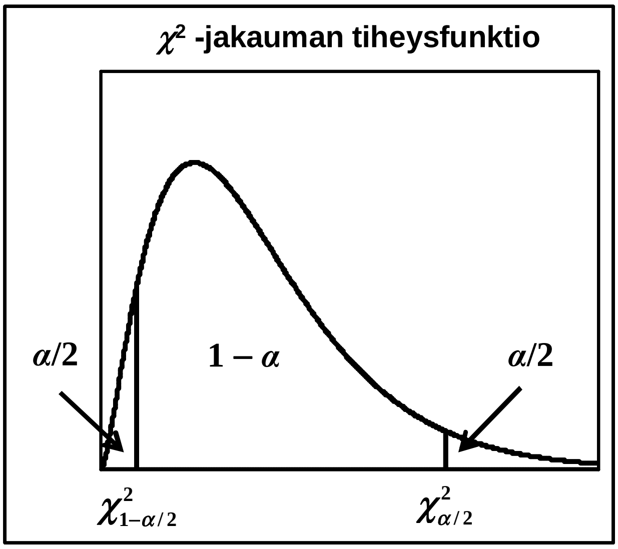
$$\Pr(\chi^2 \ge \chi^2_{\alpha/2}) = \frac{\alpha}{2}$$

jossa satunnaismuuttuja χ^2 noudattaa χ^2-jakaumaa vapausastein (n – 1):

$$\chi^2 \sim \chi^2(n-1)$$

Siten luottamuskertoimet $\chi^2_{1-\alpha/2}$ ja $\chi^2_{\alpha/2}$ toteuttavat ehdon

$$\Pr(\chi^2_{1-\alpha/2} \le \chi^2 \le \chi^2_{\alpha/2}) = 1-\alpha$$



Normaalijakauman **varianssiparametrin σ^2 luottamusväli luottamustasolla (1 − α)** on muotoa

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \right)$$

jossa

s^2	=	*otosvarianssi*
n	=	*havaintojen lukumäärä*
$\chi^2_{1-\alpha/2}$ ja $\chi^2_{\alpha/2}$	=	luottamustasoon (1 – α) liittyvät *luottamuskertoimet* χ^2*-jakaumasta* vapausastein (n – 1)

Perustelu:

Olkoon

$$X_1, X_2, \dots, X_n$$

satunnaisotos normaalijakaumasta N(μ, σ^2) ja olkoon

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

havaintojen $X_1, X_2, \dots, X_n$ *aritmeettinen keskiarvo* ja

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

havaintojen $X_1, X_2, \dots, X_n$ (harhaton) *otosvarianssi*.

Määritellään satunnaismuuttuja

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Satunnaismuuttujan χ^2 jakauma johdettiin luvun **Otokset ja otosjakaumat** kappaleessa **Aritmeettisen keskiarvon ja otosvarianssin otosjakaumat**. Tällöin todettiin, että satunnaismuuttuja χ^2 noudattaa χ^2*-jakaumaa* vapausastein (n – 1):

$$\chi^2 \quad \chi^2(n-1)$$

Määrätään χ^2-jakaumasta vapausastein (n – 1) piste $\chi^2_{1-\alpha/2}$ siten, että

$$\Pr(\chi^2 \le \chi^2_{1-\alpha/2}) = \frac{\alpha}{2}$$

ja piste $\chi^2_{\alpha/2}$ siten, että

$$\Pr(\chi^2 \ge \chi^2_{\alpha/2}) = \frac{\alpha}{2}$$

jolloin

$$\Pr(\chi^2_{1-\alpha/2} \le \chi^2 \le \chi^2_{\alpha/2}) = 1-\alpha$$

Tarkastellaan epäyhtälöketjua

$$\chi^2_{1-\alpha/2} \le \chi^2 \le \chi^2_{\alpha/2}$$

Sijoittamalla tähän epäyhtälöketjuun satunnaismuuttujan χ^2 lauseke, saadaan epäyhtälöketju

$$\chi^2_{1-\alpha/2} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{\alpha/2}$$

Tästä epäyhtälöketjusta saadaan sen kanssa *yhtäpitävä* epäyhtälöketju

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

Yhdistämällä saatu epäyhtälö siihen, että

$$\Pr(\chi^2_{1-\alpha/2} \le \chi^2 \le \chi^2_{\alpha/2}) = 1-\alpha$$

saadaan vihdoin

$$\Pr\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha$$

■

Luottamusvälin pituus on

$$(n-1)s^2\left(\frac{1}{\chi^2_{1-\alpha/2}} - \frac{1}{\chi^2_{\alpha/2}}\right)$$

Luottamusvälin konstruktiosta seuraa, että

$$\Pr\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha$$

Siten luottamusväli *peittää* parametrin σ^2 todellisen arvon todennäköisyydellä $(1-\alpha)$ ja se *ei peitä* parametrin σ^2 todellista arvoa todennäköisyydellä α.

Luottamusvälin ominaisuudet

(i) Normaalijakauman varianssin σ^2 luottamusvälin *pituus* vaihtelee otoksesta toiseen.

(ii) Luottamusvälin *pituus* riippuu valitusta luottamustasosta $(1-\alpha)$, havaintojen lukumäärästä n ja otosvarianssista s^2 .

(iii) Luottamusväli *lyhenee* (*pitenee*), jos *luottamustasoa* $(1-\alpha)$ *pienennetään* (*kasvatetaan*).

(iv) Luottamusväli *lyhenee* (*pitenee*), jos *havaintojen lukumäärää n pienennetään* (*kasvatetaan*).

(v) Luottamusväli *lyhenee* (*pitenee*), jos *otosvarianssi* s^2 *pienenee* (*kasvaa*).

Luottamusvälin frekvenssitulkinta

Normaalijakauman odotusarvon σ^2 luottamusvälillä on seuraava *frekvenssitulkinta*:

(i) Jos otantaa jakaumasta $N(\mu,\sigma^2)$ toistetaan, niin keskimäärin

$100\times(1-\alpha)$ %

otoksista konstruoiduista luottamusväleistä *peittää* parametrin σ^2 todellisen arvon.

(ii) Jos otantaa jakaumasta $N(\mu,\sigma^2)$ toistetaan, niin keskimäärin

$$100\times\alpha\,\%$$

otoksista konstruoiduista luottamusväleistä *ei peitä* parametrin σ^2 todellista arvoa.

Johtopäätökset luottamusvälistä

Oletetaan, että olemme tehneet *johtopäätöksen*, että konstruoitu luottamusväli peittää varianssiparametrin σ^2 todellisen arvon:

(i) Luottamusvälin konstruktiosta seuraa, että tehty johtopäätös on *oikea* keskimäärin

$$100\times(1-\alpha)\ \%\text{:ssa}$$

tapauksia.

(ii) Luottamusvälin konstruktiosta seuraa, että tehty johtopäätös on *väärä* keskimäärin

$$100\times\alpha\,\%\text{:ssa}$$

tapauksia.

Vaatimukset luottamusvälille

Olisi toivottavaa pystyä konstruoimaan varianssiparametrille σ^2 mahdollisimman *lyhyt* luottamus-väli, johon liittyvä luottamustaso olisi samanaikaisesti mahdollisimman *korkea*.

Vaatimusten samanaikainen täyttäminen *ei ole* kuitenkaan mahdollista:

(i) *Luottamustason kasvattaminen pidentää luottamusväliä*, jolloin tieto parametrin σ^2 todellisen arvon sijainnista tulee *epätarkemmaksi*.

(ii) *Luottamusvälin lyhentäminen pienentää luottamustasoa*, jolloin tieto parametrin σ^2 todellisen arvon sijainnista tulee *epävarmemmaksi*.

7.6. Bernoulli-jakauman odotusarvon luottamusväli

Bernoulli-jakauma

Olkoon A on jokin *tapahtuma* ja olkoon

$$\Pr(A) = p$$
$$\Pr(A^c) = 1 - p = q$$

Määritellään satunnaismuuttuja

$$X = \begin{cases} 1, \text{ jos } A \text{ tapahtuu} \\ 0, \text{ jos } A \text{ ei tapahdu} \end{cases}$$

Tällöin satunnaismuuttuja X noudattaa **Bernoulli-jakaumaa** parametrinaan

$$p = \Pr(A) = \mathrm{E}(X)$$

Merkitään:

$$X \quad \mathrm{Ber}(p)$$

Bernoulli-jakauman *pistetodennäköisyysfunktio* on

$$f(x; p) = p^x (1-p)^{1-x}\ ,\ x = 0, 1\ ;\ 0 < p < 1$$

Otos Bernoulli-jakaumasta

Olkoon

$$X_i\ ,\ i = 1, 2, \ldots, n$$

satunnaisotos Bernoulli-jakaumasta Ber(p). Tällöin satunnaismuuttujat X_i , $i = 1, 2, \ldots, n$ ovat *riippumattomia* ja noudattavat *samaa Bernoulli-jakaumaa* Ber(p):

$$X_1, X_2, \ldots, X_n \perp$$
$$X_i \quad \mathrm{Ber}(p)\ ,\ i = 1, 2, \ldots, n$$

Bernoulli-jakauman odotusarvoparametrin estimointi

Estimoidaan Bernoulli-jakauman Ber(p) odotusarvoparametri p sen *harhattomalla estimaattorilla*:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Koska

$$X_i = \begin{cases} 1, \text{ jos } A \text{ tapahtuu} \\ 0, \text{ jos } A \text{ ei tapahdu} \end{cases} \quad ,\quad i = 1, 2, \ldots, n$$

niin

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{f}{n}$$

jossa f on tapahtuman A *frekvenssi* otoksessa. Siten Bernoulli-jakauman odotusarvoparametrin p *estimaattori* $\hat{p}$ on tapahtuman A *suhteellinen frekvenssi* otoksessa.

Huomaa, että

$$f \quad \mathrm{Bin}(n, p)$$

Bernoulli-jakauman odotusarvoparametrin luottamusväli

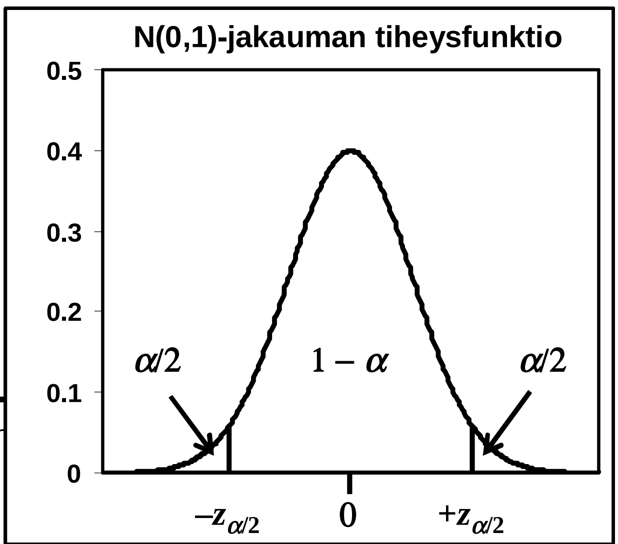
Valitaan *luottamustasoksi*

$$1-\alpha$$

Luottamustaso kiinnittää todennäköisyyden, jolla konstruoitava luottamusväli peittää Bernoulli-jakauman odotusarvoparametrin p todellisen arvon.

Määrätään *luottamuskertoimet* $-z_{\alpha/2}$ ja $+z_{\alpha/2}$ siten, että

$$\Pr(z \le -z_{\alpha/2}) = \frac{\alpha}{2}$$
$$\Pr(z \ge +z_{\alpha/2}) = \frac{\alpha}{2}$$



jossa satunnaismuuttuja Z noudattaa standardoitua normaalijakaumaa:

$$Z \quad \mathrm{N}(0,1)$$

Siten luottamuskertoimet $-z_{\alpha/2}$ ja $+z_{\alpha/2}$ toteuttavat ehdon

$$\Pr(-z_{\alpha/2} \le Z \le +z_{\alpha/2}) = 1 - \alpha$$

Bernoulli-jakauman odotusarvoparametrin p approksimatiivinen luottamusväli luottamustasolla $(1 - \alpha)$ on muotoa

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ,\ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

jossa

$\hat{p}$	=	odotusarvoparametrin p *harhaton estimaattori*
n	=	*havaintojen lukumäärä*
$-z_{\alpha/2}$ ja $+z_{\alpha/2}$	=	luottamustasoon $(1 - \alpha)$ liittyvät *luottamuskertoimet standardoidusta normaalijakaumasta* N(0,1)

Perustelu:

Olkoon

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

satunnaisotos Bernoulli-jakaumasta Ber(p) ja olkoon

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

harhaton estimaattori parametrille p. Huomaa, että $\hat{p}$ on havaintojen $X_1, X_2, \ldots, X_n$ *aritmeettinen keskiarvo*.

Määritellään satunnaismuuttuja

$$Z = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}}$$

Voidaan osoittaa, että satunnaismuuttuja Z noudattaa *suurissa otoksissa approksimatiivisesti* (*asymptoottisesti*) *standardoitua normaalijakaumaa* N(0,1) :

$$Z \quad {}_{a}\ \mathrm{N}(0,1)$$

Huomaa, että tarkastelimme satunnaismuuttujan $\hat{p}$ otosjakaumaa luvun **Otokset ja otosjakaumat** kappaleessa **Suhteellisen frekvenssin otosjakauma**; ks. myös monisteen **Todennäköisyyslaskenta** lukua **Stokastiikan konvergenssikäsitteet ja raja-arvolauseet**. Satunnaismuuttujan $\hat{p}$ otosjakaumaa ei voida suoraan soveltaa tässä (miksi?), mutta tulos voidaan modifioida tässä tarvittavaan muotoon.

Määrätään standardoidusta normaalijakaumasta piste $+ z_{\alpha/2}$ siten, että

$$\Pr(Z \geq +z_{\alpha/2}) = \frac{\alpha}{2}$$

jolloin standardoidun normaalijakauman *symmetrian* perusteella

$$\Pr(Z \leq -z_{\alpha/2}) = \frac{\alpha}{2}$$

ja edelleen

$$\Pr(-z_{\alpha/2} \leq Z \leq +z_{\alpha/2}) = 1-\alpha$$

Tarkastellaan epäyhtälöketjua

$$-z_{\alpha/2} \leq Z \leq +z_{\alpha/2}$$

Sijoittamalla tähän epäyhtälöketjuun satunnaismuuttujan Z lauseke, saadaan epäyhtälöketju

$$-z_{\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \leq +z_{\alpha/2}$$

Tästä epäyhtälöketjusta saadaan sen kanssa *yhtäpitävä* epäyhtälöketju

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Yhdistämällä saatu epäyhtälö siihen, että

$$\Pr(-z_{\alpha/2} \leq Z \leq +z_{\alpha/2}) = 1-\alpha$$

saadaan vihdoin

$$\Pr\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) =_a 1-\alpha$$

■

Koska luottamusväli

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\,,\ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

on *symmetrinen* keskipisteensä $\hat{p}$ suhteen, luottamusväli esitetään usein muodossa

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Luottamusvälin pituus on

$$2 \times z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Luottamusvälin konstruktiosta seuraa, että

$$\Pr\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) =_a 1-\alpha$$

Siten luottamusväli *peittää* parametrin p todellisen arvon approksimatiivisesti todennäköisyydellä $(1-\alpha)$ ja se *ei peitä* parametrin p todellista arvoa approksimatiivisesti todennäköisyydellä α.

Luottamusvälin ominaisuudet

(i) Bernoulli-jakauman odotusarvoparametrin p luottamusvälin *keskipiste* $\hat{p}$ vaihtelee otoksesta toiseen.

(ii) Luottamusvälin *pituus vaihtelee* otoksesta toiseen.

(iii) Luottamusvälin *pituus* riippuu valitusta luottamustasosta $(1-\alpha)$, havaintojen lukumäärästä n ja estimaattorista $\hat{p}$.

(iv) Luottamusväli *lyhenee* (*pitenee*), jos *luottamustasoa* $(1-\alpha)$ *pienennetään* (*kasvatetaan*).

(v) Luottamusväli *lyhenee* (*pitenee*), jos *havaintojen lukumäärää n kasvatetaan* (*pienennetään*).

(vi) Luottamusväli on *lyhimmillään*, kun

$$\hat{p} \approx 0 \text{ tai } 1$$

(vii) Luottamusväli on *pisimmillään*, kun

$$\hat{p} = \frac{1}{2}$$

Luottamusvälin frekvenssitulkinta

Bernoulli-jakauman odotusarvoparametrin p approksimatiivisella luottamusvälillä on seuraava *frekvenssitulkinta*:

(i) Jos otantaa jakaumasta Ber(p) toistetaan, niin keskimäärin

$$100\times(1-\alpha)\ \%$$

otoksista konstruoiduista luottamusväleistä *peittää* parametrin p todellisen arvon.

(ii) Jos otantaa jakaumasta Ber(p) toistetaan, niin keskimäärin

$$100\times\alpha\ \%$$

otoksista konstruoiduista luottamusväleistä *ei peitä* parametrin p todellista arvoa.

Johtopäätökset luottamusvälistä

Oletetaan, että olemme tehneet *johtopäätöksen*, että luottamusväli peittää odotusarvoparametrin p todellisen arvon:

(i) Luottamusvälin konstruktiosta seuraa, että tehty johtopäätös on *oikea* keskimäärin

$$100\times(1-\alpha)\ \%\text{:ssa}$$

tapauksia.

(ii) Luottamusvälin konstruktiosta seuraa, että tehty johtopäätös on *väärä* keskimäärin

$$100\times\alpha\ \%\text{:ssa}$$

tapauksia.

Virheellisen johtopäätöksen mahdollisuutta ei saada häviämään, ellei luottamusväliä tehdä äärettömän leveäksi, jolloin väli *ei enää sisällä informaatiota* odotusarvoparametrin p todellisesta arvosta.

Vaatimukset luottamusvälille

Olisi toivottavaa pystyä konstruoimaan parametrille p mahdollisimman *lyhyt* luottamusväli, johon liittyvä luottamustaso olisi samanaikaisesti mahdollisimman *korkea*.

Molempien vaatimusten samanaikainen täyttäminen *ei ole* kuitenkaan mahdollista, *jos otoskoko pidetään kiinteänä*:

(i) *Luottamustason kasvattaminen pidentää luottamusväliä*, jolloin tieto parametrin p todellisen arvon sijainnista tulee *epätarkemmaksi*.

(ii) *Luottamusvälin lyhentäminen pienentää luottamustasoa*, jolloin tieto parametrin p todellisen arvon sijainnista tulee *epävarmemmaksi*.

Otoskoon määrääminen

Oletetaan, että Bernoulli-jakauman odotusarvoparametrille p halutaan konstruoida luottamusväli, jonka *toivottu pituus* on

$$2A$$

Tarvittava *otoskoko* saadaan kaavasta

$$n = \left(\frac{z_{\alpha/2}\sqrt{p(1-p)}}{A} \right)^2$$

Tarvittava otoskoko *saavuttaa maksiminsa*

$$n = \left(\frac{z_{\alpha/2}}{2A} \right)^2$$

kun

$$p = \frac{1}{2}$$