Todennäköisyyslaskennan ja tilastotieteen peruskurssi

Esimerkkikokoelma 4

Aiheet:

Tilastollisten aineistojen kerääminen ja mittaaminen
Tilastollisten aineistojen kuvaaminen
Otokset ja otosjakaumat
Estimointi
Estimointimenetelmät
Väliestimointi

Avainsanat:

Aritmeettinen keskiarvo
Bernoulli-jakauma
Estimaatti
Estimaattori
Frekvenssi
Frekvenssijakauma
Frekvenssitulkinta
Geometrinen keskiarvo
Harhaton estimaattori
Harmoninen keskiarvo
Hajontaluku
Histogrammi
Intervalliasteikko
Järjestysasteikko
Järjestystunnusluku
Keskihajonta
Keskeinen raja-arvolause
Keskiluku
χ^2-jakauma
Kvalitatiivinen muuttuja
Kvantitatiivinen muuttuja
Laatueroasteikko
Luokiteltu frekvenssijakauma
Maksimi
Mediaani
Minimi
Mittaaminen
Mitta-asteikko
Mittari
Nominaaliasteikko
Normaalijakauma
Ordinaaliasteikko
Otos
Otoskeskihajonta
Otosjakauma
Otostunnusluku
Otosvarianssi
Perusjoukko
Pylväsdiagrammi
Satunnaisotos
Standardipoikkeama
Suhdeasteikko
Suhteellinen frekvenssi
t-jakauma
Tilastollinen aineisto
Tilastollinen muuttuja
Vaihteluväli
Vaihteluvälin pituus
Varianssi
Välimatka-asteikko
Yksinkertainen satunnaisotos
Logaritminen uskottavuusfunktio
Luottamuskerroin
Luottamustaso
Luottamusväli
Maksimointi
Momentti
Odotusarvo
Otoskoko
Riippumattomuus
Standardoitu normaalijakauma
Suhteellinen osuus
Suurimman uskottavuuden estimaattori
Suurimman uskottavuuden menetelmä
Todennäköisyys
Uskottavuusfunktio
Yksinkertainen satunnaisotos

Tilastollisten aineistojen kerääminen ja mittaaminen

Tilastolliset aineistot

Tilastollisen tutkimuksen *kaikki mahdolliset kohteet* muodostavat tutkimuksen (*kohde-*) **perusjoukon**. Tutkimuksen kohteita tarkastellaan aina *perusjoukon muodostamassa kehikossa*. Tutkimuksen kohteiksi valittuja *perusjoukon alkioita* kutsutaan **havaintoyksiköiksi**. **Tilastollinen aineisto** koostuu havaintoyksiköiden *ominaisuuksia* ja *olosuhteita* kuvaavista *numeerisista* tai *kvantitatiivisista tiedoista*. Havaintoyksiköitä koskevia numeerisia tai kvantitatiivisia tietoja kutsutaan **havaintoarvoiksi** tai **havainnoiksi**.

Tilastollisten aineistojen kerääminen

Muutetaanko tutkimuksessa tutkimuksen kohteiden olosuhteita *aktiivisesti*?

(i) Tutkimus on **koe**, jos tutkimuksen tavoitteena on selvittää, miten kohteiden olosuhteiden *aktiivinen muuttaminen* vaikuttaa tutkimuksen kohteisiin.

(ii) Tutkimus perustuu **suoriin havaintoihin**, jos tutkimuksen tavoitteena on *vain seurata*, miten kohteiden *olosuhteet* ja *niissä tapahtuvat muutokset* vaikuttavat kohteisiin.

Kohdistuuko tutkimus *kaikkiin* perusjoukon alkioihin vai johonkin perusjoukon *osaan*?

(i) Tutkimusta kutsutaan **kokonaistutkimukseksi**, jos *kaikki perusjoukon alkiot tutkitaan*.

(ii) Tutkimusta kutsutaan **otantatutkimukseksi**, jos *tutkimus kohdistuu johonkin perusjoukon osajoukkoon*.

Mittaaminen ja mittarit

Tilastollisen tutkimuksen *kohteiden ominaisuuksia* ja *olosuhteita* sekä niiden muutoksia kuvaavat *numeeriset* tai *kvantitatiiviset tiedot* saadaan selville *mittaamalla*. **Mittaaminen** tarkoittaa *numeeristen arvojen liittämistä* tutkimuksen kohteiden ominaisuuksiin ja olosuhteisiin. **Mittaria** voidaan pitää *funktiona*, joka *liittää numeeriset arvot* tutkimuksen kohteiden ominaisuuksiin ja olosuhteisiin.

Mittauksen tulos voidaan siis aina ilmaista jonkin tutkimuksen kohteen ominaisuutta tai olosuhdetta kuvaavan *muuttujan arvona*. Siten tutkimuksen kohteiden ominaisuuksia ja olosuhteita kuvataan mittaustapahtumassa *numeerisilla muuttujilla*.

Mittarin validiteetti ja tarkkuus

Mittari on **validi** eli *oikea*, jos se esittää mittauksen kohteena olevaa ominaisuutta *oikein, merkityksellisesti* ja *tarkoituksenmukaisesti*.

Mittari on **tarkka**, jos se on *harhaton* ja *reliaabeli*:

(i) Mittari on **harhaton**, jos se *ei systemaattisesti ali- tai yliarvioi* mitattavan ominaisuuden määrää.

(ii) Mittari on **reliaabeli** eli *luotettava*, jos mittaustulos *ei muutu*, kun mittausta toistetaan.

Mitta-asteikot

Mittaus on tehty **nominaali-** eli **laatueroasteikolla**, jos mittaus kertoo *mihin luokkaan* mittauksen kohde kuuluu.

Mittaus on tehty **ordinaali-** eli **järjestysasteikolla**, jos mittaus kertoo onko mittauksen kohteella mitattavaa ominaisuutta *enemmän* tai *vähemmän* kuin jollakin toisella kohteella.

Mittaus on tehty **intervalli-** eli **välimatka-asteikolla**, jos mittaus kertoo *kuinka paljon* kahden mitattavan kohteen ominaisuudet *eroavat* toisistaan.

Mittaus on tehty **ratio**- eli **suhdeasteikolla**, jos mittaus kertoo *kuinka monta kertaa enemmän* tai *vähemmän* mittauksen kohteella on mitattavaa ominaisuutta kuin jollakin toisella kohteella.

Kvalitatiiviset ja kvantitatiiviset muuttujat

Ominaisuutta ja sitä kuvaavaa muuttujaa kutsutaan **kvalitatiiviseksi**, jos mittauksen kohteet voidaan *luokitella* mittauksen perusteella toisistaan eroaviin *kategorioihin* tai *luokkiin*. Kvalitatiivisia ominaisuuksia kuvataan *laatueroasteikollisilla muuttujilla*.

Ominaisuutta ja sitä kuvaavaa muuttujaa kutsutaan **kvantitatiiviseksi**, jos mittaus tuottaa ominaisuuden *määrällisen arvon*. Kvantitatiivisia ominaisuuksia kuvataan *välimatka*- tai *suhde-asteikollisilla muuttujilla*.

Diskreetit ja jatkuvat muuttujat

Mitattavaa ominaisuutta vastaava muuttuja on **diskreetti**, jos se voi saada vain *erillisiä arvoja*. Diskreettejä muuttujia ovat esimerkiksi laatueroasteikollisten muuttujien ja sijalukuja kuvaavien järjestysasteikollisten muuttujien lisäksi myös sellaiset kvantitatiiviset muuttujat kuten lukumäärä-muuttujat.

Mitattavaa ominaisuutta vastaava muuttuja on **jatkuva**, jos se voi saada *kaikki arvot joltakin väliltä*. Jatkuvia muuttujia ovat esimerkiksi useimmat fysikaaliset suureet kuten pituus, pinta-ala, tilavuus, paino, aika, nopeus ja paine sekä myös monet talouselämää kuvaavat suureet kuten rahamäärä ja korko.

Huomautus:

> Muuttujien mitta-asteikollisilla ominaisuuksilla (*kvalitatiivisuudella*/*kvantitatiivisuudella* tai *diskreettiydellä*/*jatkuvuudella*) on syvällinen vaikutus siihen, mitä tilastollisia menetelmiä kyseisessä tilanteessa on luvallista (tai suotavaa) soveltaa.

Tilastollisten aineistojen kuvaaminen

Frekvenssit

Olkoon muuttuja *x* ***diskreetti*** ja oletetaan, että sen *mahdolliset arvot* ovat

$$y_1, y_2, \ldots, y_m$$

Olkoot

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

muuttujan *x havaitut arvot*. Muuttujan *x* mahdollisen arvon y_k , $k = 1, 2, \ldots, m$ **frekvenssi**

$$f_k$$

kertoo *kuinka monta kertaa* y_k esiintyy havaintoarvojen $x_1, x_2, \ldots, x_n$ joukossa.

Frekvenssijakauma

Muuttujan *x mahdolliset arvot*

$$y_1, y_2, \ldots, y_m$$

yhdessä niiden *frekvenssien*

$$f_1, f_2, \ldots, f_m$$

kanssa muodostavat muuttujan *x havaittujen arvojen* $x_1, x_2, \ldots, x_n$ **frekvenssijakauman**. Huomaa, että

$$f_1 + f_2 + \cdots + f_m = n$$

jossa n on havaintojen kokonaislukumäärä.

Pylväsdiagrammi

Frekvenssijakaumaa

$$(y_k, f_k)\ ,\ k = 1, 2, \ldots, m$$

voidaan kuvata graafisesti **pylväsdiagrammilla**, jossa muuttujan x mahdollisen arvon y_k havaintoarvojen $x_1, x_2, \ldots, x_n$ joukossa esittää *pylväs*, jonka *korkeus vastaa frekvenssiä* f_k .

Huomautus:

> Pylväsdiagrammin tulkinta on samantapainen kuin *diskreetin todennäköisyysjakauman pistetodennäköisyysfunktion* tulkinta.

Luokkafrekvenssit

Olkoon muuttuja x ***jatkuva*** ja oletetaan, että sen *mahdolliset arvot* ovat *välillä*

$$(a, b)$$

jossa voi olla $a = -\infty$, $b = +\infty$. Jaetaan väli (a, b) pisteillä

$$a = a_0 < a_1 < a_2 < \quad \cdots \quad < a_{m-1} < a_m = b$$

pistevieraisiin *osaväleihin*

$$(a_{k-1}, a_k]\ ,\ k = 1, 2, \ldots, m$$

Olkoot

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

muuttujan *x havaitut arvot*. Muuttujan x havaittujen arvojen **frekvenssi**

$$f_k$$

luokassa k kertoo niiden havaintoarvojen $x_1, x_2, \ldots, x_n$ *lukumäärän*, jotka kuuluvat väliin

$$(a_{k-1}, a_k]\ ,\ k = 1, 2, \ldots, m$$

Luokiteltu frekvenssijakauma

Luokkavälit

$$(a_{k-1}, a_k]\ ,\ k = 1, 2, \ldots, m$$

yhdessä vastaavien *luokkafrekvenssien*

$$f_1, f_2, \ldots, f_m$$

kanssa muodostavat muuttujan *x havaittujen arvojen* $x_1, x_2, \ldots, x_n$ **luokitellun frekvenssijakauman**. Huomaa, että

$$f_1 + f_2 + \cdots + f_m = n$$

jossa n on havaintojen kokonaislukumäärä.

Histogrammi

Luokiteltua frekvenssijakaumaa

$$((a_{k-1}, a_k], f_k)\ ,\ k = 1, 2, \ldots , m$$

voidaan kuvata graafisesti **histogrammilla**, jossa muuttujan x havaittujen arvojen

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

frekvenssiä f_k luokassa $(a_{k-1}, a_k]$, esittää *suorakaide* (*nelikulmio*), jonka *kantana* on väli

$$(a_{k-1}, a_k]$$

ja jonka *pinta-ala* vastaa luokkafrekvenssiä f_k .

Huomautuksia:

Histogrammin tulkinta on samantapainen kuin *jatkuvan todennäköisyysjakauman tiheysfunktion* tulkinta.

Jos luokiteltua frekvenssijakaumaa muodostettaessa käytetään *tasavälistä luokitusta*, niin luokiteltua frekvenssijakaumaa kuvaavan histogrammikuvion muodostavien nelikulmioiden *korkeudet* vastaavat luokkafrekvenssejä f_k .

Suhdeasteikollisten muuttujien tunnusluvut

Aritmeettinen keskiarvo

Olkoot

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

välimatka- tai *suhdeasteikollisen* muuttujan x havaittuja arvoja. Lukujen $x_1, x_2, \ldots , x_n$ **aritmeettinen keskiarvo** on

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

Aritmeettinen keskiarvo on havaintoarvojen *painopiste* ja kuvaa havaintoarvojen *keskimääräistä* arvoa.

Varianssi

Olkoot

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

välimatka- tai *suhdeasteikollisen* muuttujan x havaittuja arvoja. Lukujen $x_1, x_2, \ldots , x_n$ (**otos-**) **varianssi** on

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)$$

jossa

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

on lukujen $x_1, x_2, \ldots , x_n$ *aritmeettinen keskiarvo*. Otosvarianssi kuvaa havaintoarvojen *hajaantuneisuutta* (tai *keskittyneisyyttä*) niiden aritmeettisen keskiarvon (painopisteen) ympärillä.

Aritmeettisen keskiarvon ja varianssi laskeminen

Olkoot

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

välimatka- tai *suhdeasteikollisen* muuttujan *x* havaittuja arvoja. Jos havaintoarvojen $x_1, x_2, \ldots, x_n$ aritmeettinen keskiarvo ja varianssi joudutaan laskemaan *käsin* tai *laskinta* käyttäen, kannattaa laskut järjestää alla olevan taulukon muotoon ja käyttää taulukon vieressä esitettyjä kaavoja.

i	x_i	x_i^2
1	x_1	x_1^2
2	x_2	x_2^2
⋮	⋮	⋮
n	x_n	x_n^2
Summa	$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} x_i^2$

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

$$s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)$$

Keskihajonta

Olkoot

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

välimatka- tai *suhdeasteikollisen* muuttujan *x* havaittuja arvoja. Lukujen $x_1, x_2, \ldots, x_n$ (**otos-**) **keskihajonta** on

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)} = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)} = \sqrt{s_x^2}$$

jossa

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

on lukujen $x_1, x_2, \ldots, x_n$ *aritmeettinen keskiarvo* ja s_x^2 on lukujen $x_1, x_2, \ldots, x_n$ (*otos-*) *varianssi*. Otoskeskihajonta kuvaa (kuten otosvarianssi) havaintoarvojen *hajaantuneisuutta* (tai *keskittyneisyyttä*) niiden aritmeettisen keskiarvon (painopisteen) ympärillä.

Standardointi

Olkoon $\bar{x}$ *välimatka-* tai *suhdeasteikollisen muuttujan x* havaittujen arvojen $x_1, x_2, \ldots, x_n$ *aritmeettinen keskiarvo* ja s_x^2 niiden *varianssi*. Tällöin **standardoitujen** *havaintoarvojen*

$$z_i = \frac{x_i - \bar{x}}{s_x},\ i = 1, 2, \ldots, n$$

aritmeettinen keskiarvo ja varianssi ovat

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = 0$$

$$s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2 = 1$$

Tilastollinen etäisyys

Olkoot $\bar{x}$ *välimatka-* tai *suhdeasteikollisen muuttujan x* havaittujen arvojen $x_1, x_2, \ldots, x_n$ *aritmeettinen keskiarvo* ja s_x^2 niiden *varianssi*. Tällöin havaintoarvojen x_k ja x_l *tilastollinen etäisyys* on

$$d_{kl} = \frac{x_k - x_l}{s_x}$$

Origomomentit

Olkoot

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

välimatka- tai *suhdeasteikollisen* muuttujan *x* havaittuja arvoja. Lukujen $x_1, x_2, \ldots, x_n$ ***k*. origomomentti** on

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

Keskusmomentit

Olkoot

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

välimatka- tai *suhdeasteikollisen* muuttujan *x* havaittuja arvoja. Lukujen $x_1, x_2, \ldots, x_n$ ***k*. keskusmomentti** on

$$m_x^k = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^k$$

jossa

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

on lukujen $x_1, x_2, \ldots, x_n$ *aritmeettinen keskiarvo*.

Vinous

Olkoot

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

välimatka- tai *suhdeasteikollisen* muuttujan *x* havaittuja arvoja. Havaintoarvojen $x_1, x_2, \ldots, x_n$ jakauman **vinoutta** voidaan kuvata otostunnusluvulla

$$c_1 = \frac{m_3}{m_2^{3/2}}$$

jossa

m_2 = 2. *keskusmomentti* luvuille $x_1, x_2, \ldots, x_n$

m_3 = 3. *keskusmomentti* luvuille $x_1, x_2, \ldots, x_n$

Huipukkuus

Olkoot

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

välimatka- tai *suhdeasteikollisen* muuttujan x havaittuja arvoja. Havaintoarvojen $x_1, x_2, \ldots, x_n$ jakauman **huipukkuutta** voidaan kuvata otostunnusluvulla

$$c_2 = \frac{m_4}{m_2^2}$$

jossa

m_2 = 2. *keskusmomentti* luvuille $x_1, x_2, \ldots, x_n$

m_4 = 4. *keskusmomentti* luvuille $x_1, x_2, \ldots, x_n$

Geometrinen keskiarvo

Olkoot

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

positiivisia lukuja. Lukujen $x_1, x_2, \ldots, x_n$ **geometrinen keskiarvo** on

$$G = \sqrt[n]{x_1 x_2 \cdots x_n}$$

Lukujen $x_1, x_2, \ldots, x_n$ geometrisen keskiarvon logaritmi on lukujen $x_1, x_2, \ldots, x_n$ logaritmien aritmeettinen keskiarvo:

$$\log(G) = \frac{\log(x_1) + \log(x_2) + \cdots + \log(x_n)}{n} = \frac{1}{n}\sum_{i=1}^{n} \log(x_i)$$

Harmoninen keskiarvo

Olkoot

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

positiivisia lukuja. Lukujen $x_1, x_2, \ldots, x_n$ **harmoninen keskiarvo** on

$$H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}}$$

Lukujen $x_1, x_2, \ldots, x_n$ harmonisen keskiarvon käänteisluku on lukujen $x_1, x_2, \ldots, x_n$ käänteislukujen aritmeettinen keskiarvo:

$$\frac{1}{H} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}$$

Aritmeettinen, harmoninen ja geometrinen keskiarvo

Oletetaan, että aritmeettinen keskiarvo M, harmoninen keskiarvo H ja geometrinen keskiarvo G lasketaan *samoista luvuista* $x_1, x_2, \ldots, x_n$.

Tällöin

$$H \le G \le M$$

ja

$$H = G = M$$

vain, jos

$$x_1 = x_2 = \cdots = x_n$$

Järjestysasteikollisten muuttujien tunnusluvut

Järjestystunnusluvut

Olkoot

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

järjestys-, *välimatka-* tai *suhdeasteikollisen* muuttujan x havaittuja arvoja. *Järjestetään* havainto-arvot $x_1, x_2, \ldots, x_n$ suuruusjärjestykseen pienimmästä suurimpaan ja olkoot

$$z_1, z_2, \ldots, z_n$$

järjestykseen asetetut havaintoarvot. Suuruusjärjestyksessä k. havaintoarvoa z_k kutsutaan ***k.* järjestystunnusluvuksi**.

Minimi, maksimi, vaihteluväli

Olkoot

$$z_1, z_2, \ldots, z_n$$

järjestys-, *välimatka-* tai *suhdeasteikollisen muuttujan x* havaitut arvot järjestettyinä *suuruus-järjestykseen* pienimmästä suurimpaan. Tällöin

z_1 = **minimiarvo**

z_n = **maksimiarvo**

(z_1 , z_n) = **vaihteluväli**

$z_n - z_1$ = **vaihteluvälin pituus**

Prosenttipisteet

Olkoot

$$z_1, z_2, \ldots, z_n$$

järjestys-, *välimatka-* tai *suhdeasteikollisen muuttujan x* havaitut arvot järjestettyinä *suuruus-järjestykseen* pienimmästä suurimpaan. Havaintoarvojen ***p. prosenttipiste***

$$z_{(p)}\ ,\ p = 1, 2, \ldots, 99$$

on piste, joka jakaa havaintoaineiston *kahteen osaan*:

(i) p % havaintoarvoista on lukua $z_{(p)}$ *pienempiä* tai korkeintaan yhtä suuria kuin $z_{(p)}$.

(ii) $(100 - p)$ % havaintoarvoista on lukua $z_{(p)}$ *suurempia*.

Mediaani

Olkoot

$$z_1, z_2, \ldots, z_n$$

järjestys-, *välimatka-* tai *suhdeasteikollisen muuttujan x* havaitut arvot järjestettyinä *suuruusjärjestykseen* pienimmästä suurimpaan.

Mediaani *Me* on havaintoarvojen 50. *prosenttipiste*:

$$Me = z_{(50)}$$

Mediaani jakaa havaintoaineiston *kahteen yhtä suureen osaan* niin, että toisessa *kaikki* havaintoarvot ovat mediaania *pienempiä*, toisessa *kaikki* havaintoarvot ovat mediaania *suurempia*.

Havaintoarvojen mediaani *Me* voidaan määrätä seuraavalla tavalla:

(1) Järjestetään havaintoarvot *suuruusjärjestykseen* pienimmästä suurimpaan.

(2a) Jos havaintoarvojen lukumäärä on *pariton*, mediaani on järjestetyistä havaintoarvoista *keskimmäinen*.

(2b) Jos havaintoarvojen lukumäärä on *parillinen*, mediaani on järjestetyistä havaintoarvoista *kahden keskimmäisen aritmeettinen keskiarvo*.

Oletetaan, että *aritmeettinen keskiarvo M* ja *mediaani Me* määrätään *samasta* jatkuvan muuttujan havaittujen arvojen *luokitellusta frekvenssijakaumasta*. Jos havaintoarvojen jakauma on *yksihuippuinen*, pätee seuraava:

Vasemmalle vinoilla jakaumilla

$$M < Me$$

Symmetrisillä jakaumilla

$$M \approx Me$$

Oikealle vinoilla jakaumilla

$$Me < M$$

Kvartiilit

Olkoot

$$z_1, z_2, \ldots, z_n$$

järjestys-, *välimatka-* tai *suhdeasteikollisen muuttujan x* havaitut arvot järjestettyinä *suuruusjärjestykseen* pienimmästä suurimpaan. Tällöin

Q_1 = **Alakvartiili** = 25. prosenttipiste = $z_{(25)}$

Q_2 = **Keskikvartiili** = 50. prosenttipiste = $z_{(50)}$

Q_3 = **Yläkvartiili** = 75. prosenttipiste = $z_{(75)}$

Kvartiilit Q_1, Q_2, Q_3 jakavat suuruusjärjestykseen asetetun havaintoaineiston *neljään yhtä suureen osaan*.

Erityisesti:

Alakvartiili Q_1 = Havaintoarvojen mediaania *Me pienempien havaintoarvojen mediaani*

Keskikvartiili Q_2 = Havaintoarvojen *mediaani Me*

Yläkvartiili Q_3 = Havaintoarvojen mediaania *Me suurempien havaintoarvojen mediaani*

Kvartiilit, kvartiiliväli, kvartiilipoikkeama

Olkoot havaintoarvojen *kvartiilit* Q_1, Q_2, Q_3. Tällöin

(Q_1 , Q_3) = **kvartiiliväli**

$Q_3 - Q_1$ = IQR = **kvartiilivälin pituus**

$(Q_3 - Q_1)/2$ = $IQR/2$ = **kvartiilipoikkeama**

Kvartiiliväliä, kvartiilivälin pituutta (IQR = interquartile range) ja kvartiilipoikkeamaa voidaan käyttää kuvaamaan havaintoarvojen *hajaantuneisuutta* (*keskittyneisyyttä*). Jos havaintoarvojen jakaumaa kuvaavana *keskilukuna* on käytetty *mediaania*, *hajontalukuna* käytetään usein *kvartiilipoikkeamaa*.

Laatueroasteikollisten muuttujien tunnusluvut

Frekvenssi

Olkoon *otoskoko* eli kerättyjen *havaintoarvojen lukumäärä n*. Olkoon *A* jokin perusjoukon osajoukko ja olkoon *f* otokseen kuuluvien *A*-tyyppisten havaintoarvojen *frekvenssi* eli *lukumäärä*. Tällöin *A*-tyyppisten havaintoarvojen **suhteellinen frekvenssi** eli **osuus** otoksessa on

$$\frac{f}{n}$$

Moodi

Frekvenssijakauman **moodi** eli *tyyppiarvo Mo* on *yleisin* havaintoarvo. *Luokitellun frekvenssijakauman* **moodi** eli *tyyppiarvo Mo* on siinä luokassa, jossa luokiteltua frekvenssijakaumaa *vastaava histogrammi saavuttaa maksiminsa*.

Huomautuksia:

Jos käytetty luokitus *on tasavälinen*, luokitellun frekvenssijakauman moodi *on* siinä luokassa, jota vastaava frekvenssi on suurin.

Jos käytetty luokitus *ei ole tasavälinen*, luokitellun frekvenssi jakauman moodi *ei välttämättä ole* siinä luokassa, jota vastaava frekvenssi on suurin.

Oletetaan, että *aritmeettinen keskiarvo M*, *mediaani Me* ja *moodi Mo* määrätään *samasta* jatkuvan muuttujan havaittujen arvojen *luokitellusta frekvenssijakaumasta*. Jos havaintoarvojen jakauma on *yksihuippuinen*, pätee seuraava:

Vasemmalle vinoilla jakaumilla

$$M < Me < Mo$$

Symmetrisillä jakaumilla

$$M \approx Me \approx Mo$$

Oikealle vinoilla jakaumilla

$$Mo < Me < M$$

Otokset ja otosjakaumat

Yksinkertainen satunnaisotos

Olkoot havainnot $X_1, X_2, \ldots, X_n$ *riippumattomia, identtisesti jakautuneita satunnaismuuttujia*, joilla on *sama pistetodennäköisyys-* tai *tiheysfunktio* $f(x)$:

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

Tällöin sanomme, että havainnot

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

muuodostavat (**yksinkertaisen**) **satunnaisotoksen** jakaumasta, jonka *pistetodennäköisyys-* tai *tiheysfunktio* on $f(x)$.

Otostunnusluku

Olkoon

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

yksinkertainen satunnaisotos jakaumasta, jonka *pistetodennäköisyys-* tai *tiheysfunktio* on $f(x)$.

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 *yksinkertaisen satunnaisotoksen* jakaumasta $f(x)$ 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.*

Aritmeettisen keskiarvon ja otosvarianssin otosjakaumat

Aritmeettinen keskiarvo ja otosvarianssi

Oletetaan, että havainnot

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

muodostavat *yksinkertaisen 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 odotusarvo* ja *varianssi*:

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

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

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

Otoksen ominaisuuksia voidaan kuvata havaintoarvojen *aritmeettisella keskiarvolla* ja *varianssilla*.

Määritellään havaintojen $X_1, X_2, \ldots, X_n$ *aritmeettinen keskiarvo* kaavalla

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

Määritellään havaintojen $X_1, X_2, \ldots, X_n$ *otosvarianssi* kaavalla

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

Huomaa, että sekä aritmeettinen keskiarvo $\bar{X}$ että otosvarianssi s^2 ovat havaintojen $X_1, X_2, \ldots, X_n$ funktioina *satunnaismuuttujia*, joiden saamat arvot vaihtelevat satunnaisesti otoksesta toiseen.

Summan odotusarvo ja varianssi

Havaintojen $X_1, X_2, \ldots, X_n$ *summalla* $\sum X_i$ on em. oletusten pätiessä seuraava *odotusarvo* ja *varianssi*:

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

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = n\sigma^2$$

Aritmeettisen keskiarvon odotusarvo ja varianssi

Havaintojen $X_1, X_2, \ldots, X_n$ *aritmeettisella keskiarvolla* $\bar{X}$ on em. oletusten pätiessä seuraava *odotusarvo* ja *varianssi*:

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

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

Huomaa, että havaintojen $X_1, X_2, \ldots, X_n$ aritmeettisen keskiarvon $\bar{X}$ varianssi otoksessa on aina pienempi kuin havaintojen varianssi, jos otoskoko $n > 1$. Lisäksi aritmeettisen keskiarvon varianssi $\bar{X}$ *pienenee*, jos otoskoon n annetaan kasvaa.

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.

Otosvarianssin odotusarvo

Havaintojen $X_1, X_2, \ldots, X_n$ *otosvarianssilla* s^2 on em. oletusten pätiessä seuraava *odotusarvo*:

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

Normaalijakautuneiden havaintojen summan otosjakauma

Oletetaan, että havainnot

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

muodostavat *yksinkertaisen satunnaisotoksen* normaalijakaumasta $\mathrm{N}(\mu,\sigma^2)$. Tällöin havainnot $X_1, X_2, \ldots, X_n$ ovat *riippumattomia satunnaismuuttujia*, jotka *noudattavat samaa normaalijakaumaa* $\mathrm{N}(\mu,\sigma^2)$:

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

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

Havaintojen $X_1, X_2, \ldots, X_n$ summa $\sum X_i$ noudattaa em. oletusten pätiessä *normaalijakaumaa* parametrein $n\mu$ ja $n\sigma^2$:

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

Normaalijakautuneiden havaintojen aritmeettisen keskiarvon otosjakauma

Oletetaan, että havainnot

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

muodostavat *yksinkertaisen satunnaisotoksen* normaalijakaumasta $\mathrm{N}(\mu,\sigma^2)$. Tällöin havainnot $X_1, X_2, \ldots, X_n$ ovat *riippumattomia satunnaismuuttujia*, jotka *noudattavat samaa normaalijakaumaa* $\mathrm{N}(\mu,\sigma^2)$:

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

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

Havaintojen $X_1, X_2, \ldots, X_n$ aritmeettinen keskiarvo $\bar{X}$ noudattaa em. oletusten pätiessä *normaalijakaumaa* parametrein μ ja σ^2/n :

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

Erityisesti

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

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

mikä pätee myös ilman normaalisuusoletusta.

Aritmeettisen keskiarvon approksimatiivinen (asymptoottinen) otosjakauma

Oletetaan, että havainnot

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

muodostavat *yksinkertaisen 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 :

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

Normaalijakautuneiden havaintojen otosvarianssin otosjakauma

Oletetaan, että havainnot

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

muodostavat *yksinkertaisen satunnaisotoksen* normaalijakaumasta $\mathrm{N}(\mu,\sigma^2)$. Tällöin havainnot $X_1, X_2, \ldots, X_n$ ovat *riippumattomia satunnaismuuttujia*, jotka *noudattavat samaa normaalijakaumaa* $\mathrm{N}(\mu,\sigma^2)$:

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

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

Olkoon s^2 havaintojen $X_1, X_2, \ldots, X_n$ *otosvarianssi*. Satunnaismuuttuja $(n-1)s^2/\sigma^2$ noudattaa em. oletusten pätiessä χ^2*-jakaumaa* vapausastein $(n-1)$:

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

Lisäksi voidaan osoittaa, että aritmeettinen keskiarvo $\bar{X}$ ja otosvarianssi s^2 ovat satunnaismuuttujina *riippumattomia*:

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

Siten suoraan Studentin *t-jakauman* määritelmän mukaan

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

em. oletusten pätiessä.

Suhteellisen frekvenssin otosjakauma

Frekvenssi ja suhteellinen frekvenssi

Olkoon $A \subset S$ jokin otosavaruuden S tapahtuma ja olkoon

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

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

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

vastaava *suhteellinen frekvenssi* eli *osuus*.

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

Frekvenssin odotusarvo, varianssi ja otosjakauma

Frekvenssin *f odotusarvo* ja *varianssi*:

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

jossa $q = 1 - p$.

Frekvenssi f noudattaa otoksessa *binomijakaumaa* parametrein *n* ja $\Pr(A) = p$:

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

Suhteellisen frekvenssin odotusarvo ja varianssi

Suhteellisen frekvenssin $\hat{p} = f / n$ *odotusarvo* ja *varianssi*:

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

jossa $q = 1 - p$. Huomaa, että suhteellisen frekvenssin $\hat{p}$ varianssi *pienenee*, jos otoskoon *n* annetaan kasvaa.

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

Keskeisestä raja-arvolauseesta seuraa, että suhteellinen frekvenssi $\hat{p}$ otoksessa noudattaa em. oletusten pätiessä suurissa otoksissa *approksimatiivisesti normaalijakaumaa*:

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

Estimointi

Satunnaisotos

Olkoon

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

(**yksinkertainen**) **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 todennäköisyysjakauman $f(x;\theta)$ *parametri* θ *on tuntematon* ja sen *estimoimiseen* käytetään havaintojen $X_i\,,\ i = 1,\ 2,\ \ldots\ ,\ n$ funktiota eli (*otos-*) *tunnuslukua*

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

Funktiota $T = g(\ X_1,\ X_2,\ \ldots\ ,\ X_n)$ kutsutaan parametrin θ **estimaattoriksi** ja funktion *g havainto-arvoista*

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

laskettua arvoa

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

kutsutaan parametrin θ **estimaatiksi**.

Otosjakauma

Oletetaan, että havainnot

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

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

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

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*.

Estimaattorin hyvyys

Hyvyyskriteerit estimaattoreille

Todennäköisyysjakaumien parametreille on tavallisesti tarjolla useita vaihtoehtoisia estimaattoreita. Tällöin nousee esiin kysymys siitä, mikä tarjolla olevista estimaattoreista on *paras*. Tärkeimmät estimaattoreiden vertailussa käytetyt kriteerit ovat *tyhjentävyys, harhattomuus, tehokkuus* ja *tarkentuvuus*.

Tyhjentävyys

Olkoon T parametrin θ estimaattori. Estimaattori T on **tyhjentävä** parametrille θ, jos se käyttää kaiken otoksessa olevan *informaation* parametrista θ.

Huomaa, että tässä esitetty määritelmä on pikemminkin tyhjentävyyden kuvaus kuin matemaattinen määritelmä tyhjentävyydelle.

Harhattomuus

Olkoon T parametrin θ estimaattori. Estimaattori T on **harhaton** parametrille θ, jos

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

Jos T on parametrin θ harhaton estimaattori, niin se tuottaa parametrille θ arvoja (estimaatteja), jotka vaihtelevat kyllä otoksesta toiseen, mutta kuitenkin *parametrin θ oikean arvon ympärillä*.

Tehokkuus

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

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

Jos estimaattori T_1 on tehokkaampi kuin estimaattori T_2, niin estimaattori T_1 tuottaa parametrille θ arvoja (estimaatteja), jotka vaihtelevat otoksesta toiseen *vähemmän* kuin estimaattori T_2.

Täystehokkuus (minimivarianssisuus)

Olkoon T parametrin θ estimaattori. Estimaattori T on **täystehokas**, jos sen varianssi

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

on pienempi kuin minkä tahansa muun estimaattorin. Jos estimaattori T on täystehokas, niin se tuottaa parametrille θ arvoja (estimaatteja), jotka vaihtelevat otoksesta toiseen *vähemmän* kuin muut parametrin θ estimaattorit.

Täystehokkuus voidaan tavallisesti saavuttaa vain jossakin tietyssä estimaattoreiden luokassa kuten esimerkiksi harhattomien estimaattoreiden luokassa.

Tarkentuvuus

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

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

Jos T on tarkentuva, se tuottaa parametrille θ arvoja (estimaatteja), jotka lähestyvät havaintojen lukumäärän kasvaessa (eräässä mielessä) parametrin θ oikeata arvoa.

Huomaa, että melkein varma konvergenssi ei ole tavanomaista alkeismatematiikan konvergenssia, vaan kuuluu todennäköisyyslaskennan konvergenssikäsitteiden piiriin. Emme täsmennä tässä melkein varman konvergenssin käsitettä.

Estimointimenetelmät

Suurimman uskottavuuden menetelmä

Yksinkertainen satunnaisotos

Olkoon

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

(**yksinkertainen**) **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* $f(x;\theta)$:

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

Uskottavuusfunktio

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*. Uskottavuusfunktio L sisältää kaiken informaation otoksesta.

Koska havainnot $X_i\ ,\ i = 1,\ 2,\ \ldots\ ,\ n$ muodostavat yksinkertaisen satunnaisotoksen jakaumasta $f(x;\theta)$, otoksen yhteisjakauman tiheysfunktio on muotoa

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

jossa

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

on yksittäiseen havaintoon X_i liittyvä pistetodennäköisyys- tai tiheysfunktio.

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. Huomaa, että uskottavuusfunktion L maksimin antava parametrin θ arvo t on havaintoarvojen (muuttujien) $x_1,\ x_2,\ \ldots\ ,\ x_n$ 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 SU-estimaattori $\hat{\theta}$ tuottaa parametrille θ arvon, joka *maksimoi poimitun otoksen* eli *saatujen havaintoarvojen uskottavuuden* (*todennäköisyyden*). Siten parametrin θ SU-estimaattorin $\hat{\theta}$ otoskohtainen arvo *maksimoi todennäköisyyden saada juuri se otos, joka on saatu*.

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*

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

Logaritminen uskottavuusfunktio

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

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

parametrin θ suhteen.

Koska havainnot $X_1, X_2, \ldots, X_n$ on oletettu tässä *riippumattomiksi*, *logaritminen uskottavuusfunktio* voidaan kirjoittaa seuraavaan muotoon:

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

jossa

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

on yhteen havaintoarvoon x_i liittyvä logaritminen uskottavuusfunktio.

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

Hyvin yleisin ehdoin pätee:

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

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

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

Normaalijakauman parametrien suurimman uskottavuuden estimaattorit

Satunnaismuuttuja X noudattaa **normaalijakaumaa**, 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*

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

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

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

ja *otosvarianssi*

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

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

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

Normaalijakauman $\mathrm{N}(\mu,\sigma^2)$ 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 (voidaan osoittaa, että parametrilla σ^2 ei ole minimivarianssista estimaattoria, ellei parametria μ tunneta).

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

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

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

Bernoulli-jakauman odotusarvoparametrin suurimman uskottavuuden estimaattori

Olkoon *A tapahtuma*, jonka todennäköisyys on *p*:

$$\Pr(A) = p$$

Määritellään satunnaismuuttuja *X* seuraavasti:

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

Satunnaismuuttuja *X* noudattaa **Bernoulli-jakaumaa** parametrilla *p*:

$$X : \ \text{Ber}(p)$$

jossa

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

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

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

Huomaa, että

$$\bar{X} = \frac{f}{n}$$

jossa *f* on kiinnostuksen kohteena olevan tapahtuman *A suhteellinen frekvenssi* otoksessa.

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*:

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

Väliestimointi

Normaalijakauman odotusarvon luottamusväli, kun jakauman varianssi on tunnettu

Otos normaalijakaumasta

Olkoon

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

satunnaisotos normaalijakaumasta N(μ, σ^2). Tällöin satunnaismuuttujat X_i , i = 1, 2, … , n ovat *riippumattomia* ja noudattavat *samaa normaalijakaumaa* N(μ, σ^2):

$$X_1, X_2, \cdots, X_n \perp$$
$$X_i :\ \ \mathrm{N}(\mu, \sigma^2)\ ,\ i = 1, 2, \ldots, n$$

Normaalijakauman parametrien estimointi

Oletetaan, että normaalijakauman N(μ, σ^2) varianssi σ^2 on *tunnettu* ja *estimoidaan* odotusarvo-parametri E(X) = μ sen *harhattomalla estimaattorilla*: Havaintojen *aritmeettinen keskiarvo*

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

on *odotusarvoparametrin* E(X) = μ *harhaton estimaattori*.

Normaalijakauman odotusarvon luottamusväli, kun jakauman varianssi on tunnettu

Valitaan *luottamustasoksi*

$$1 - \alpha$$

Luottamustaso kiinnittää todennäköisyyden, jolla konstruoitava luottamusväli peittää normaali-jakauman odotusarvon μ todellisen arvon.

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

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

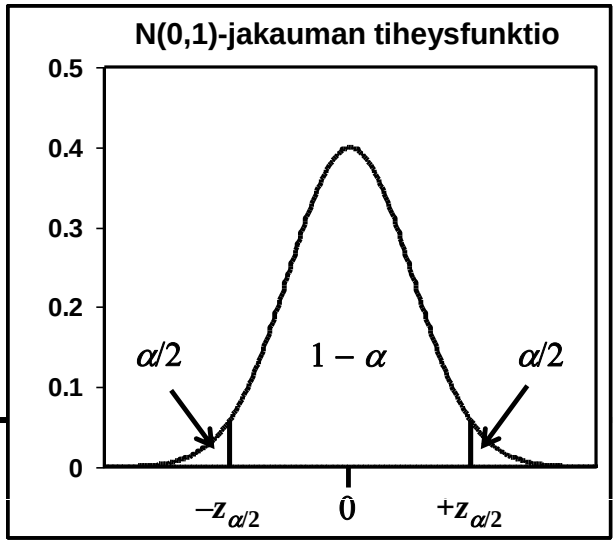
ja

$$\Pr(z \ge +z_{\alpha/2}) = \frac{\alpha}{2}$$

jossa satunnaismuuttuja z noudattaa *standardoitua normaalijakaumaa*:

$$z :\ \ \mathrm{N}(0,1)$$

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



$$\Pr(-z_{\alpha/2} \le z \le +z_{\alpha/2}) = 1-\alpha$$

Normaalijakauman *odotusarvoparametrin* μ *luottamusväli luottamustasolla* $(1-\alpha)$ 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* otoksessa

σ^2 = jakauman *varianssi*

n = havaintojen *lukumäärä*

$-z_{\alpha/2}$ ja $+z_{\alpha/2}$ = luottamustasoon $(1-\alpha)$ liittyvät *luottamuskertoimet standardoidusta normaalijakaumasta* N(0,1)

Luottamusvälin konstruktio perustuu siihen, että

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

Koska luottamusväli 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(μ, σ^2) toistetaan, *keskimäärin* $100\times(1-\alpha)$ % otoksista konstruoiduista luottamusväleistä *peittää* parametrin μ todellisen arvon.

(ii) Jos otantaa jakaumasta N(μ, σ^2) toistetaan, *keskimäärin* $100\times\alpha$ % otoksista konstruoiduista luottamusväleistä *ei peitä* parametrin μ todellista arvoa.

Johtopäätökset luottamusvälistä

Oletetaan, että olemme tehneet *johtopäätöksen*, että konstruoitu luottamusväli peittää odotusarvo-parametrin μ todellisen arvon:

(i) Luottamusvälin konstruktiosta seuraa, että tehty johtopäätös *on oikea* $100\times(1-\alpha)$ %:ssa tapauksia.

(ii) Luottamusvälin konstruktiosta seuraa, että tehty johtopäätös *on väärä* $100\times\alpha$ %: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 $2A$. Tarvittava *otoskoko* saadaan kaavasta

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

jossa

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

Normaalijakauman odotusarvon ja varianssin luottamusvälit, kun jakauman varianssi ei ole tunnettu

Otos normaalijakaumasta

Olkoon

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

satunnaisotos normaalijakaumasta N(μ, σ^2). Tällöin satunnaismuuttujat X_i , i = 1, 2, … , n ovat *riippumattomia* ja noudattavat *samaa normaalijakaumaa* N(μ, σ^2):

$$X_1, X_2, \cdots, X_n \perp$$
$$X_i :\ \ \mathrm{N}(\mu, \sigma^2)\ ,\ i = 1, 2, \ldots, n$$

Normaalijakauman parametrien estimointi

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

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

on *odotusarvoparametrin* E(X) = μ harhaton estimaattori ja havaintojen *otosvarianssi*

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

on *varianssiparametrin* Var(X) = σ^2 harhaton estimaattori.

Normaalijakauman odotusarvon luottamusväli, kun jakauman varianssi ei ole tunnettu

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

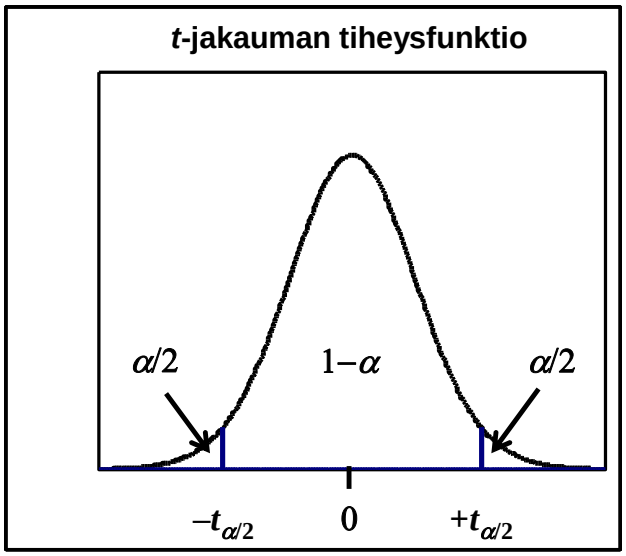
$$\Pr(t \le -t_{\alpha/2}) = \frac{\alpha}{2}$$

ja

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

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

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



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* otoksessa

s^2 = *otosvarianssi*

n = havaintojen *lukumäärä*

$-t_{\alpha/2}$ ja $+t_{\alpha/2}$ = luottamustasoon $(1 - \alpha)$ liittyvät *luottamuskertoimet t-jakaumasta vapausastein* $(n - 1)$

Luottamusvälin konstruktio perustuu siihen, että

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

Koska luottamusväli 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.

Odotusarvon 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*).

Odotusarvon luottamusvälin frekvenssitulkinta

Normaalijakauman odotusarvon μ luottamusvälillä on seuraava *frekvenssitulkinta*:

(i) Jos otantaa jakaumasta $N(\mu, \sigma^2)$ toistetaan, *keskimäärin* $100 \times (1 - \alpha)$ % otoksista konstruoiduista luottamusväleistä *peittää* parametrin μ todellisen arvon.

(ii) Jos otantaa jakaumasta N(μ, σ^2) toistetaan, *keskimäärin* $100\times\alpha$ % otoksista konstruoiduista luottamusväleistä *ei peitä* parametrin μ todellista arvoa.

Johtopäätökset odotusarvon luottamusvälistä

Oletetaan, että olemme tehneet *johtopäätöksen*, että konstruoitu luottamusväli peittää odotusarvo-parametrin μ todellisen arvon:

(i) Luottamusvälin konstruktiosta seuraa, että tehty johtopäätös *on oikea* $100\times(1-\alpha)$ %:ssa tapauksia.

(ii) Luottamusvälin konstruktiosta seuraa, että tehty johtopäätös *on väärä* $100\times\alpha$ %:ssa tapauksia.

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

Vaatimukset odotusarvon 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 $2A$. Tarvittava *otoskoko* saadaan kaavasta

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

jossa

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

Normaalijakauman varianssin luottamusväli

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

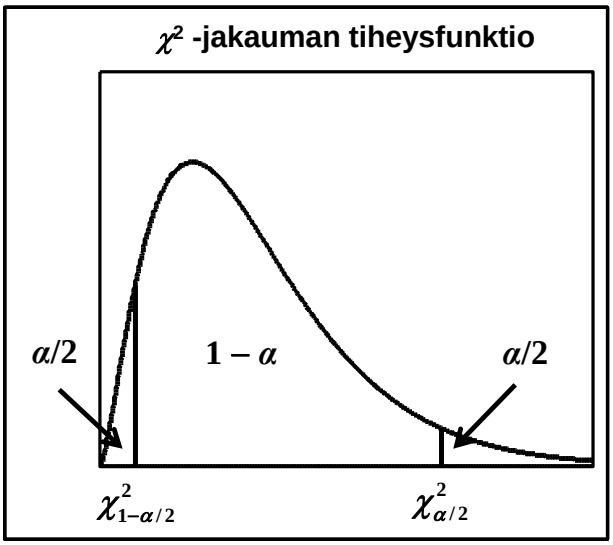
$$\Pr(\chi^2 \le \chi^2_{1-\alpha/2}) = \frac{\alpha}{2}$$

ja

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

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

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



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* $(1 - \alpha)$ 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 $(1 - \alpha)$ liittyvät *luottamuskertoimet* χ^2*-jakaumasta vapausastein* $(n - 1)$

Luottamusvälin konstruktio perustuu siihen, että

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

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

Varianssin 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) Luottamusvälin *lyhenee* (*pitenee*), jos *otoskokoa n kasvatetaan* (*pienennetään*).

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

Varianssin luottamusvälin frekvenssitulkinta

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

(i) Jos otantaa jakaumasta $N(\mu,\sigma^2)$ toistetaan, *keskimäärin* $100\times(1 - \alpha)$ % otoksista konstruoiduista luottamusväleistä *peittää* parametrin σ^2 todellisen arvon.

(ii) Jos otantaa jakaumasta $N(\mu,\sigma^2)$ toistetaan, *keskimäärin* $100\times\alpha$ % otoksista konstruoiduista luottamusväleistä *ei peitä* parametrin σ^2 todellista arvoa.

Johtopäätökset varianssin luottamusvälistä

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

(i) Luottamusvälin konstruktiosta seuraa, että tehty johtopäätös *on oikea* $100\times(1 - \alpha)$ %:ssa tapauksia.

(ii) Luottamusvälin konstruktiosta seuraa, että tehty johtopäätös *on väärä* $100\times\alpha$ %:ssa tapauksia.

Vaatimukset varianssin luottamusvälille

Olisi toivottavaa pystyä konstruoimaan parametrille σ^2 mahdollisimman *lyhyt* luottamusväli, 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*.

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

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):

$$X_1, X_2, \cdots, X_n \perp$$
$$X_i : \ \mathrm{Ber}(p), \ i = 1, 2, \ldots, n$$

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

$$X_i = \begin{cases} 1, \text{ jos } A \text{ tapahtuu} \\ 0, \text{ jos } A \text{ ei tapahdu} \end{cases}, \quad i = 1, 2, \ldots, n$$

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

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

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

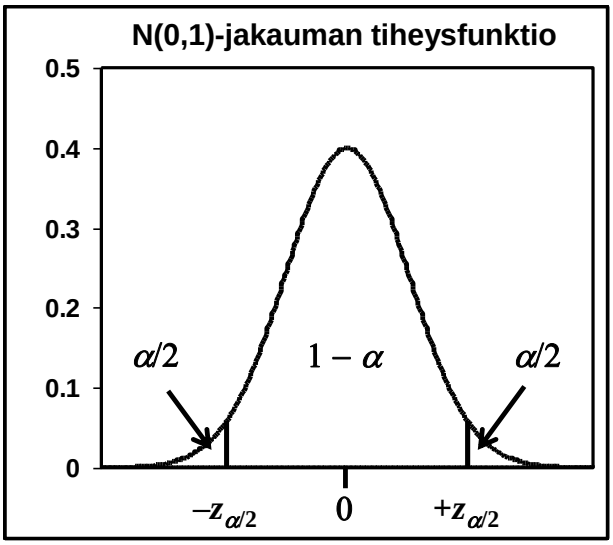
$$\Pr(z \le -z_{\alpha/2}) = \frac{\alpha}{2}$$

ja

$$\Pr(z \ge +z_{\alpha/2}) = \frac{\alpha}{2}$$

jossa satunnaismuuttuja *z* noudattaa *standardoitua normaalijakaumaa*:

$$z: \ \mathrm{N}(0,1)$$



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)

Luottamusvälin konstruktio perustuu siihen, että *keskeisen raja-arvolauseen* mukaan

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} :\ {}_{a}\ \mathrm{N}(0,1)$$

Koska luottamusväli 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

$$\Pr\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) =_a 1-\alpha$$

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

$$\hat{p} = 0.5$$

Luottamusvälin frekvenssitulkinta

Bernoulli-jakauman odotusarvoparametrin *p* approksimatiivisella luottamusvälillä on seuraava *frekvenssitulkinta*:

(i) Jos otantaa jakaumasta Ber(*p*) toistetaan, *keskimäärin* $100\times(1 - \alpha)$ % otoksista konstruoiduista luottamusväleistä *peittää* parametrin *p* todellisen arvon.

(ii) Jos otantaa jakaumasta Ber(*p*) toistetaan, *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* $100\times(1 - \alpha)$ %:ssa tapauksia.

(ii) Luottamusvälin konstruktiosta seuraa, että tehty johtopäätös *on väärä* $100\times\alpha$ %: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

$$p = 0.5$$

Esimerkki 7.1.

Alla on lueteltu joukko tilastollisia muuttujia.

1. Mansikoiden C-vitamiinipitoisuus; yksikkö: mg/100 g
2. Alvarin aukiolta kasvavan kasvin laji
3. Paine, joka vaaditaan teräksisen säiliön murtumiseen; yksikkö: kg/cm^2
4. Suomalaisten reaktiot väitteeseen

   ”Suomen on liityttävä NATO:on”

   mitattuna asteikolla: täysin eri mieltä, yhdentekevää, täysin samaa mieltä
5. Jokereiden sijoitus jääkiekkoliigassa; asteikkona 1, 2, 3, …
6. Teekkarin koulutusohjelma
7. Teekkarin älykkyysosamäärä äo-pisteinä; yksikkö: äo-piste
8. Teekkarin pistemäärä kurssin 1. välikokeesta; asteikkona 0, 1, 2, … , 24
9. Lentokoneen nopeus; yksikkö: km/h

(a) Mitkä ovat muuttujien 1-9 mitta-asteikot?

(b) Mitkä muuttujista 1-9 ovat kvalitatiivisia ja mitkä kvantitatiivisia?

(c) Mitkä muuttujista 1-9 ovat diskreettejä ja mitkä jatkuvia?

Esimerkki 7.1. – Mitä opimme?

Esimerkissä 7.1. tarkastellaan tilastollisten muuttujien *mitta-asteikollisia ominaisuuksia* sekä tilastollisten muuttujien luokittelua toisaalta *kvalitatiivisiin* ja *kvantitatiivisiin* muuttujiin ja toisaalta *diskreetteihin* ja *jatkuviin* muuttujiin.

Esimerkki 7.1. – Ratkaisu:

(a) Laatueroasteikolliset muuttujat: 2, 6

Järjestysasteikolliset muuttujat: 4, 5, 7, 8

Suhdeasteikolliset muuttujat: 1, 3, 9

(b) Kvalitatiiviset muuttujat: 2, (4), (5), 6, (7)

Kvantitatiiviset muuttujat: 1, 3, (4), (5), (7), 8, 9

Kvalitatiivisten ja kvantitatiivisten muuttujien välimaastoon sijoittuvat järjestysasteikolliset muuttujat on merkitty sulkuihin.

(c) Diskreetit muuttujat: 2, 4, 5, 6, 8

Jatkuvat muuttujat: 1, 3, 7, 9

Huomautus:

Muuttujan "oikea" luokittelu ei ole aina helppoa (muuttuja 7).

Esimerkki 7.2.

Erään talon asukkailla on seuraavat tulot (€/vuosi):

20100	19400	10100	23000	24200	25100	8200	8900	10300
26000	11400	12900	13200	14300	15800	16100	17200	18900
5200	10100	12300	14000	15100	16000	11100	10800	9100
7200	4300	38000	51100	9600	10900	12000	13200	15100

Määrää aineistosta seuraavat tunnusluvut:

(a) minimi, maksimi

(b) vaihteluväli, vaihteluvälin pituus

(c) mediaani

Esimerkki 7.2. – Mitä opimme?

Esimerkissä 7.2. tarkastellaan *järjestystunnuslukujen* määräämistä.

Esimerkki 7.2. – Ratkaisu:

Kaikki määrättäviksi pyydetyt tunnusluvut ovat *järjestystunnuslukuja* tai niihin perustuvia tunnuslukuja.

Järjestetään havaintoarvot suuruusjärjestykseen pienimmästä suurimpaan järjestys-tunnuslukujen määräämistä varten:

4300	5200	7200	8200	8900	9100	9600	10100	10100
10300	10800	10900	11100	11400	12000	12300	12900	13200
13200	14000	14300	15100	15100	15800	16000	16100	17200
18900	19400	20100	23000	24200	25100	26000	38000	51100

(a) *Minimi* ja *maksimi*:

$Min = 4300$, $Max = 51100$

(b) *Vaihteluväli*:

$(Min, Max) = (4300, 51100)$

Vaihteluvälin pituus:

$Max - Min = 51100 - 4300 = 46800$

(c) Etsitään havaintoarvojen *mediaani Me*.

Mediaani *Me* jakaa havaintoaineiston kahteen yhtä suureen osaan siten, että puolet niistä havaintoarvoista, jotka eivät ole yhtä suuria kuin mediaani, ovat mediaania pienempiä, ja puolet niistä havaintoarvoista, jotka eivät ole yhtä suuria kuin mediaani, ovat mediaania suurempia.

Oletetaan, että n havaintoarvoa on järjestetty suuruusjärjestykseen pienimmästä suurimpaan.

(i) Jos n on pariton, niin mediaaniksi valitaan havaintoarvo, joka löytyy paikasta

$$(n + 1)/2$$

(ii) Jos n on parillinen, mediaaniksi valitaan kahden keskimmäisen havainnon aritmeettinen keskiarvo.

Koska havaintojen lukumäärä on tässä *parillinen*, niin

$$Me = (13200 + 13200)/2 = 13200$$

Esimerkki 7.3.

Muodosta esimerkin 7.2. aineistosta luokiteltu frekvenssijakauma, jonka luokkaväleinä ovat

[4000, 12000]

(12000, 28000]

(28000, 60000]

Määrää myös frekvenssijakaumaa vastaavan histogrammin suorakaiteiden korkeudet, kun luokkaväliä [4000, 12000] vastaavan suorakaiteen korkeudeksi valitaan 15 yksikköä. Hahmottele histogrammi ruudulliselle paperille. Missä luokassa on jakauman moodi?

Esimerkki 7.3. – Mitä opimme?

Esimerkissä 7.3. tarkastellaan jatkuvan muuttujan havaittujen arvojen *luokitellun frekvenssijakauman* ja sitä vastaavan graafisen esityksen eli *histogrammin* muodostamista.

Esimerkki 7.3. – Ratkaisu:

Jatkuvan muuttujan havaittujen arvojen jakaumaa kuvataan *luokitellulla frekvenssi-jakaumalla*. Luokiteltua frekvenssijakaumaa voidaan kuvata graafisesti *histogrammilla*. Histogrammi koostuu luokitellun frekvenssijakauman luokkia vastaavista *suorakaiteista* (*nelikulmioista*), *joiden pinta-alat ovat suhteessa vastaaviin luokkafrekvensseihin*.

Tehtävän 7.2. aineistosta saadaan seuraava luokiteltu frekvenssijakauma, kun luokkaväleinä ovat [4000, 12000], (12000, 28000], (28000, 60000] :

Luokkaväli	Luokkafrekvenssi	Suorakaiteen korkeus (yksikköä)
[4000, 12000]	15	15
(12000, 28000]	19	19/2 = 9.5
(28000, 60000]	2	2/4 = 0.5

Histogrammikuvion suorakaiteiden korkeuksien määrääminen:

(1) Valitaan luokkaväliin [4000, 12000] liittyvän suorakaiteen korkeudeksi 15 yksikköä.

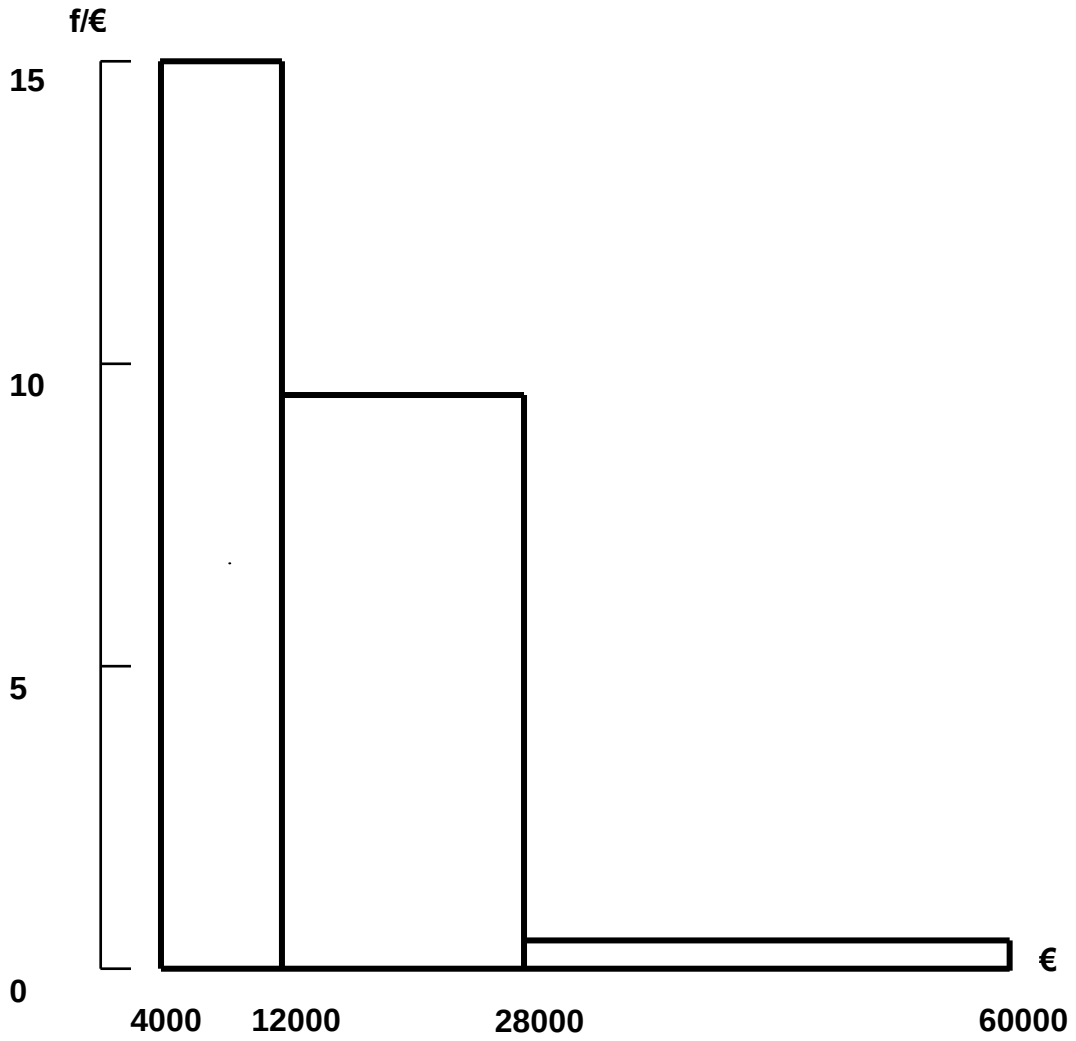
(2) Luokkaväli (12000, 28000] on kaksi kertaa niin pitkä kuin luokkaväli (4000, 12000].

Siksi luokkaväliin (12000, 28000] liittyvän suorakaiteen korkeus saadaan jakamalla luokkaväliä vastaava frekvenssi 19 luvulla 2.

(3) Luokkaväli (28000, 60000] on neljä kertaa niin pitkä kuin luokkaväli (4000, 12000].

Siksi luokkaväliin (18000, 60000] liittyvän suorakaiteen korkeus saadaan jakamalla luokkaväliä vastaava frekvenssi 2 luvulla 4.

Alla oleva kuvio esittää yo. luokiteltua frekvenssijakaumaa vastaavaa histogrammia.



Jakauman *moodi* on luokassa (4000, 12000], koska siinä histogrammi saavuttaa maksiminsa. Huomaa, että moodi *ei ole* luokassa (12000, 28000], vaikka sitä vastaava frekvenssi on suurin.

Huomautuksia:

(i) Histogrammin *suorakaiteiden* (*nelikulmioiden*) *pinta-alat – eivät siis korkeudet – ovat suhteessa luokkafrekvensseihin*.

(ii) Histogrammissa suorakaiteiden korkeudet ovat suhteessa luokkafrekvensseihin *vain, jos luokitus on tasavälinen*.

(iii) Oikea *laatu* pystyakselille on tehtävän 8.3. tapauksessa frekvenssi/€ :

Vaaka-akselin laatu:	€
Pystyakselin laatu:	frekvenssi/€
Tällöin suorakaiteen pinta-ala:	€ × frekvenssi/€ = frekvenssi

Esimerkki 7.4.

Määrää esimerkin 7.2. aineiston kahden ensimmäisen sarakkeen 8:sta luvusta aritmeettinen keskiarvo, otosvarianssi ja otoshajonta.

Esimerkki 7.4. – Mitä opimme?

Esimerkissä 7.4. tarkastellaan havaintoarvojen *aritmeettisen keskiarvon, otosvarianssin* ja *otoshajonnan* määräämistä.

Esimerkki 7.4. – Ratkaisu:

Laskutoimitukset voidaan suorittaa kahdella tavalla.

Tapa 1:

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

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

$$s_x = \sqrt{s_x^2}$$

Tapa 2:

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

$$s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)$$

$$s_x = \sqrt{s_x^2}$$

Jos havaintoarvojen aritmeettisen keskiarvon ja varianssi laskemista varten laaditaan tietokoneohjelma, laskutoimitukset voidaan järjestää laskutavassa 2 niin, että havainnot käydään läpi vain kerran, kun taas laskutavassa 1 havainnot on käytävä läpi kaksi kertaa. Sen sijaan laskutavan 1 kaavat ovat numeerisesti vakaampia kuin laskutavassa 2.

Tehtävän lopussa on kopio laskutoimitusten tekemisessä apuna käytetystä Microsoft Excel -taulukosta.

Taulukosta saadaan:

$$\sum_{i=1}^{8} x_i = 103700$$

$$\overline{x} = \frac{1}{8}\sum_{i=1}^{8} x_i = 12962.5$$

$$\sum_{i=1}^{8} x_i^2 = 1785710000$$

$$s_x^2 = \frac{1}{8-1}\left(\sum_{i=1}^{8} x_i^2 - \frac{1}{8}\left(\sum_{i=1}^{8} x_i \right)^2 \right) = 63071250$$

$$\sum_{i=1}^{8} (x_i - \overline{x})^2 = 441498750$$

$$s_x^2 = \frac{1}{8-1}\sum_{i=1}^{8} (x_i - \overline{x})^2 = 63071250$$

$$s_x = \sqrt{s_x^2} = 7941.741$$

	Palkka			
i	**x**	**x-Ka**	**(x-Ka)^2**	**x^2**
1	20100	7137.5	50943906.25	404010000
2	26000	13037.5	169976406.3	676000000
3	5200	-7762.5	60256406.25	27040000
4	7200	-5762.5	33206406.25	51840000
5	19400	6437.5	41441406.25	376360000
6	11400	-1562.5	2441406.25	129960000
7	10100	-2862.5	8193906.25	102010000
8	4300	-8662.5	75038906.25	18490000
Summa	**103700**	**0**	**441498750**	**1785710000**

Ka = 12962.5

Tapa 1:

Var = 63071250

Hajonta = 7941.741

Tapa 2:

Var = 63071250

Hajonta = 7941.741

Esimerkki 7.5.

Olet ottanut pankista 10000 euron lainan, jota ei saa lyhentää kahden ensimmäisen vuoden aikana. Alkuperäisen sopimuksen mukainen korko on 1. vuotena 10 % ja 2. vuotena 20 %, jolloin takaisin maksettava lainapääoma kasvaa kahdessa vuodessa x %.

Oletetaan, että pankin vaatimuksesta sopimusta muutetaan niin, että kahden ensimmäisen vuoden aikana käytetään *samaa* korkoprosenttia, joka määrätään niin, että lainapääoma kasvaa tänä aikana kuitenkin samaksi kuin alkuperäisen sopimuksen mukaan.

(a) Määrää x.

(b) Näytä, että uuden sopimuksen korkoa *ei saada* kaavalla

$$(10 + 20)/2\ \%$$

(c) Näytä, että uuden sopimuksen korko *saadaan* kaavalla

$$(\sqrt{1.1\times 1.2}-1)\times 100\ \%$$

jossa

$$\sqrt{1.1\times 1.2}$$

on lukujen 1.1 ja 1.2 geometrinen keskiarvo.

Esimerkki 7.5. – Mitä opimme?

Esimerkissä 7.5. näytetään, että *aritmeettinen keskiarvo* ei ole joka tilanteessa käypä tunnusluku. Oikea keskiluku tehtävän ongelman ratkaisemiseen on *geometrinen keskiarvo*.

Esimerkki 7.5. – Ratkaisu:

(a) Olkoon korko 1. vuotena 10 % ja 2. vuotena 20 %.

Lainapääoma 1. vuoden lopussa:

$$(1 + 10/100)\times 10000 = (1 + 0.1)\times 10000 = 1.1\times 10000 = 11000$$

Lainapääoma 2. vuoden lopussa:

$$(1 + 20/100)\times 11000 = (1 + 0.2)\times 11000 = 1.2\times 11000 = 13200$$

Siten lainapääoma kasvaa kahdessa vuodessa

$$100\times(13200 - 10000)/10000\ \% = 32\ \%$$

joten

$$x = 32$$

(b) Määrätään 1. ja 2. vuoden korkoprosenttien aritmeettinen keskiarvo:

$$\frac{10+20}{2}\ \% = 15\ \%$$

Olkoon korkona molempina vuosina siis 15 %.

Lainapääoma 1. vuoden lopussa:

$(1 + 15/100)\times 10000 = (1 + 0.15)\times 10000 = 1.15\times 10000 = 11500$

Lainapääoma 2. vuoden lopussa:

$(1 + 15/100)\times 11500 = (1 + 0.15)\times 11500 = 1.15\times 11500 = 13225$

Tällöin lainapääoma kasvaisi siis kahdessa vuodessa

$100\times(13225 - 10000)/10000\ \% = 32.25\ \% > 32\ \%$

Huomaa, että oikea korkoprosentti ei myöskään ole

$32/2\ \% = 16\ \%$

(c) Määrätään korkoprosentti kaavalla

$(\sqrt{1.1\times 1.2} - 1)\times 100\ \%$

jossa

$\sqrt{1.1\times 1.2}$

on lukujen 1.1 ja 1.2 geometrinen keskiarvo:

$(\sqrt{1.1\times 1.2} - 1)\times 100 \approx 14.8913$

Olkoon korko siis molempina vuosina 14.8913 %.

Lainapääoma 1. vuoden lopussa:

$(1 + 14.8913/100)\times 10000 = (1 + 0.148913)\times 10000 = 1.148913\times 10000$
$= 11489.13$

Lainapääoma 2. vuoden lopussa:

$(1 + 14.8913/100)\times 11489.13 = (1 + 0.148913)\times 11489.13 = 1.148913\times 11489.13$
≈ 13200

Siten lainapääoma kasvaa kahdessa vuodessa

$100\times(13200 - 10000)/10000\ \% = 32\ \%$

kuten pitääkin.

Huomautus:

Olkoon korko 1. vuotena p % ja toisena vuotena q %.

Yleisesti pätee:

$$\left[1+\left(\sqrt{\left(1+\frac{p}{100}\right)\left(1+\frac{q}{100}\right)}-1\right)\right]^2 = \left(1+\frac{p}{100}\right)\left(1+\frac{q}{100}\right)$$

Sen sijaan

$$\left(1+\frac{(p+q)/2}{100}\right)\left(1+\frac{(p+q)/2}{100}\right)\neq\left(1+\frac{p}{100}\right)\left(1+\frac{q}{100}\right)$$

paitsi, jos

$p = q$

Esimerkki 7.6.

Paikkakuntien A ja B välimatka on 120 km. Henkilö ajaa A:sta B:hen keskinopeudella 60 km/h ja B:stä A:han keskinopeudella 120 km/h.

(a) Määrää keskinopeus edestakaisella matkalla.

(b) Näytä, että keskinopeutta edestakaisella matkalla *ei saada* kaavalla

(60 + 120)/2 = 90 km/h

(c) Näytä, että oikea keskinopeus *saadaan* määräämällä lukujen 60 ja 100 harmoninen keskiarvo

$$\frac{1}{\frac{1}{2}\left(\frac{1}{60}+\frac{1}{120}\right)}$$

Esimerkki 7.6. – Mitä opimme?

Esimerkissä 7.6. näytetään, että *aritmeettinen keskiarvo* ei ole joka tilanteessa käypä tunnusluku. Oikea keskiluku tehtävän ongelman ratkaisemiseen on *harmoninen keskiarvo*.

Esimerkki 7.6. – Ratkaisu:

(a) A:n ja B:n välimatka:

120 km

Ajoaika A:sta B:hen (60 km/h):

120/60 = 2 h

Ajoaika B:stä A:han (120 km/h):

120/120 = 1 h

Matka edestakaisin:

240 km

Ajoaika edestakaisin:

2 + 1 = 3 h

Keskinopeus edestakaisella matkalla:

240/3 = 80 km/h

(b) Määrätään keskinopeuksien *aritmeettinen keskiarvo*:

$$\frac{1}{2}(60+120)\ \text{km/h} = 90\ \text{km/h} \neq 80\ \text{km/h}$$

(c) Määrätään keskinopeuksien *harmoninen keskiarvo*:

$$\frac{1}{\frac{1}{2}\left(\frac{1}{60}+\frac{1}{120}\right)}\ \text{km/h} = 80\ \text{km/h}$$

Esimerkki 7.7.

Kone valmistaa kuulalaakerin kuulia, joiden halkaisijat vaihtelevat satunnaisesti noudattaen normaalijakaumaa parametrein

$$\mu = 10\ \text{mm}\ ,\ \sigma = 0.01\ \text{mm}$$

Poimitaan kuulien joukosta yksinkertainen satunnaisotos, jonka koko $n = 10$. Olkoot $\bar{X}$ ja s^2 kuulien halkaisijoiden aritmeettinen keskiarvo ja otosvarianssi otoksessa. Mitkä ovat aritmeettisen keskiarvon $\bar{X}$ ja otosvarianssin muunnoksen $(n-1)s^2/\sigma^2$ otosjakaumat?

Esimerkki 7.7. – Mitä opimme?

Esimerkissä 7.7. tarkastellaan *aritmeettisen keskiarvon* ja *otosvarianssin otosjakaumia*.

Esimerkki 7.7. – Ratkaisu:

Oletuksen mukaan havainnot $X_1, X_2, \ldots, X_n$ muodostavat *yksinkertaisen satunnaisotoksen* normaalijakaumasta $\mathrm{N}(\mu,\sigma^2)$, jossa

$$n = 10$$

$$\mu = 10\ \text{mm}$$

$$\sigma^2 = 0.01^2\ \text{mm} = 0.0001\ \text{mm}^2$$

Siten kuulien halkaisijoiden aritmeettinen keskiarvo $\bar{X}$ noudattaa otoksessa *normaali-jakaumaa* $\mathrm{N}(\mu,\sigma^2/n)$, jossa

$$\mu = \mathrm{E}(\bar{X}) = 10\ \text{mm}$$

$$\frac{\sigma^2}{n} = \mathrm{Var}(\bar{X}) = \mathrm{D}^2(\bar{X}) = \frac{0.0001}{10} = 0.00001\ \text{mm}^2$$

Olkoon s^2 kuulien halkaisijoiden varianssi otoksessa. Tällöin satunnaismuuttuja $(n-1)s^2/\sigma^2$ noudattaa otoksessa χ^2*-jakaumaa* vapausastein

$$n - 1 = 10 - 1 = 9$$

Esimerkki 7.8.

Äänestäjistä 25 % kannattaa puoluetta ABC. Poimitaan äänestäjien joukosta yksinkertainen satunnaisotos, jonka koko $n = 1000$. Määrää puolueen ABC kannattajien suhteellisen osuuden (approksimatiivinen) otosjakauma.

Esimerkki 7.8. – Mitä opimme?

Esimerkissä 7.8. tarkastellaan *suhteellisen osuuden* (approksimatiivista) *otosjakaumaa*.

Esimerkki 7.8. – Ratkaisu:

Olkoon

A = satunnaisesti valittu äänestäjä kannattaa puoluetta ABC

Oletuksen mukaan

$$\Pr(A) = p = 0.25$$

Poimitaan äänestäjien joukosta *yksinkertainen satunnaisotos*, jonka koko on $n = 1000$.

Puoluetta ABC kannattavien äänestäjien suhteellinen frekvenssi $\hat{p} = f / n$ otoksessa noudattaa *suurissa otoksissa approksimatiivisesti normaalijakaumaa*:

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

jossa siis

$$p = \Pr(A) = 0.25$$

$$q = \Pr(A^C) = 1 - \Pr(A) = 1 - p = 0.75$$

Siten puolueen ABC kannattajien suhteellinen frekvenssi $\hat{p} = f / n$ otoksessa noudattaa *suurissa otoksissa approksimatiivisesti normaalijakaumaa* parametrein

$$\mathrm{E}(\hat{p}) = p = 0.25$$

$$\mathrm{Var}(\hat{p}) = \mathrm{D}^2(\hat{p}) = \frac{pq}{n} = \frac{0.25 \times 0.75}{1000} = \frac{0.1875}{1000} = 0.0001875$$

Huomautuksia esimerkkeihin 7.7. ja 7.8.:

(1) Esimerkkien 7.7. ja 7.8. ideana on näyttää, *millaisia ovat tavanomaisten havainnoista laskettavien otostunnuslukujen jakaumat perusjoukossa*, *jos havaintojen jakauma perusjoukossa tunnetaan*.

(2) Otostunnuslukujen jakaumia koskevat tulokset ovat kuitenkin *epäoperationaalisia*, koska *havaintojen jakauman parametrit ovat kaikissa tavanomaisissa sovellustilanteissa tuntemattomia*.

(3) Jos havaintojen jakauman parametreja ei tunneta, ne voidaan pyrkiä *estimoimaan* eli arvioimaan otoksesta saatujen tietojen perusteella; ks. lukua **Tilastollisten mallien parametrien estimointi**.

(4) Perusjoukon parametrien arvoista tehtyjä oletuksia voidaan pyrkiä *testaamaan tilastollisesti* otoksesta saatujen tietojen perusteella; ks. lukua **Tilastollisten hypoteesien testaus**.

(5) Myös havaintojen jakauman tyyppiä koskevia oletuksia voidaan pyrkiä *testaamaan tilastollisesti* otoksesta saatujen tietojen perusteella; ks. lukua **Yhteensopivuuden, homogeenisuuden ja riippumattomuuden testaaminen**.

Esimerkki 7.11.

Oletetaan, että suomalaisten miesten pituus on normaalijakautunut parametrein μ = 175 cm ja σ = 5 cm. Poimitaan miesten joukosta yksinkertainen satunnaisotos, jonka koko on 100. Määrää lukuarvo, jota *suurempia* arvoja havaintojen aritmeettinen keskiarvoa saa todennäköisyydellä 0.01.

Esimerkki 7.11. – Mitä opimme?

Esimerkissä 7.11. tarkastellaan *aritmeettisen keskiarvon otosjakaumaa*.

Esimerkki 7.11. – Ratkaisu:

Oletetaan, että havainnot

$$X_i \text{ , } i = 1, 2, \ldots , 100$$

muodostavat yksinkertaisen satunnaisotoksen normaalijakaumasta N(μ,σ^2), jossa

$$\mu = 175$$

$$\sigma^2 = 25$$

Olkoon havaintojen X_i , i = 1, 2, … , 100 *aritmeettinen keskiarvo*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{100}\sum_{i=1}^{100} X_i$$

Oletuksista seuraa, että satunnaismuuttuja $\bar{X}$ *noudattaa normaalijakaumaa* parametrein μ ja σ^2/n :

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

jossa siis

$$\mu = 175$$

$$\frac{\sigma^2}{n} = \frac{25}{100} = \frac{1}{4} = 0.25$$

Tehtävänä on määrätä lukuarvo, jota *suurempia* arvoja havaintojen aritmeettinen keskiarvo $\bar{X}$ saa todennäköisyydellä 0.01.

Koska

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

niin *standardoitu satunnaismuuttuja*

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

noudattaa *standardoitua normaalijakaumaa*:

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

Normaalijakauman taulukoista näemme, että

$$\Pr(Z \le 2.33) = 0.9901 \approx 0.99$$

Komplementtitapahtuman todennäköisyyden kaavan mukaan

$$\Pr(Z > 2.33) = 1 - \Pr(Z \le 2.33) \approx 1 - 0.99 = 0.01$$

Saamme siten epäyhtälön

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} > 2.33$$

josta aritmeettiselle keskiarvolle saadaan ehto

$$\bar{X} > \mu + \frac{\sigma}{\sqrt{n}} \times 2.33 = 175 + \frac{5}{\sqrt{100}} \times 2.33 = 176.165$$

Siten

$$\Pr(\bar{X} \ge 176.165) = 0.01$$

Esimerkki 7.12.

Oletetaan, että havainnot X_i , $i = 1, 2, \ldots, 101$ muodostavat yksinkertaisen satunnaisotoksen normaalijakaumasta N(1,4). Määrää lukuarvo, jota *pienempiä* arvoja havaintojen otos-varianssi saa todennäköisyydellä 0.01.

Esimerkki 7.12. – Mitä opimme?

Esimerkissä 7.12. tarkastellaan *otosvarianssin otosjakaumaa*.

Esimerkki 7.12. – Ratkaisu:

Oletetaan, että havainnot

X_i , $i = 1, 2, \ldots, 101$

muodostavat yksinkertaisen satunnaisotoksen normaalijakaumasta N(μ,σ^2), jossa

$\mu = 1$

$\sigma^2 = 4$

Olkoon havaintojen X_i , $i = 1, 2, \ldots, 101$ *aritmeettinen keskiarvo*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{101}\sum_{i=1}^{101} X_i$$

ja *otosvarianssi*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{100}\sum_{i=1}^{101}(X_i - \bar{X})^2$$

Oletuksista seuraa, että satunnaismuuttuja

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

jossa

$$\sigma^2 = 4$$

$$n = 101$$

noudattaa χ^2-jakaumaa vapausastein $(n - 1)$:

$$V: \ \chi^2(100)$$

Tehtävänä on määrätä lukuarvo, joka erottaa χ^2-jakauman *vasemmalle hännälle* todennäköisyysmassan, jonka koko on 0.01.

χ^2-jakauman taulukoista nähdään suoraan, että

$$\Pr(V \leq 70.065) = 0.01$$

kun

$$V: \ \chi^2(100)$$

Koska

$$V = \frac{(n-1)s^2}{\sigma^2} = \frac{100s^2}{4} = 25s^2$$

saamme epäyhtälön

$$25s^2 \leq 70.065$$

josta otosvarianssille s^2 saadaan ehto

$$s^2 \leq 2.803$$

Siten

$$\Pr(s^2 \leq 2.803) = 0.01$$

Esimerkki 7.13.

Oletetaan, että suomalaisten miesten pituus on normaalijakautunut parametrein $\mu = 175$ cm ja $\sigma = 5$ cm. Poimitaan miesten joukosta yksinkertainen satunnaisotos, jonka koko on 101. Määrää lukuarvo, jota *suurempia* arvoja otosvarianssi saa todennäköisyydellä 0.01.

Esimerkki 7.13. – Mitä opimme?

Esimerkissä 7.13. tarkastellaan *otosvarianssin otosjakaumaa*.

Esimerkki 7.13. – Ratkaisu:

Oletetaan, että havainnot

X_i , $i = 1, 2, \ldots , 101$

muodostavat *yksinkertaisen satunnaisotoksen* normaalijakaumasta $N(\mu,\sigma^2)$, jossa

$\mu = 175$

$\sigma^2 = 25$

Olkoon havaintojen X_i, $i = 1, 2, \ldots , 101$ *aritmeettinen keskiarvo*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{101}\sum_{i=1}^{101} X_i$$

ja *otosvarianssi*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{100}\sum_{i=1}^{101}(X_i - \bar{X})^2$$

Oletuksista seuraa, että satunnaismuuttuja

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

jossa

$\sigma^2 = 25$

$n = 101$

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

$V : \ \chi^2(100)$

Tehtävänä on määrätä lukuarvo, joka erottaa χ^2-jakauman *oikealle hännälle* todennäköisyysmassan, jonka koko on 0.01:

χ^2-jakauman taulukoista nähdään suoraan, että

$\Pr(V \geq 135.807) = 0.01$

kun

$V : \ \chi^2(100)$

Koska

$$V = \frac{(n-1)s^2}{\sigma^2} = \frac{100s^2}{25} = 4s^2$$

saamme epäyhtälön

$4s^2 \geq 135.807$

josta otosvarianssille s^2 saadaan ehto

$s^2 \geq 33.9518$

Siten

$\Pr(s^2 \geq 33.9518) = 0.01$

Esimerkki 7.14.

Oletetaan, että teemme 100 toisistaan riippumatonta Bernoulli-koetta, jossa kiinnostuksen kohteena olevan tapahtuman *A* todennäköisyys on 0.2. Määrää todennäköisyys, että tapahtuman *A* frekvenssi toistojen joukossa on *suurempi* kuin 10.

Esimerkki 7.14. – Mitä opimme?

Esimerkissä 7.14. tarkastellaan *suhteellisen frekvenssin* (approksimatiivista) *otos-jakaumaa*.

Esimerkki 7.14. – Ratkaisu:

Olkoon

f = tapahtuman *A frekvenssi* toistojen joukossa

$\hat{p} = \frac{f}{n}$ = tapahtumien *A suhteellinen frekvenssi* toistojen joukossa

n = toistojen lukumäärä

Koska toistojen lukumäärä $n = 100$ on näinkin suuri, voimme melko hyvin approksimoida suhteellisen frekvenssin $\hat{p}$ otantajakaumaa normaalijakaumalla:

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

jossa

$p = 0.2$

$q = 1 - p = 0.8$

$n = 100$

Oletuksista seuraa, että *standardoitu satunnaismuuttuja*

$$Z = \frac{\hat{p} - \mathrm{E}(\hat{p})}{\mathrm{D}(\hat{p})} = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

noudattaa *approksimatiivisesti standardoitua normaalijakaumaa*:

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

Koska

$$\frac{10}{100} = \frac{1}{10} = 0.1$$

tehtävänä on määrätä todennäköisyys

$$\Pr\left(\hat{p} > 0.1 \right)$$

Selvästi

$$\Pr(\hat{p} > 0.1) = \Pr\left(\frac{\hat{p} - p}{\sqrt{pq/n}} > \frac{0.1 - p}{\sqrt{pq/n}} \right) = \Pr\left(Z > \frac{0.1 - 0.2}{\sqrt{0.2 \times 0.8/100}} \right) = \Pr\left(Z > -2.5 \right)$$

jossa siis

$$Z = \frac{\hat{p} - \mathrm{E}(\hat{p})}{\mathrm{D}(\hat{p})} = \frac{\hat{p} - p}{\sqrt{pq/n}} :\ {}_{a}\ \mathrm{N}(0,1)$$

Normaalijakauman taulukoiden mukaan

$$\Pr(Z \le -2.5) = 0.0062$$

joten komplementtitapahtuman todennäköisyyden kaavan mukaan kysytty todennäköisyys on

$$\Pr(\hat{p} > 0.1) = \Pr(Z > -2.5) = 1 - \Pr(Z \le -2.5) = 1 - 0.0062 = 0.9938$$

Esimerkki 8.2.

Olkoot X_i, $i = 1, 2, \ldots, n$ riippumattomia, samaa eksponenttijakaumaa noudattavia satunnaismuuttujia, joiden odotusarvo $\mathrm{E}(X_i) = \beta$, ts. satunnaismuuttujat X_i, $i = 1, 2, \ldots, n$ muodostavat yksinkertaisen satunnaisotoksen eksponenttijakaumasta, jonka parametrina on $1/\beta$. Määrää parametrin β *suurimman uskottavuuden estimaattori*.

Esimerkki 8.2. – Mitä opimme?

Esimerkissä 8.2. tarkastellaan eksponenttijakauman parametrin *suurimman uskottavuuden estimointia*.

Esimerkki 8.2. – Ratkaisu:

Oletetaan, että X_i , $i = 1, 2, \ldots, n$ on *yksinkertainen satunnaisotos eksponenttijakaumasta*, jonka parametrina on $1/\beta$. Siten

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

$$X_i \sim \mathrm{Exp}(1/\beta)\ ,\ i = 1, 2, \ldots, n$$

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

$$L(\beta; x_1, x_2, \ldots, x_n) = f(x_1;\beta) f(x_2;\beta) \cdots f(x_n;\beta) = \frac{1}{\beta^n} \exp\left(-\frac{1}{\beta} \sum_{i=1}^{n} x_i \right)$$

jossa

$$f(x_i;\beta) = \frac{1}{\beta} \exp\left(-\frac{1}{\beta} x_i \right),\ \ i = 1, 2, \ldots, n$$

on havainnon X_i tiheysfunktio. Vastaava *logaritminen uskottavuusfunktio* on

$$l(\beta; x_1, x_2, \ldots, x_n) = \log L(\beta; x_1, x_2, \ldots, x_n) = -\frac{1}{\beta} \sum_{i=1}^{n} x_i - n \log(\beta)$$

Suurimman uskottavuuden estimaattori parametrille β löydetään maksimoimalla logaritminen uskottavuusfunktio l parametrin β suhteen.

Tämä tapahtuu derivoimalla logaritminen uskottavuusfunktio l parametrin β suhteen, merkitsemällä derivaatta nollaksi ja *ratkaisemalla saatu normaaliyhtälö* parametrin β suhteen:

$$\frac{\partial\, l(\beta; x_1, x_2, \ldots, x_n)}{\partial \beta} = \frac{1}{\beta^2} \sum_{i=1}^{n} x_i - n \frac{1}{\beta} = 0$$

Ratkaisuksi saadaan

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

Saatu ratkaisu antaa logaritmisen uskottavuusfunktion maksimin, mikä nähdään esim. sijoittamalla saatu ratkaisu logaritmisen uskottavuusfunktion 2. derivaatan lausekkeeseen.

Esimerkki 8.4.

Tehdas väittää, että normaalitilanteessa sen valmistamista tuotteista 5 % on viallisia. Asiakas poimii tuotteiden joukosta yksinkertaisen satunnaisotoksen, jonka koko on 150 ja löytää 15 viallista tuotetta. Voidaanko tehtaan väitettä viallisten suhteellisesta osuudesta pitää oikeutettuna?

Ohje: Määrää otoksesta 95 %:n ja 99 %:n *luottamusvälit* viallisten todelliselle suhteelliselle osuudelle ja tee johtopäätökset niiden perusteella.

Lisäkysymys: Miten valittu luottamustaso vaikuttaa luottamusvälin pituuteen?

Esimerkki 8.4. – Mitä opimme?

Esimerkissä 8.4. tarkastellaan *Bernoulli-jakauman odotusarvon luottamusvälin* määräämistä.

Esimerkki 8.4. – Ratkaisu:

Olkoon tapahtuma

A = {Satunnaisesti valittu tuote on viallinen}

ja olkoon tapahtuman A todennäköisyys

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

Määritellään satunnaismuuttuja

$$X = \begin{cases} 1, \text{ jos satunnaisesti valittu tuote on viallinen} \\ 0, \text{ jos satunnaisesti valittu tuote ei ole viallinen} \end{cases}$$

Tällöin satunnaismuuttuja X noudattaa *Bernoulli-jakaumaa* parametrinaan

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

Valmistettujen tuotteiden joukosta poimittiin yksinkertainen satunnaisotos, jonka koko oli

$$n = 150$$

ja otoksessa havaittiin 15 viallista tuotetta.

Konstruoidaan otoksesta saatujen tietojen perusteella $(1 - \alpha)$ %:n *luottamusväli* odotusarvo-parametrille p. Luottamusväli on muotoa

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

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)

Parametrin *p estimaatiksi* saadaan

$$\hat{p} = \frac{f}{n} = \frac{15}{150} = 0.1$$

Valitaan *luottamustasoksi*

$$1 - \alpha = 0.95$$

Koska

$$\alpha = 0.05$$

luottamustasoa 0.95 vastaavat *luottamuskertoimet* ovat

$$-z_{\alpha/2} = -z_{0.025}$$
$$+z_{\alpha/2} = +z_{0.025}$$

Luottamuskertoimet $-z_{\alpha/2} = -z_{0.025}$ ja $+z_{\alpha/2} = +z_{0.025}$ toteuttavat yhtälöt

$$\Pr(z \le -z_{\alpha/2}) = \Pr(z \le -z_{0.025}) = \frac{\alpha}{2} = 0.025$$
$$\Pr(z \ge +z_{\alpha/2}) = \Pr(z \ge -z_{0.025}) = \frac{\alpha}{2} = 0.025$$

jossa satunnaismuuttuja *z* noudattaa *standardoitua normaalijakaumaa*:

$$z: \ \mathrm{N}(0,1)$$

Siten

$$\Pr(-z_{\alpha/2} \le z \le +z_{\alpha/2}) = \Pr(-z_{0.025} \le z \le +z_{0.025}) = 1 - \alpha = 0.95$$

Standardoidun normaalijakauman N(0,1) taulukoiden mukaan

$$-z_{0.025} = -1.96$$
$$+z_{0.025} = +1.96$$

Siten 95 %:n *luottamusväli* Bernoulli-jakauman parametrille *p* on muotoa

$$\begin{aligned}\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= 0.1 \pm 1.96\sqrt{\frac{0.1\times(1-0.1)}{150}} \\ &= 0.1 \pm 1.96 \times 0.024 \\ &= 0.1 \pm 0.048 \\ &= (0.052, 0.148)\end{aligned}$$

Väli *ei peitä* parametrin *p* oletettua arvoa 0.05.

Valitaan *luottamustasoksi*

$$1 - \alpha = 0.99$$

Koska

$$\alpha = 0.01$$

luottamustasoa $1 - \alpha = 0.99$ vastaavat *luottamuskertoimet* ovat

$$-z_{\alpha/2} = -z_{0.005}$$
$$+z_{\alpha/2} = +z_{0.005}$$

Luottamuskertoimet $-z_{\alpha/2} = -z_{0.005}$ ja $+z_{\alpha/2} = +z_{0.005}$ toteuttavat yhtälöt

$$\Pr(z \le -z_{\alpha/2}) = \Pr(z \le -z_{0.005}) = \frac{\alpha}{2} = 0.005$$
$$\Pr(z \ge +z_{\alpha/2}) = \Pr(z \ge +z_{0.005}) = \frac{\alpha}{2} = 0.005$$

jossa satunnaismuuttuja *z* noudattaa *standardoitua normaalijakaumaa*:

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

Siten

$$\Pr(-z_{\alpha/2} \le z \le +z_{\alpha/2}) = \Pr(-z_{0.005} \le z \le +z_{0.005}) = 1 - \alpha = 0.99$$

Standardoidun normaalijakauman N(0,1) taulukoiden mukaan

$$-z_{0.005} = -2.58$$
$$+z_{0.005} = +2.58$$

Siten 99 %:n *luottamusväli* Bernoulli-jakauman parametrille *p* on muotoa

$$\begin{aligned} \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} &= 0.1 \pm 2.58\sqrt{\frac{0.1\times(1-0.1)}{150}} \\ &= 0.1 \pm 2.58 \times 0.024 \\ &= 0.1 \pm 0.063 \\ &= (0.037, 0.163) \end{aligned}$$

Väli *peittää* parametrin *p* oletetun arvon 0.05.

Siten otoksesta saatu evidenssi viittaa siihen suuntaan, että valmistajan väitteeseen voidaan kohdistaa jonkin verran epäilyjä. Asiaa voidaan tarkastella myös *tilastollisen testiteorian* muodostamassa kehikossa.

Huomautuksia:

(i) Suhteellisen osuuden luottamusväli *pitenee*, jos luottamustasoa $1 - \alpha$ *kasvatetaan*, jolloin luottamusvälistä tulee *epäinformatiivisempi*.

(ii) Luottamusväli *lyhenee*, jos otoskokoa *n kasvatetaan*.

(iii) Jos luottamusvälin pituus halutaan *puolittaa*, pitää havaintojen lukumäärä *n nelinkertaistaa*.

Esimerkki 8.5.

Tehdas valmistaa ruuveja. Ruuvien paino vaihtelee satunnaisesti noudattaen normaalijakaumaa. Ruuvien joukosta poimittiin yksinkertainen satunnaisotos. Otoskeskiarvoksi saatiin tällöin 25 g. Tehdään (tavallisesti epärealistinen) oletus, että normaalijakauman varianssi 0.25 g^2 on tunnettu.

Määrää 99 %:n *luottamusvälit* ruuvien painon odotusarvolle, jos otoskokona on

(a) 1

(b) 100

(c) 10000

Vertaa saatujen luottamusvälien pituuksia toisiinsa. Miten luottamusvälin pituus käyttäytyy otoskoon funktiona?

Esimerkki 8.5. – Mitä opimme?

Esimerkissä 8.5. tarkastellaan *normaalijakauman odotusarvon luottamusvälin* määräämistä (tavallisesti epärealistisessa) tilanteessa, jossa *jakauman varianssi oletetaan tunnetuksi*.

Esimerkki 8.5. – Ratkaisu:

Tehdas valmistaa ruuveja. Ruuvien paino vaihtelee satunnaisesti noudattaen normaalijakaumaa. Ruuvien joukosta poimittiin yksinkertainen satunnaisotos, jonka koko oli *n*.

Määritellään satunnaismuuttujat

X_i = Ruuvin *i* paino otoksessa, $i = 1, 2, \ldots, n$

Oletuksien mukaan

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

jossa varianssi

$$\sigma^2 = 0.25\ \mathrm{g}^2$$

on tunnettu.

Otokseen poimittujen ruuvien painojen *aritmeettinen keskiarvo* oli

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 25\ \mathrm{g}$$

Konstruoidaan otoksesta saatujen tietojen perusteella $(1 - \alpha)$ %:n *luottamusväli* odotusarvoparametrille μ. Koska varianssi σ^2 oletettiin *tunnetuksi*, luottamusväli on muotoa

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

jossa

$\bar{X}$ = havaintojen *aritmeettinen keskiarvo* otoksessa

σ^2 = jakauman *varianssi*

n = havaintojen *lukumäärä*

$-z_{\alpha/2}$ ja $+z_{\alpha/2}$ = luottamustasoon $(1-\alpha)$ liittyvät *luottamuskertoimet standardoidusta normaalijakaumasta* N(0,1)

Valitaan *luottamustasoksi*

$$1-\alpha = 0.99$$

Koska

$$\alpha = 0.01$$

luottamustasoa $1-\alpha = 0.99$ vastaavat *luottamuskertoimet* ovat

$$-z_{\alpha/2} = -z_{0.005}$$
$$+z_{\alpha/2} = +z_{0.005}$$

Luottamuskertoimet $-z_{\alpha/2} = -z_{0.005}$ ja $+z_{\alpha/2} = +z_{0.005}$ toteuttavat yhtälöt

$$\Pr(z \le -z_{\alpha/2}) = \Pr(z \le -z_{0.005}) = \frac{\alpha}{2} = 0.005$$
$$\Pr(z \ge +z_{\alpha/2}) = \Pr(z \ge +z_{0.005}) = \frac{\alpha}{2} = 0.005$$

jossa satunnaismuuttuja z noudattaa *standardoitua normaalijakaumaa*:

$$z: \ \mathrm{N}(0,1)$$

Siten

$$\Pr(-z_{\alpha/2} \le z \le +z_{\alpha/2}) = \Pr(-z_{0.005} \le z \le +z_{0.005}) = 1-\alpha = 0.99$$

Standardoidun normaalijakauman N(0,1) taulukoiden mukaan

$$-z_{0.005} = -2.58$$
$$+z_{0.005} = +2.58$$

Siten 99 %:n *luottamusväli* normaalijakauman odotusarvoparametrille μ on muotoa

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 25 \pm 2.58 \times \frac{0.5}{\sqrt{n}}$$

(a) $n = 1$:

Luottamusväliksi saadaan

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 25 \pm 2.58 \times \frac{0.5}{\sqrt{1}} = 25 \pm 1.29 = (23.71, 26.29)$$

(b) $n = 100$:

Luottamusväliksi saadaan

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 25 \pm 2.58 \times \frac{0.5}{\sqrt{100}} = 25 \pm 0.129 = (24.871, 25.129)$$

(c) $n = 10000$:

Luottamusväliksi saadaan

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{10000}} = 25 \pm 2.58 \times \frac{0.5}{\sqrt{10000}} = 25 \pm 0.0129 = (24.9871, 25.0129)$$

Jos otantaa toistetaan, niin *luottamustason frekvenssitulkinnan* mukaan otoksista konstruoidut luottamusvälit peittävät (keskimäärin) 99 %:ssa otoksia parametrin μ tuntemattoman arvon ja (keskimäärin) 1 %:ssa otoksia ei sitä tee.

Huomautuksia:

(i) Odotusarvon luottamusväli *pitenee*, jos luottamustasoa $1 - \alpha$ *kasvatetaan*, jolloin luottamisvälistä tulee *epäinformatiivisempi*.

(ii) Luottamusväli *lyhenee*, jos otoskokoa *n kasvatetaan*.

(iii) Jos luottamusvälin pituus halutaan *puolittaa*, pitää havaintojen lukumäärä *n nelinkertaistaa*.

(iv) **Luottamuskertoimet pitää valita normaalijakauman sijasta *t-jakaumasta,* jos varianssi σ^2 *ei ole tunnettu* ja se joudutaan *estimoimaan* otoksesta.**

Näin saatava *estimoituun varianssiin* σ^2 perustuva luottamusväli on *leveämpi* kuin tässä konstruoitu *tunnettuun varianssiin* σ^2 perustuva luottamusväli; ks. luentokalvot.

Jos havaintojen lukumäärän annetaan kasvaa rajatta, *estimoituun varianssiin* σ^2 perustuva luottamisväli *lähestyy tunnettuun varianssiin* σ^2 perustuvaa luottamusväliä.