Tilastolliset menetelmät

Osa 1: Johdanto

- **Johdanto tilastotieteeseen**

Mitä tilastotiede on?

- Tilastotiede kehittää ja soveltaa menetelmiä ja malleja, joiden avulla reaalimaailman ilmiöistä voidaan tehdä johtopäätöksiä tilanteissa, joissa ilmiöitä koskeviin tietoihin liittyy epävarmuutta ja satunnaisuutta.

- Koska tietoihin liittyy epävarmuutta tai satunnaisuutta tilastolliset menetelmät ja mallit perustuvat todennäköisyyslaskentaan.

- Tilastotiede on todennäköisyyslaskennan tärkeimpiä sovellusalueita.

Mitä tilastotiede on?

- Tilastotiedettä voidaan soveltaa kaikkialla missä tuotetaan reaalimailmaa ja sen ilmiöitä kuvaavaa numeerista tai kvantitatiivista tietoa. Jokainen empiirisen tutkimuksen havaintoaineisto on tilastotieteen tutkimuksen mahdollinen kohde.

Tilastollinen aineisto

- Tilastollisen tutkimuksen *kaikki mahdolliset kohteet* muodostavat tutkimuksen (**kohde-**) **perusjoukon**.
- Tutkimuksen kohteiksi valittuja perusjoukon alkioita kutsutaan **havaintoyksiköiksi**.
- **Tilastollinen aineisto** koostuu havaintoyksiköitä kuvaavien muuttujien **havaituista arvoista**.

Tilastollisen kuvailun ja päättelyn menetelmiä

- Kuvailun menetelmiä:
  - Tilastografiikka
  - Tilastolliset tunnusluvut
  - Tilastolliset mallit

- Päättelyn menetelmiä:
  - Tilastolliset mallit
  - Tilastollinen estimointi
  - Tilastollinen testaus

Aineiston kerääminen

- Kohdistuuko tutkimus koko perusjoukkoon vai vain johonkin sen osaan?
  - Jos tutkimuksen kohteena on koko perusjoukko, tutkimusta kutsutaan **kokonaistutkimukseksi**, muuten kyseessä on **otantatutkimus**.

- Muutetaanko tutkimuksessa aktiivisesti tutkimuksen kohteiden olosuhteita?
  - Tutkimus on **koe**, jos tutkitaan olosuhteiden aktiivisen muuttamisen vaikutusta tutkimuksen kohteisiin.
  - Jos olosuhteita ei muuteta aktiivisesti, sanomme, että tutkimus perustuu **suoriin havaintoihin**.

Kontrolloidut kokeet

- Kokeessa ei voida tehdä luotettavia johtopäätöksiä, ellei koe ole kontrolloitu:
  - Kokeessa on vertailtava vähintään kahden erilaisen käsittelyn vaikutuksia.
  - Käsittelyjen kohdistamisessa on käytettävä satunnaistusta.
  - Kokeessa on tehtävä riittävästi koetoistoja.

- Jos koe on kontrolloitu, koetuloksien analysointi tilastotieteen keinoin on mahdollista. Kutsumme kontrolloituja kokeita tavallisesti **tilastollisiksi kokeiksi**.

Mittaaminen

- Tilastollisen tutkimuksen kohdetta kuvaavat numeeriset ja kvantitatiiviset tiedot saadaan mittaamalla.

- Mittari voidaan pitää funktiona, joka liittää mitattavan kohteen ominaisuuteen numeerisen arvon.

Tilastolliset mitta-asteikot

- Mittaus on tehty **nominaali-** eli **laatueroasteikolla**, jos mittaus kertoo mihin luokkaan mittauksen kohde kuuluu. Esim. sukupuoli, asuinpaikka.
- **Ordinaali-** eli **järjestysasteikkoa** käytetään silloin, kun tilastoyksiköt voidaan jakaa luokkiin, joiden välillä on järjestys. Esim. oppiarvot: ylioppilas, kandidaatti, di/maisteri, lisensiaatti, tohtori.
- **Intervalli**- eli **välimatka-asteikko** kertoo kuinka paljon kahden mitattavan kohteen ominaisuudet eroavat toisistaan. Esim. lämpötila Celsius-asteina.
- Mittaus on tehty **suhdeasteikolla**, jos mittaus kertoo kuinka monta kertaa enemmän tai vähemmän mittauksen kohteella on mitattavaa ominaisuutta kuin jollakin toisella kohteella. Esim. pituus, nopeus, lämpötila Kelvin-asteikolla.

Kvalitatiiviset ja kvantitatiiviset muuttujat

- Ominaisuutta ja sitä kuvaavaa muuttujaa kutsutaan
  - **Kvalitatiiviseksi**, jos mittauksen kohteet voidaan luokitella mittauksen perusteella toisistaan eroaviin luokkiin.
  - **Kvantitatiivisiksi**, jos mittaus tuottaa ominaisuuden määrällisen arvon.

Havaintoarvot

- Olkoon tutkimuksen kohteiksi valittujen **havainto-yksiköiden lukumäärä** *n*.
- Olkoon

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

  kohdeperusjoukon alkioiden ominaisuutta kuvaavan muuttujan x **havaittu arvo** havaintoyksikössä i.
- Kutsumme muuttujan x havaittuja arvoja

  $$x_1 \, , x_2 \, , \dots , x_n$$

  tavallisesti **havaintoarvoiksi**.
- Havaintoarvo x_i saadaan *mittaamalla* muuttujan x arvo havaintoyksikölle i.

Havaintoarvojen jakauma ja sen kuvaaminen

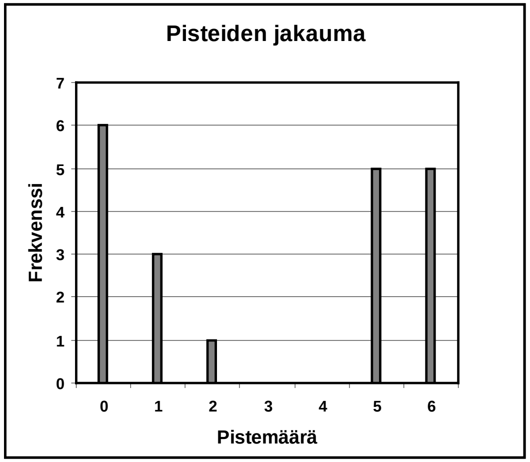
- Perusjoukon alkioiden ominaisuutta kuvaavan muuttujan *x* *havaittujen arvojen*

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

  vaihtelua havaintoyksiköiden joukossa kuvaa parhaiten havaintoarvojen **jakauma**.
- Havaittujen arvojen jakaumaa voidaan kuvailla ja esitellä *tiivistämällä* havaintoarvoihin sisältyvä *informaatio* sopivaan muotoon:
  - Havaintoarvojen *jakaumaa kokonaisuutena* voidaan kuvata sopivasti valitulla **graafisella esityksellä**.
  - *Jakauman karakteristisia ominaisuuksia* voidaan kuvata sopivasti valituilla **tunnusluvuilla**.

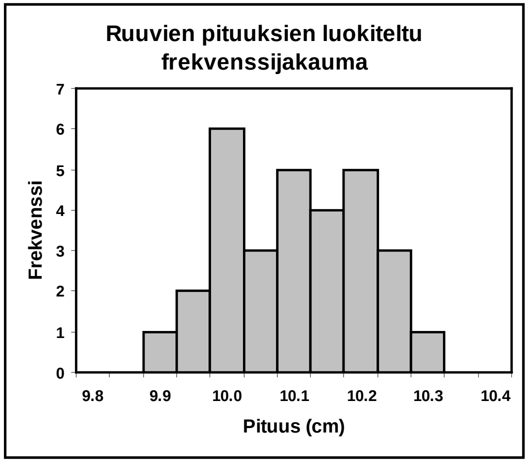
Havaintoarvojen jakauma ja sen kuvaaminen

Muistutus eri graafisista esityksistä ja tunnusluvuista löytyvät Liitteessä **Tilastollisten aineistojen kuvaaminen.**



$$\overline{x} \qquad \frac{f}{n}$$

$$m_k \qquad \textbf{Mediaani}$$



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

Tilastolliset menetelmät

Osa 2: Otokset, otosjakaumat ja estimointi

- **Otokset ja otosjakaumat**

Otokset ja otosjakaumat

>> **Satunnaisotanta ja satunnaisotokset**

Otostunnusluvut ja otosjakaumat

Aritmeettisen keskiarvon ja otosvarianssin otosjakaumat

Suhteellisen frekvenssin otosjakauma

Tilastollinen aineisto

- **Tilastollinen aineisto** koostuu tutkimuksen kohteita kuvaavien muuttujien *havaituista arvoista*.
- Tilastollisissa tutkimusasetelmissa havaintoarvoihin ajatellaan aina liittyvän *epävarmuutta* ja *satunnaisuutta*.
- Seurauksia:

(i) Tilastollisissa tutkimusasetelmissa ajatellaan, että *havaintoarvot on generoinut jokin satunnaisilmiö*.

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

Tilastollisen aineiston tilastollinen malli

- Tilastollisen havaintoaineiston **tilastollisella mallilla** tarkoitetaan tutkimuksen kohteita kuvaavien satunnaismuuttujien *todennäköisyysjakaumaa*.
- Tämän todennäköisyysjakauman ajatellaan *generoineen* ko. satunnaismuuttujien *havaitut arvot*.
- Havaintoarvojen ajatellaan syntyneen *arpomalla* käyttäen arvontatodennäköisyyksinä tilastollisena mallina käytetystä todennäköisyysjakaumasta saatavia todennäköisyyksiä.
- **Huomautus:**

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

Tilastolliset mallit ja tilastollinen päättely

- Kun tilastollista mallia sovelletaan jotakin reaalimaailman ilmiötä kuvaavan havaintoaineiston analysointiin, kohdataan lähes aina seuraavat mallin **parametreja** koskevat ongelmat:
  - (i) Parametrien arvoja *ei tunneta* ja ne on **estimoitava** eli *arvioitava* havaintoaineistosta.
  - (ii) Parametrien arvoista on esitetty *oletuksia* tai *väitteitä*, joita halutaan **testata** eli asettaa koetteelle havainto-aineistosta saatua informaatiota vastaan.
- Tilastollisten mallien parametrien estimointi ja testaus muodostavat keskeisen osan **tilastollista päättelyä**.

Satunnaisotanta ja satunnaisotokset

- **Satunnaisotos** poimitaan perusjoukosta *arpomalla* tutkittavat havaintoyksiköt perusjoukosta otokseen.
- Arvonnassa käytettävää menetelmää kutsutaan **satunnaisotannaksi**.
- Satunnaisotannassa *sattuma* määrää mitkä perusjoukon alkioista tulevat poimituiksi otokseen.

Satunnaisotanta: Kommentteja

- Jos havaintoyksiköt poimitaan perusjoukosta satunnaisotannalla, pätee seuraava:

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

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

Yksinkertainen satunnaisotanta

- Olkoot

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

  riippumattomia, *identtisesti jakautuneita* satunnais-muuttujia, joilla on *sama* pistetodennäköisyys- tai tiheysfunktio

  $$f(x)$$

- Tällöin sanomme, että satunnaismuuttujat

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

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

Havainnot ja havaintoarvot

- Olkoon

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

  (*yksinkertainen*) *satunnaisotos* jakaumasta $f(x)$.
- Kutsumme satunnaismuuttujia $X_1 , X_2 , \ldots , X_n$ tavallisesti **havainnoiksi**.
- *Otoksen poimimisen jälkeen* satunnaismuuttujat $X_1 , X_2 , \ldots , X_n$ saavat havaituiksi arvoikseen **havainto-arvot**

  $$x_1 , x_2 , \ldots , x_n$$
- Merkitään:

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

Yksinkertainen satunnaisotanta: Kommentteja 1/2

- Olkoon

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

  satunnaisotos jakaumasta $f(x)$.
- Tällöin havaintoarvot

  $$x_1 , x_2 , \dots , x_n$$

  on saatu *toistamalla arvontaa toisistaan riippumattomin toistoin n kertaa samoin, jakaumasta* $f(x)$ *saatavin toden-näköisyyksin*.

Yksinkertainen satunnaisotanta: Kommentteja 2/2

- Havaintoarvot $x_1 , x_2 , \ldots , x_n$ ovat *kiinteitä* eli *ei-satunnaisia*, mutta ne *vaihtelevat toisistaan riippumatta ja satunnaisesti otoksesta toiseen*.
- **Satunnaisuus *ei siis liity* otannan tuloksena saatuihin havaintoarvoihin, vaan poiminnassa sovellettuun arvontaan.**

Tilastollinen malli yksinkertaiselle satunnaisotannalle 1/2

- Olkoon

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

  satunnaisotos jakaumasta $f(x)$.
- Satunnaismuuttujien $X_1 , X_2 , \dots , X_n$ *yhteisjakauma* muodostaa **tilastollisen mallin** *havaintoarvojen satunnaiselle vaihtelulle otoksesta toiseen*.

Tilastollinen malli yksinkertaiselle satunnaisotannalle 2/2

- Koska satunnaismuuttujat

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

  on oletettu riippumattomiksi, niin satunnaismuuttujien $X_1 , X_2 , \ldots , X_n$ **yhteisjakauma** on muotoa

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

  jossa

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

Otokset ja otosjakaumat

Satunnaisotanta ja satunnaisotokset

>> Otostunnusluvut ja otosjakaumat

Aritmeettisen keskiarvon ja otosvarianssin otosjakaumat

Suhteellisen frekvenssin otosjakauma

Otostunnusluvut 1/3

- Olkoon

  $X_1 , X_2 , \dots , X_n$

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

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

  $X_i \sim f(x) \, , i = 1, 2, \dots, n$

Otostunnusluvut 2/3

- Olkoon

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

  jokin *satunnaismuuttujien* X_1 , X_2 , … , X_n (mitallinen) *funktio*.
- Kutsumme satunnaismuuttujaa T seuraavassa (**otos-**) **tunnusluvuksi**.

Otostunnusluvut 3/3

- 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 satunnaismuuttujat

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

  muodostavat *satunnaisotoksen* jakaumasta $f(x)$ ja olkoon funktio

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

  jokin *otostunnusluku*.
- Tunnusluvun T jakaumaa kutsutaan *tunnusluvun T* **otosjakaumaksi**.
- Tunnusluvun T otosjakauma muodostaa **tilastollisen mallin** *tunnusluvun T arvojen satunnaiselle vaihtelulle otoksesta toiseen*.

Eräiden tavallisten tunnuslukujen otosjakaumat

- Olkoon

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

  satunnaisotos jakaumasta $f(x)$.
- Jatkossa tarkastellaan seuraavien tunnuslukujen (ks. lukua **Tilastollisten aineistojen kuvaaminen**) otosjakaumia:
  - **Aritmeettinen keskiarvo**
  - **Otosvarianssi**
  - **Suhteellinen frekvenssi**

Otokset ja otosjakaumat

Satunnaisotanta ja satunnaisotokset

Otostunnusluvut ja otosjakaumat

>> Aritmeettisen keskiarvon ja otosvarianssin otosjakaumat

Suhteellisen frekvenssin otosjakauma

Aritmeettinen keskiarvo: Määritelmä 1/2

- Oletetaan, että havainnot

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

muodostavat *satunnaisotoksen* satunnaismuuttujan X jakaumasta, jonka *odotusarvo* ja *varianssi* ovat

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

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

- Tällöin kaikilla satunnaismuuttujilla X_i , $i = 1, 2, \ldots , n$ on *sama* odotusarvo μ ja *sama* varianssi σ^2 .

Aritmeettinen keskiarvo: Määritelmä 2/2

- Olkoon

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

havaintojen X_1 , X_2 , … , X_n **aritmeettinen keskiarvo**.

- Aritmeettinen keskiarvo $\bar{X}$ kuvaa havaintojen *keski-määräistä arvoa*.
- Aritmeettinen keskiarvo $\bar{X}$ on satunnaismuuttuja, jonka saamat *arvot vaihtelevat satunnaisesti otoksesta toiseen*.

Aritmeettisen keskiarvon odotusarvo ja varianssi

- **Aritmeettisen keskiarvon** $\bar{X}$ **odotusarvo** ja **varianssi**:

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

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

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

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

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

Aritmeettisen keskiarvon otosjakauman käyttäytyminen otoskoon kasvaessa

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

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

ja varianssi on

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

niin aritmeettisen keskiarvon otosjakauma *keskittyy yhä voimakkaammin havaintojen yhteisen odotusarvon* μ *ympärille*, *jos otoskoon* n *annetaan kasvaa*.

Aritmeettisen keskiarvon otosjakauma normaalisen otoksen tapauksessa

- Oletetaan, että havainnot

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

  muodostavat *satunnaisotoksen normaalijakaumasta*

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

  Tällöin havaintojen *aritmeettinen keskiarvo* $\bar{X}$ *noudattaa eksaktisti* (eli myös äärellisissä otoksissa) **normaali-jakaumaa**:

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

Aritmeettisen keskiarvon asymptoottinen jakauma

- Oletetaan, että havainnot

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

  muodostavat *satunnaisotoksen* satunnaismuuttujan X jakaumasta, jonka *odotusarvo* on μ ja *varianssi* on σ^2.
- Tällöin havaintojen *aritmeettinen keskiarvo* $\bar{X}$ *noudattaa keskeisen raja-arvolauseen mukaan suurissa otoksissa approksimatiivisesti* **normaalijakaumaa**, jonka odotusarvo on μ ja varianssi on σ^2/n:

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

Aritmeettisen keskiarvon jakauma: Kommentteja

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

Otosvarianssi: Määritelmä 1/2

- Oletetaan, että havainnot

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

muodostavat *satunnaisotoksen* satunnaismuuttujan X jakaumasta, jonka *odotusarvo* ja *varianssi* ovat

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

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

- Tällöin kaikilla satunnaismuuttujilla X_i , $i = 1, 2, \ldots , n$ on *sama* odotusarvo μ ja *sama* varianssi σ^2 .

Otosvarianssi: Määritelmä 2/2

- Olkoon

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

havaintojen X_1 , X_2 , … , X_n **otosvarianssi**, jossa

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

on havaintojen X_1 , X_2 , … , X_n *aritmeettinen keskiarvo*.

- Otosvarianssi S^2 kuvaa havaintoarvojen *vaihtelua niiden aritmeettisen keskiarvon ympärillä*.
- Otosvarianssi S^2 on satunnaismuuttuja, jonka saamat *arvot vaihtelevat satunnaisesti otoksesta toiseen*.

Otosvarianssin odotusarvo ja varianssi

- **Otosvarianssin** S^2 **odotusarvo**:

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

- Jos lisäksi voidaan olettaa, että havainnot X_1 , X_2 , … , X_n noudattavat normaalijakaumaa $\mathrm{N}(\mu, \sigma^2)$, niin **otosvarianssin** S^2 **varianssi** on

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

- Siten otosvarianssin S^2 *standardipoikkeama* on normaalisen otoksen tapauksessa

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

Otosvarianssin otosjakauma normaalisen otoksen tapauksessa 1/2

- Oletetaan, että havainnot

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

muodostavat *satunnaisotoksen normaalijakaumasta*

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

- Tällöin *satunnaismuuttuja*

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

noudattaa $\boldsymbol{\chi}^{\mathbf{2}}$**-jakaumaa** *vapausastein n*:

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

Otosvarianssin otosjakauma normaalisen otoksen tapauksessa 2/2

- Oletetaan, että havainnot

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

muodostavat *satunnaisotoksen normaalijakaumasta*

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

- Tällöin *satunnaismuuttuja*

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

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

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

Aritmeettisen keskiarvon ja otosvarianssin otosjakaumat: Seuraus

- Oletetaan, että havainnot

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

  muodostavat *satunnaisotoksen normaalijakaumasta*

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

- Tällöin

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

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

Otokset ja otosjakaumat

Satunnaisotanta ja satunnaisotokset

Otostunnusluvut ja otosjakaumat

Aritmeettisen keskiarvon ja otosvarianssin otosjakaumat

>> Suhteellisen frekvenssin otosjakauma

Frekvenssi ja suhteellinen frekvenssi: Määritelmät 1/3

- Olkoon P jokin otosavaruuden S alkioiden *ominaisuus*.
- Jos otosavaruuden S alkiolla x on ominaisuus P, merkitään

  $$P(x)$$

- Olkoon

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

  niiden otosavaruuden S alkioiden *osajoukko*, joilla on ominaisuus P.
- Oletetaan, että **tapahtuman *A* todennäköisyys** on

  $$\Pr(A) = p$$

Frekvenssi ja suhteellinen frekvenssi: Määritelmät 2/3

- Poimitaan otosavaruudesta S *satunnaisotos*, jonka *koko* on n.
- Olkoon

  $$f$$

  niiden havaintoyksiköiden **frekvenssi**, joilla on ominaisuus P ja olkoon

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

  vastaava **suhteellinen frekvenssi**.

Frekvenssi ja suhteellinen frekvenssi: Määritelmät 3/3

- Frekvenssi

  $$f$$

  kuvaa *A-tyyppisten alkioiden* **lukumäärää** otoksessa ja vastaava suhteellinen frekvenssi

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

  kuvaa *A-tyyppisten alkioiden* **suhteellista osuutta** otoksessa.
- Frekvenssi f ja vastaava suhteellinen frekvenssi $\hat{p}$ ovat *satunnaismuuttujia*, joiden saamat *arvot vaihtelevat satunnaisesti otoksesta toiseen*.

Frekvenssi: Odotusarvo, varianssi ja jakauma 1/2

- Olkoon A jokin otosavaruuden S *tapahtuma*:

$$A \subset S$$

- Poimitaan otosavaruudesta S *satunnaisotos*, jonka *koko* on n.
- Olkoon

$$f$$

A-tyyppisten alkioiden *lukumäärä* eli *frekvenssi* otoksessa.

Frekvenssi: Odotusarvo, varianssi ja jakauma 2/2

- **Frekvenssin** f **odotusarvo** ja **varianssi**:

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

$$\mathrm{Var}(f) = npq$$

jossa $q = 1 - p$.

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

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

Ks. monisteen **Todennäköisyyslaskenta** lukua **Diskreettejä jakaumia** tai lukua **Satunnaismuuttujien muunnosten jakaumat.**

Suhteellinen frekvenssi: Odotusarvo ja varianssi

- **Suhteellisen frekvenssin** $\hat{p}$ **odotusarvo** ja **varianssi**:

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

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

jossa $q = 1 - p$.

- Suhteellisen frekvenssin $\hat{p}$ *standardipoikkeamaa*

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

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

Suhteellisen frekvenssin otosjakauma: Jakauman käyttäytyminen otoskoon kasvaessa

- Koska suhteellisen frekvenssin $\hat{p}$ odotusarvo

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

  ja varianssi on

  $$\mathrm{Var}(\hat{p}) = pq/n\,,\ q = 1 - p$$

  niin suhteellisen frekvenssin otosjakauma *keskittyy yhä voimakkaammin tapahtuman A todennäköisyyden p ympärille*, *jos otoskoon n annetaan kasvaa*.

Suhteellisen frekvenssin otosjakauma: Asymptoottinen jakauma

- *Suhteellinen frekvenssi* $\hat{p}$ *noudattaa keskeisen raja-arvo-lauseen mukaan suurissa otoksissa approksimatiivisesti* **normaalijakaumaa**:

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

- Siten *standardoitu satunnaismuuttuja*

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

noudattaa suurissa otoksissa approksimatiivisesti standardoitua normaalijakaumaa:

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

Tilastolliset menetelmät

Osa 2: Otokset, otosjakaumat ja estimointi

- **Estimointi**

Estimointi

>> **Todennäköisyysjakaumien parametrit ja niiden estimointi**

Hyvän estimaattorin ominaisuudet

Estimointimenetelmät

Todennäköisyysjakaumat tilastollisten aineistojen kuvaajina

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

Todennäköisyysjakaumien parametrit 1/2

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

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

  riippuu **parametrista** θ.
- Merkintä:

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

Todennäköisyysjakaumien parametrit 2/2

- Satunnaismuuttujan X pistetodennäköisyys- tai tiheysfunktio

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

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

Yksinkertainen satunnaisotos

- Olkoon

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

(**yksinkertainen**) **satunnaisotos** jakaumasta, jonka pistetodennäköisyys- tai tiheysfunktio $f(x ; \theta)$ riippuu parametrista θ.

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

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

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

Havainnot ja havaintoarvot

- Oletetaan, että satunnaismuuttujat (havainnot)

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

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

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

- **Havaintoarvot**

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

  vaihtelevat satunnaisesti otoksesta toiseen.

Estimaattorit ja estimaatit

- Oletetaan, että *parametrin* θ *estimoimiseen* käytetään satunnaismuuttujien $X_1 , X_2 , \dots , X_n$ funktiota eli *tunnuslukua*

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

- Tällöin funktiota $T = g(X_1 , X_2 , \dots , X_n)$ kutsutaan parametrin θ **estimaattoriksi**

  ja *havaintoarvoista* $x_1 , x_2 , \dots , x_n$ laskettua arvoa $t = g(x_1 , x_2 , \dots , x_n)$ kutsutaan parametrin θ **estimaatiksi**.

Estimaattorit ja estimaatit: Kommentti

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

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

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

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

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

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

  parametrille θ.

Estimaattorin otosjakauma

- Estimaattorin

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

  havaintoarvoista

  $$x_1 , x_2 , \dots , x_n$$

  lasketut arvot eli estimaatit

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

  vaihtelevat satunnaisesti otoksesta toiseen.
- Estimaattorin T arvojen satunnaista vaihtelua otoksesta toiseen voidaan kuvata *estimaattorin T* **otosjakaumalla**; ks. lukua **Otokset ja otosjakaumat.**

Estimointi

Todennäköisyysjakaumien parametrit ja niiden estimointi

\>> Hyvän estimaattorin ominaisuudet

Estimointimenetelmät

Hyvä estimaattori

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

Hyvän estimaattorin ominaisuudet

Harhattomuus ja harha

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

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

- Harhaton estimaattori tuottaa *keskimäärin* oikean kokoisia arvoja parametrille.

- Estimaattorin $\hat{\theta}$ **harha** on

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

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

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

Estimaattorin keskineliövirhe ja tarkkuus

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

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

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

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

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

niin

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

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

Tehokkuus

- Olkoot $\hat{\theta}_1$ ja $\hat{\theta}_2$ kaksi parametrin θ estimaattoria.
- Estimaattori $\hat{\theta}_1$ on **tehokkaampi** kuin estimaattori $\hat{\theta}_2$, jos

$$\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2)$$

- Parametrin θ estimaattori $\hat{\theta}$ on **täystehokas**, jos sen varianssi on *pienempi kuin minkä tahansa muun parametrin θ estimaattorin*.

Hyvän estimaattorin ominaisuudet

Harhattomuus ja tehokkuus: Esimerkki 1/2

- Olkoon

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

  satunnaisotos normaalijakaumasta N(μ, σ^2).
- Voidaan osoittaa, että sekä havaintojen *aritmeettinen keskiarvo*

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

  että niiden *mediaani Me* ovat molemmat *harhattomia* normaalijakauman odotusarvolle μ :

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

Hyvän estimaattorin ominaisuudet

Harhattomuus ja tehokkuus: Esimerkki 2/2

- Sen sijaan

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

joten normaalijakautuneiden havaintojen aritmeettisen keskiarvon $\bar{X}$ varianssi on *pienempi* kuin niiden mediaanin *Me* varianssi.

- Siten aritmeettinen keskiarvo $\bar{X}$ on normaalijakauman odotusarvon μ estimaattorina *tehokkaampi* kuin mediaani *Me*.
- Voidaan osoittaa, että havaintoarvojen aritmeettinen keskiarvo $\bar{X}$ on normaalijakauman odotusarvoparametrin μ estimaattorina *täystehokas*.

Tarkentuvuus

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

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

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

Estimointi

Todennäköisyysjakaumien parametrit ja niiden estimointi

Hyvän estimaattorin ominaisuudet

\>> Estimointimenetelmät

Estimaattoreiden johtaminen

- *Hyvien estimaattoreiden johtaminen* todennäköisyys-jakaumien tuntemattomille parametreille *on teoreettisen tilastotieteen keskeisiä ongelmia*.

- **Suurimman uskottavuuden menetelmä** *on ehkä tärkein kaikista estimointimenetelmistä*. **Momenttimenetelmä** on toinen hyvin yleinen estimaattoreiden johtamiseen käytettävä menetelmä.

Estimointimenetelmät

>> **Suurimman uskottavuuden menetelmä**

- **Normaalijakauman parametrien estimointi**
- **Bernoulli-jakauman parametrien estimointi**

Uskottavuusfunktio 1/2

- Olkoon $X_1 , X_2 , \ldots , X_n$ *yksinkertainen satunnaisotos* jakaumasta $f(x\ ;\ \theta)$, jonka *parametrina* on θ.

- Koska havainnot $X_1 , X_2 , \ldots , X_n$ on tässä oletettu riippumattomiksi, niiden *yhteisjakauman piste-todennäköisyys*- tai *tiheysfunktio* on

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

  jossa

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

  on havaintoon X_i liittyvä pistetodennäköisyys- tai tiheysfunktio.

Uskottavuusfunktio 2/2

- Otoksen X_1 , X_2 , … , X_n **uskottavuusfunktio**

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

  on havaintojen

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

  yhteisjakauman pistetodennäköisyys- tai *tiheysfunktion f* arvo pisteessä x_1 , x_2 , … , x_n tulkittuna *parametrin* θ *arvojen funktioksi*.

Suurimman uskottavuuden estimaattori 1/2

- Olkoon

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

  parametrin θ arvo, joka *maksimoi uskottavuusfunktion*

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

  parametrin θ suhteen.

- **Huomautus:**

  Uskottavuusfunktion L maksimin antava parametrin θ arvo t on muuttujien X_1 , X_2 , … , X_n havaittujen arvojen x_1 , x_2 , … , x_n funktio.

Suurimman uskottavuuden estimaattori 2/2

- Sijoittamalla uskottavuusfunktion L maksimin parametrin θ suhteen antavassa lausekkeessa

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

  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**

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

Kommentteja

- Suurimman uskottavuuden menetelmän idea on antaa mallin parametreille sellaiset arvot, jotka tekevät saadun havaintoaineiston mahdollisimman todennäköiseksi.

SU-estimaattorin määrääminen 1/2

- Parametrin θ suurimman uskottavuuden estimaattori määrätään *maksimoimalla uskottavuusfunktio*

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

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

  $$L'(\theta)$$

  nollaksi ja ratkaisemalla θ saadusta *yhtälöstä*

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

SU-estimaattorin määrääminen 2/2

- Jos parametrin θ arvo

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

  tuottaa uskottavuusfunktion $L(\theta)$ maksimin, parametrin θ *suurimman uskottavuuden estimaattori* on

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

  jossa X_1 , X_2 , … , X_n on yksinkertainen satunnaisotos siitä jakaumasta, johon uskottavuusfunktio $L(\theta)$ liittyy.

Logaritminen uskottavuusfunktio 1/3

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

$$l(\theta) = \log L(\theta)$$

- Tämä johtuu seuraavista seikoista:

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

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

Logaritminen uskottavuusfunktio 2/3

- Koska havainnot X_1 , X_2 , … , X_n on tässä oletettu *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 havaintoarvoon x_i liittyvä logaritminen uskottavuus-funktio.

Logaritminen uskottavuusfunktio 3/3

- Logaritmisen uskottavuusfunktion summaesityksen
$$l(\theta) = l(\theta\,;x_1) + l(\theta\,;x_2) + \cdots + l(\theta\,;x_n)$$
maksimointi on usein *ratkaisevasti helpompaa* kuin uskottavuusfunktion itsensä maksimointi.

SU-estimaattorin ominaisuudet

- Olkoon X_1 , X_2 , … , X_n *yksinkertainen satunnaisotos* jakaumasta $f(x\,;\,\theta)$, jonka *parametrina* on θ.
- Olkoon $\hat{\theta}$ parametrin θ **suurimman uskottavuuden** eli **SU-estimaattori**.
- Hyvä estimaattori on *harhaton*, *tehokas* ja *tarkentuva* .

- **SU-estimaattori ei välttämättä täytä yhtäkään hyvän estimaattorin kriteereistä, joten suurimman uskottavuuden menetelmää käytettäessä on aina erikseen varmistettava tuloksena saadun estimaattorin hyvyys.**

SU-estimaattorin asymptoottiset ominaisuudet

- Jos SU-estimaattori ei täytä hyvän estimaattorin kriteereitä *äärellisillä havaintojen lukumäärillä*, sen käyttöä 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**.

Esimerkkejä

- Olkoon

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

  yksinkertainen satunnaisotos jakaumasta $f(x\,;\,\theta)$.
- Tarkastellaan seuraavien jakaumien parametrien SU-estimointia eli estimointia suurimman uskottavuuden menetelmällä
  - **Normaalijakauma**
  - **Bernoulli-jakauma**

Normaalijakauma ja sen parametrointi

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

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

$$-\infty < \mu < +\infty\ ,\ \sigma > 0$$

- Normaalijakauman *parametreina* ovat jakauman *odotusarvo*

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

ja *varianssi*

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

Otos normaalijakaumasta

- Olkoon

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

  yksinkertainen satunnaisotos normaalijakaumasta

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

- Tällöin havainnot $X_1 , X_2 , \ldots , X_n$ ovat *riippumattomia, samaa normaalijakaumaa*

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

  noudattavia satunnaismuuttujia.

Normaalijakautuneen otoksen uskottavuusfunktio ja log-uskottavuusfunktio

- Otoksen X_1 , X_2 , … , X_n *uskottavuusfunktio* on

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

- Otoksen X_1 , X_2 , … , X_n *logaritminen uskottavuusfunktio* on

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

Normaalijakauman parametrien SU-estimaattorit

- Normaalijakauman N(μ, σ^2) odotusarvon μ ja varianssin σ^2 **suurimman uskottavuuden estimaattorit** ovat havaintojen X_1 , X_2 , … , X_n *aritmeettinen keskiarvo*

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

ja *otosvarianssi*

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

SU-estimaattoreiden johto 1/2

- *Derivoidaan* logaritminen uskottavuusfunktio

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

ensin parametrin μ suhteen ja merkitään derivaatta nollaksi:

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

- Derivaatan ainoa *nollakohta*

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

antaa log-uskottavuusfunktion *maksimin* parametrin μ suhteen.

Normaalijakauman parametrien estimointi SU-menetelmällä

SU-estimaattoreiden johto 2/2

- *Sijoitetaan* ratkaisu $\mu = \bar{x}$ logaritmiseen uskottavuusfunktioon:

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

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

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

- Derivaatan ainoa *nollakohta*

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

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

SU-estimaattoreiden ominaisuudet 1/2

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

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

  (ii) $\hat{\mu}$ *on tehokas* eli minimivarianssinen estimaattori.

  (iii) $\hat{\mu}$ on *tarkentuva*.

  (iv) $\hat{\mu}$ noudattaa *normaalijakaumaa*:

$$\hat{\mu} \sim \mathrm{N}\left(\mu , \frac{\sigma^2}{n} \right)$$

SU-estimaattoreiden ominaisuudet 2/2

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

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

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

on *harhaton*.

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

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

(iv) (n – 1) S^2/σ^2 noudattaa χ^2-*jakaumaa*:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

Bernoulli-jakauma ja sen parametrointi

- 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** B(p) ja sen *pistetodennäköisyysfunktio* on

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

- Bernoulli-jakauman ainoa *parametri p* yhtyy jakauman *odotusarvoon*:

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

Otos Bernoulli-jakaumasta

- Olkoon

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

  yksinkertainen satunnaisotos Bernoulli-jakaumasta

  $$\mathrm{B}(p)$$

- Tällöin havainnot $X_1 , X_2 , \ldots , X_n$ ovat *riippumattomia*, s*amaa Bernoulli-jakaumaa*

  $$\mathrm{B}(p)$$

  noudattavia satunnaismuuttujia.

Bernoulli-jakautuneen otoksen uskottavuusfunktio ja log-uskottavuusfunktio

- Otoksen X_1 , X_2 , …, X_n *uskottavuusfunktio* on

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

- Otoksen X_1 , X_2 , …, X_n *logaritminen uskottavuusfunktio* on

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

Bernoulli-jakauman odotusarvoparametrin SU-estimaattori 1/2

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

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

Bernoulli-jakauman odotusarvoparametrin SU-estimaattori 2/2

- Koska

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

niin

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

jossa f on tapahtuman A *frekvenssi* otoksessa.

- Siten Bernoulli-jakauman odotusarvoparametrin p *suurimman uskottavuuden estimaattori*

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

on tapahtuman A *suhteellinen frekvenssi* otoksessa.

SU-estimaattorin johto

- *Derivoidaan* logaritminen uskottavuusfunktio

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

parametrin *p* suhteen ja merkitään derivaatta nollaksi:

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

- Derivaatan ainoa *nollakohta*

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

antaa uskottavuusfunktion *maksimin*.

SU-estimaattorin ominaisuudet

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

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

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

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

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

$$\hat{p} \sim_a \mathrm{N}\left(p\,,\, \frac{p(1-p)}{n} \right)$$

Tilastolliset menetelmät

Osa 2: Otokset, otosjakaumat ja estimointi

- **Väliestimointi**

Piste-estimointi ja väliestimointi

- Todennäköisyysjakauman *parametrin arvon estimointia* kutsutaan tavallisesti **piste-estimoinniksi**.
- Estimaatin arvo vaihtelee otoksesta toiseen, joten siihen liittyy tietty epätarkkuus. Herää kysymys, missä *rajoissa* parametrin todellinen arvo θ on.

- Ei ole järkevää yrittää antaa rajoja, joiden sisällä θ:n arvo on ’täysin varmasti’. Sen sijaan voidaan yrittää antaa väli, jonka uskotaan peittävän alleen parametrin tarkan arvon tietyllä suurella todennäköisyydellä. Tämmöistä väliä kutsutaan **luottamusväliksi.**

- Luottamusvälin määräämistä kutsutaan **väliestimoinniksi**.

Väliestimointi

>> **Luottamusväli**

Normaalijakauman odotusarvon luottamusväli

Normaalijakauman varianssin luottamusväli

Bernoulli-jakauman odotusarvon luottamusväli

Luottamusväli ja luottamustaso

- *Väliestimoinnissa* todennäköisyysjakauman $f(x\,;\,\theta)$ tuntemattomalle parametrille θ pyritään määräämään *havainnoista riippuva väli*, *joka tietyllä*, *soveltajan valitsemalla todennäköisyydellä*, *peittää parametrin todellisen arvon*.

- Konstruoitua väliä kutsutaan **luottamusväliksi** ja valittua todennäköisyyttä kutsutaan **luottamustasoksi**.

Luottamusvälin määrääminen

- Olkoon $f(x\,;\,\theta)$ todennäköisyysjakauma, jonka määrää tuntematon parametri θ, ja $X_1, X_2, \ldots, X_n$ *satunnaisotos* jakaumasta $f(x\,;\,\theta)$.

- Valitaan **luottamustaso** $(1 - \alpha)$ siten, että

$$0 < 1 - \alpha < 1$$

- Määrätään satunnaismuuttujat

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

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

siten, että

$$\Pr(\theta \in (L, U)) = 1 - \alpha$$

- Tällöin sanomme, että väli (L, U) on **parametrin $\boldsymbol{\theta}$ luottamusväli luottamustasolla $(1 - \alpha)$**.

Luottamusvälin määrääminen: Kommentit

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

- Käytännössä θ:n arvo on tuntematon eikä ole tietoa, onko sen arvo saadun luottamusvälin sisällä. Näin kuitenkin uskotaan. Käytettäessä esim. 95%:n luottamusväliä keskimäärin 5 kertaa sadasta käy niin, että tarkka arvo jää saadun välin ulkopuolelle.

Johtopäätökset luottamisväleistä

- Tehdään *johtopäätös*, että konstruoitu luottamusväli peittää parametrin θ tuntemattoman 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.
- **Huomautus:**

  Virheellisen johtopäätöksen mahdollisuutta *ei saada häviämään*, ellei luottamusväliä tehdä *niin leveäksi*, että väli *ei enää sisällä informaatiota* parametrin θ todellisesta arvosta.

Esimerkki: Normaalijakauman odotusarvon luottamusväli, (kun σ*:n arvo tiedossa)* 1/2

- Jos estimaattori $\hat{\theta}$ on jatkuva ja sen jakauma on tiedossa, luottamusväli voidaan ilmoittaa havaintojen avulla.

- Olkoon X ~ N(μ, σ^2) ja X_1 , X_2 , … , X_n (*yksinkertainen*) *satunnaisotos* normaalijakaumasta N(μ, σ^2). Silloin μ:n harhaton estimaattori on

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

ja tiedetään että

$$\bar{X} \sim N(\mu, \sigma^2 / n)$$

Esimerkki: Normaalijakauman odotusarvon luottamusväli, (kun σ:*n arvo tiedossa)* 2/2

- Oletetaan että σ:n arvo on tiedossa.
- Lasketaan μ:n 95%:n luottamusväli: Tiedetään taulukoista että

$$P\left(\left|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right| \leq 1.96\right) = 0.95$$

- Tämä voi muokata muotoon

$$P(\bar{X} - 1.96\sigma/\sqrt{n} \leq \mu \leq \bar{X} + 1.96\sigma/\sqrt{n}) = 0.95$$

- Joten väli $\left[\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n}\right]$ on μ:n 95%:n luottamusväli.
- Otoksesta $(x_1,....,x_n)$ lasketaan μ:n 95%:n luottamusvälin arvoksi

$$\left[\bar{x} - 1.96\sigma/\sqrt{n}, \bar{x} + 1.96\sigma/\sqrt{n}\right]$$

Väliestimointi

Luottamusväli

\>> Normaalijakauman odotusarvon luottamusväli

Normaalijakauman varianssin luottamusväli

Bernoulli-jakauman odotusarvon luottamusväli

Normaalijakauman parametrien estimointi

- Olkoon X_1 , X_2 , … , X_n (*yksinkertainen*) *satunnaisotos* normaalijakaumasta $N(\mu, \sigma^2)$

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

  (i) *Odotusarvoparametrin* μ harhaton estimaattori:

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

  (ii) *Varianssiparametrin* σ^2 harhaton estimaattori:

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

Luottamuskertoimet

- Olkoon valittu luottamustaso $(1 - \alpha)$.
- Määrätään **luottamuskertoimet** $-t_{\alpha/2}$ ja $+t_{\alpha/2}$ siten, että

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

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

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

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

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

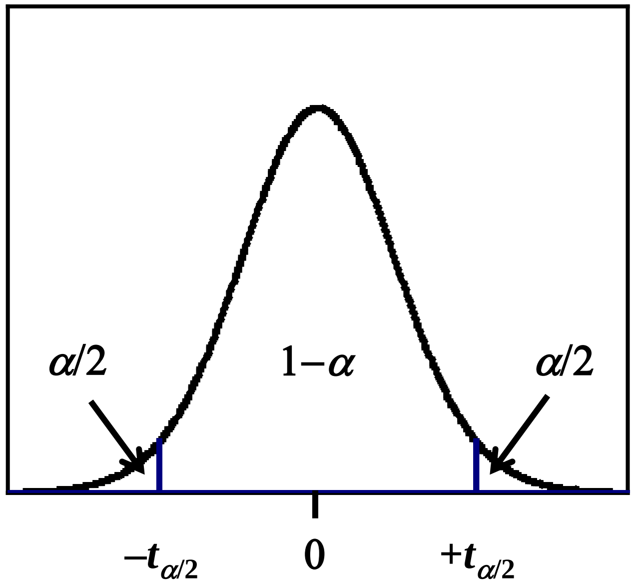
$$\Pr(-t_{\alpha/2} \leq t \leq +t_{\alpha/2}) = 1-\alpha$$

Luottamuskertoimien määrääminen: Havainnollistus

- Luottamuskertoimet $-t_{\alpha/2}$ ja $+t_{\alpha/2}$ jakavat *t*-jakauman tiheysfunktion kuvaajan alle jäävän todennäköisyysmassan (= 1) *kolmeen* osaan:
  - (1) Pisteen $-t_{\alpha/2}$ *vasemmalle* puolelle jää $\alpha/2$ % massasta.
  - (2) Pisteen $+t_{\alpha/2}$ *oikealle* puolelle jää $\alpha/2$ % massasta.
  - (3) Pisteiden $-t_{\alpha/2}$ ja $+t_{\alpha/2}$ *väliin* jää $(1 - \alpha)$ % massasta.
- Huomaa, että

$$\alpha/2 + \alpha/2 + (1 - \alpha) = 1$$

t-jakauman tiheysfunktio



Luottamusväli normaalijakauman odotusarvolle

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

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

jossa

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

S^2 = havaintojen harhaton *otosvarianssi*

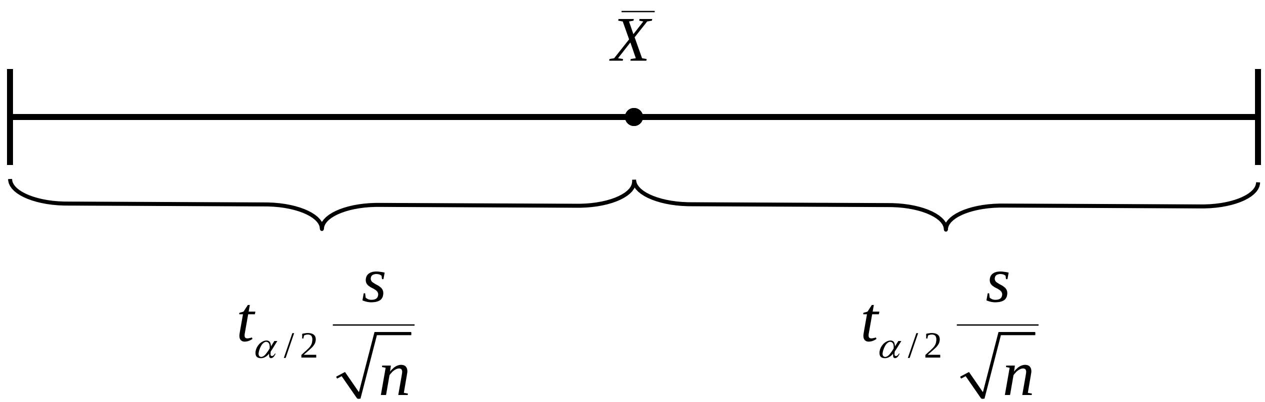
n = havaintojen *lukumäärä*

$-t_{\alpha/2}$, $+t_{\alpha/2}$

= luottamustasoon $(1 - \alpha)$ liittyvät *luottamuskertoimet t-jakaumasta vapausastein* $(n - 1)$

Luottamusväli ja sen pituus

- Normaalijakauman odotusarvon μ luottamusväli luottamustasolla $(1 - \alpha)$:



Luottamusvälin johto 1/5

- Olkoon

  $X_1, X_2, \ldots, X_n$

  yksinkertainen satunnaisotos normaalijakaumasta

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

- Olkoon

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

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

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

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

Luottamusvälin johto 2/5

- Määritellään satunnaismuuttuja

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

. Voidaan osoittaa että t noudattaa *Studentin t-jakaumaa* vapausastein $(n - 1)$:

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

ks. lukua **Otokset ja otosjakaumat**.

Luottamusvälin johto 3/5

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

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

jolloin (*t*-jakauman symmetrian perusteella)

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

ja edelleen

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

Luottamusvälin johto 4/5

- Tarkastellaan epäyhtälöketjua

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

- Sijoittamalla tähän epäyhtälöketjuun satunnaismuuttujan t lauseke, saadaan epäyhtälöketju

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

- Tästä epäyhtälöketjusta saadaan sen kanssa *yhtäpitävä* epäyhtälöketju

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

Luottamusvälin johto 5/5

- Yhdistämällä saatu epäyhtälö siihen, että

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

  saadaan vihdoin

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

  Siten konstruoitu luottamusväli *peittää* parametrin μ todellisen arvon todennäköisyydellä (1 – α) ja se *ei peitä* parametrin μ todellista arvoa todennäköisyydellä α.

Luottamusvälin ominaisuudet

- Normaalijakauman odotusarvon μ luottamusvälin *keskipiste* $\bar{X}$ vaihtelee otoksesta toiseen.
- Luottamusvälin *pituus*

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

vaihtelee otoksesta toiseen.
- Luottamusvälin *pituus* riippuu valitusta luottamustasosta $(1 - \alpha)$, havaintojen lukumäärästä n ja otosvarianssista S^2 .

Vaatimukset luottamusvälille

- Olisi toivottavaa pystyä konstruoimaan odotusarvolle μ mahdollisimman *lyhyt* luottamusväli, johon liittyvä luottamustaso olisi mahdollisimman *korkea*.
- Vaatimusten samanaikainen täyttäminen *ei ole* kuitenkaan *mahdollista*, *jos otoskoko pidetään kiinteänä*:

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

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

Normaalijakauman odotusarvon luottamusväli

Esimerkki – 1/6

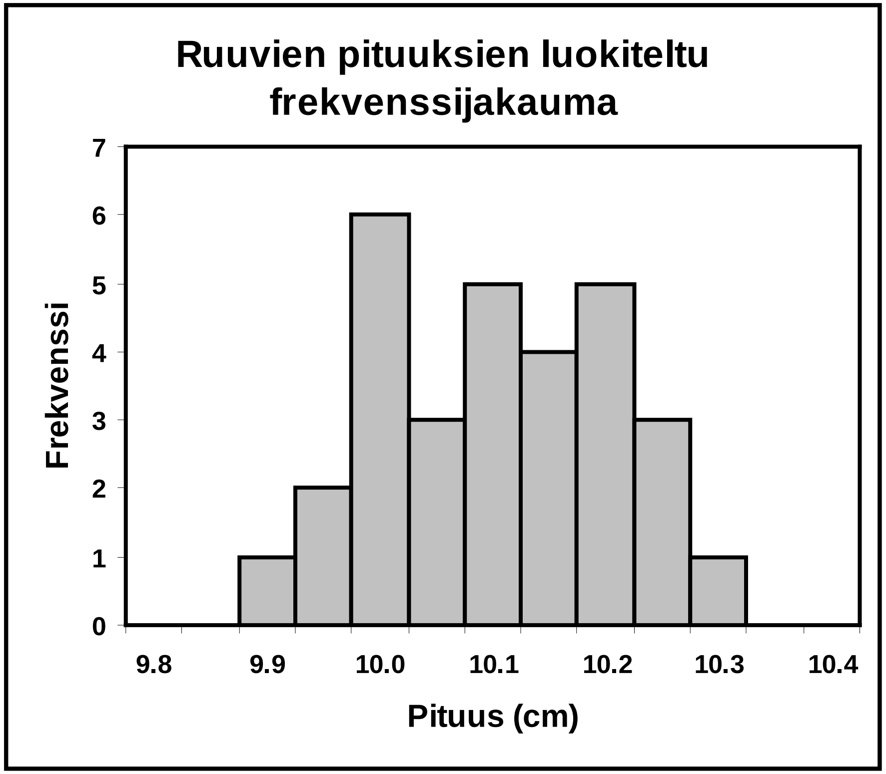
- Kone tekee *ruuveja*, joiden *pituudet vaihtelevat satunnaisesti noudattaen normaalijakaumaa*;
- Ruuvien joukosta poimitaan *yksinkertainen satunnaisotos*, jonka *koko* $n = 30$ ja otokseen poimittujen ruuvien pituudet mitataan.
- Taulukko oikealla esittää pituuksien *luokiteltua frekvenssi-jakaumaa*.

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

Normaalijakauman odotusarvon luottamusväli

Esimerkki – 2/6

- Kuva oikealla esittää otokseen poimittujen ruuvien pituuksien luokiteltua frekvenssijakaumaa vastaavaa *histogrammia*.
- *Luokkavälit* määräävät histogrammin suorakaiteiden *kannat*.
- Suorakaiteiden *korkeudet* on valittu niin, että suorakaiteiden *pinta-alat* suhtautuvat toisiinsa kuten vastaavat *luokka-frekvenssit*.



Normaalijakauman odotusarvon luottamusväli

Esimerkki – 3/6

Otoskoko:	$n = 30$
Pituuksien aritmeettinen keskiarvo:	$\bar{x} = 10.09$ cm
Pituuksien otoskeskihajonta:	$s = 0.1038$ cm

- Konstruoidaan ruuvien *todelliselle keskipituudelle* μ *luottamusväli luottamustasolla* 0.95.

- Valitaan *luottamuskertoimet* $-t_{0.025}$ ja $+t_{0.025}$ siten, että

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

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

Normaalijakauman odotusarvon luottamusväli

Esimerkki – 4/6

- *t*-jakauman *taulukoista* nähdään, että

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

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

  kun vapausasteiden lukumäärä

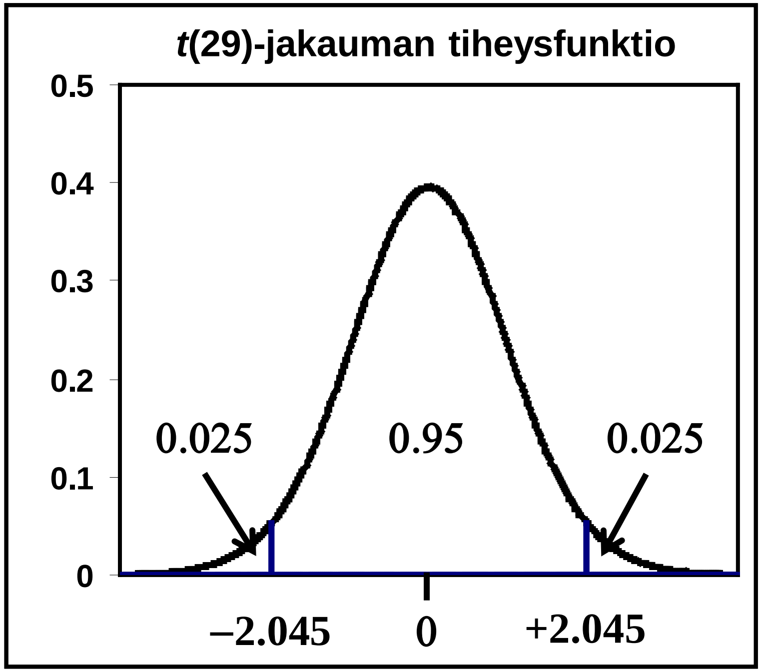
  $$n - 1 = 29$$

- Siten *luottamuskertoimet* ovat:

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

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

- Kuvio oikealla havainnollistaa luottamuskertoimien valintaa.



Normaalijakauman odotusarvon luottamusväli

Esimerkki – 5/6

- *Luottamusväliksi* saadaan:

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

- Siten tiedämme, että *ruuvien todellinen keskipituus on todennäköisyydellä* 0.95 *välillä*

  (10.05, 10.13)

Esimerkki – 6/6

- *Luottamustason* 0.95 *tulkinta*:

  Oletetaan, että poimimme koneen tekemien ruuvien joukosta *toistuvasti* yksinkertaisia satunnaisotoksia, joiden koko on 30 ja konstruoimme *jokaisesta* otoksesta 95 %:n luottamusvälin edellä esitetyllä menetelmällä.

  Tällöin:

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

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

Väliestimointi

Luottamusväli

Normaalijakauman odotusarvon luottamusväli

>> Normaalijakauman varianssin luottamusväli

Bernoulli-jakauman odotusarvon luottamusväli

Otos normaalijakaumasta

- Olkoon

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

  (*yksinkertainen*) *satunnaisotos* normaalijakaumasta

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

- Tällöin satunnaismuuttujat $X_1 , X_2 , \ldots , X_n$ ovat *riippumattomia* ja noudattavat *samaa normaalijakaumaa*

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

Normaalijakauman parametrien estimointi

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

  (i) *Odotusarvoparametrin* μ harhaton estimaattori:

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

  (ii) *Varianssiparametrin* σ^2 harhaton estimaattori:

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

Luottamustaso

- Määrätään **luottamusväli normaalijakauman varianssiparametrille** σ^2.
- Valitaan **luottamustasoksi**

$$1-\alpha$$

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

Luottamuskertoimet

- Olkoon valittu luottamustaso $(1 - \alpha)$.
- Määrätään **luottamuskertoimet** $\chi^2_{1-\alpha/2}$ ja $\chi^2_{\alpha/2}$ siten, että

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

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

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

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

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

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

Luottamuskertoimien määrääminen: Havainnollistus

- Luottamuskertoimet $\chi^2_{1-\alpha/2}$ ja $\chi^2_{\alpha/2}$ jakavat χ^2-jakauman tiheysfunktion kuvaajan alle jäävän todennäköisyysmassan (= 1) *kolmeen* osaan:

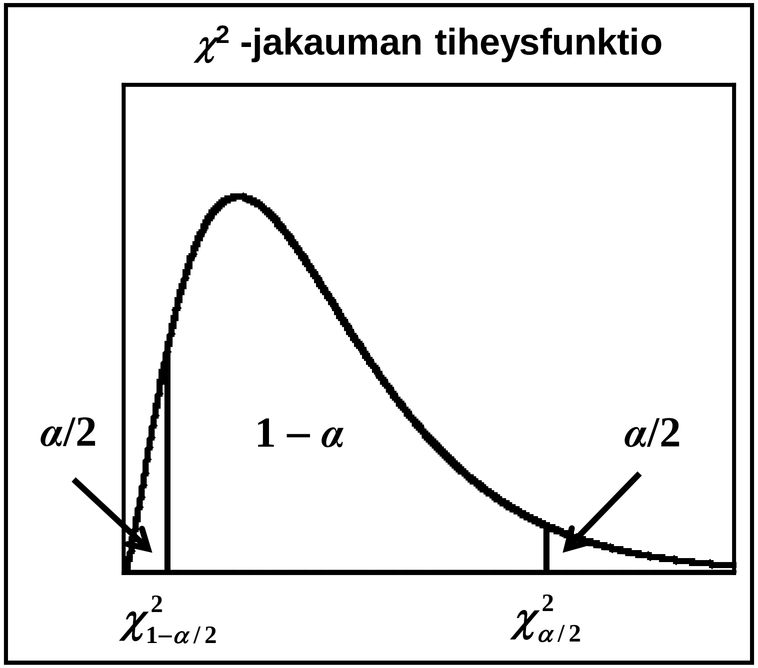
  (1) Pisteen $\chi^2_{1-\alpha/2}$ *vasemmalle* puolelle jää $\alpha/2$ % massasta.

  (2) Pisteen $\chi^2_{\alpha/2}$ *oikealle* puolelle jää $\alpha/2$ % massasta.

  (3) Pisteiden $\chi^2_{1-\alpha/2}$ ja $\chi^2_{\alpha/2}$ *väliin* jää $(1-\alpha)$ % massasta.

- Huomaa, että

$$\alpha/2 + \alpha/2 + (1-\alpha) = 1$$



Luottamusväli normaalijakauman varianssille

- Normaalijakauman **varianssiparametrin** σ^2 **luottamus-väli 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 = havaintojen harhaton *otosvarianssi*

n = havaintojen *lukumäärä*

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

= luottamustasoon $(1 - \alpha)$ liittyvät *luottamus-kertoimet* χ^2*-jakaumasta vapausastein* $(n - 1)$

Normaalijakauman varianssin luottamusväli

Luottamusvälin ominaisuudet 1/2

- Normaalijakauman varianssin σ^2 luottamusvälin *pituus*

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

  vaihtelee otoksesta toiseen.

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

Luottamusvälin ominaisuudet 2/2

- Luottamusväli *lyhenee* (*pitenee*), jos *luottamustasoa* $(1 - \alpha)$ *pienennetään* (*kasvatetaan*).
- Luottamusväli *lyhenee* (*pitenee*), jos *otosvarianssi* s^2 *pienenee* (*kasvaa*).
- Luottamusväli *lyhenee*, jos *otoskokoa* n *kasvatetaan*.

Vaatimukset luottamusvälille

- Olisi toivottavaa pystyä konstruoimaan varianssiparametrille 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*.

Väliestimointi

Luottamusväli
Normaalijakauman odotusarvon luottamusväli
Normaalijakauman varianssin luottamusväli
\>> Bernoulli-jakauman odotusarvon luottamusväli

Bernoulli-jakauma ja sen parametrointi 1/2

- Olkoon A kiinnostuksen kohteena oleva *tapahtuma* ja

$$\Pr(A) = p$$

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

- Määritellään satunnaismuuttuja

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

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

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

- Merkitään:

$$X \sim \mathrm{B}(p)$$

Bernoulli-jakauma ja sen parametrointi 2/2

- *Bernoulli-jakauman* B(p) *pistetodennäköisyysfunktio* on

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

$$0 < p < 1$$

$$q = 1 - p$$

Bernoulli-jakauman odotusarvoparametrin estimointi 1/2

- Olkoon

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

  (*yksinkertainen*) *satunnaisotos* Bernoulli-jakaumasta

  $$\mathrm{B}(p)$$

- *Estimoidaan* Bernoulli-jakauman $\mathrm{B}(p)$ odotusarvo-parametri p sen *harhattomalla estimaattorilla*:

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

Bernoulli-jakauman odotusarvoparametrin estimointi 2/2

- Koska

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

niin

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

jossa f on tapahtuman A *frekvenssi* otoksessa.

- Siten Bernoulli-jakauman parametrin p *estimaattori*

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

on tapahtuman A *suhteellinen frekvenssi* otoksessa.

Luottamustaso

- Määrätään **luottamusväli Bernoulli-jakauman odotus-arvoparametrille** p.
- Valitaan **luottamustasoksi**

$$1-\alpha$$

- Luottamustaso kiinnittää todennäköisyyden, jolla konstruoitava luottamusväli peittää Bernoulli-jakauman parametrin p todellisen arvon.

Luottamuskertoimet

- Olkoon valittu luottamustaso $(1 - \alpha)$.
- Määrätään **luottamuskertoimet** $-z_{\alpha/2}$ ja $+z_{\alpha/2}$ siten, että

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

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

  jossa satunnaismuuttuja z noudattaa *standardoitua normaalijakaumaa*:

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

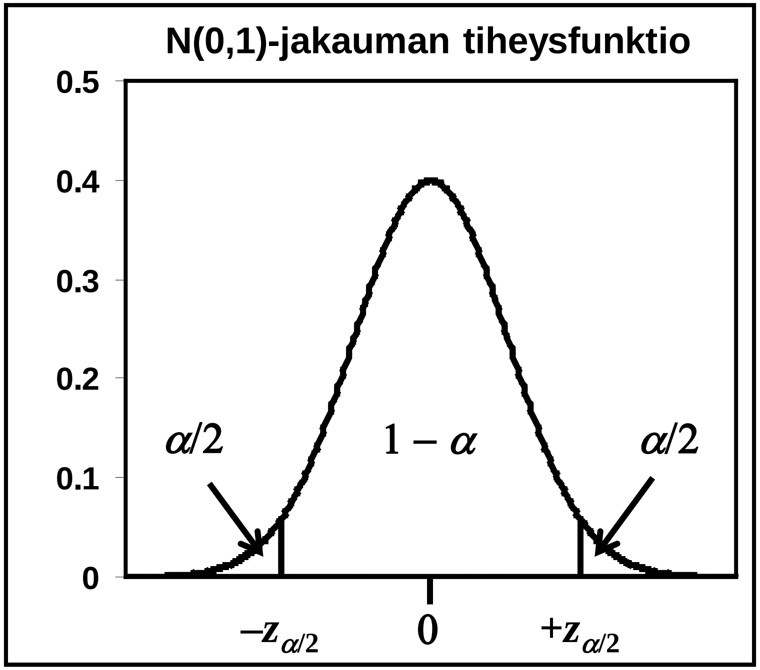
- Luottamuskertoimet $-z_{\alpha/2}$ ja $+z_{\alpha/2}$ toteuttavat ehdon

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

Luottamuskertoimien määrääminen: Havainnollistus

- Luottamuskertoimet $-z_{\alpha/2}$ ja $+z_{\alpha/2}$ jakavat normaalijakauman tiheysfunktion kuvaajan alle jäävän todennäköisyysmassan (= 1) *kolmeen* osaan:
  (1) Pisteen $-z_{\alpha/2}$ *vasemmalle* puolelle jää $\alpha/2$ % massasta.
  (2) Pisteen $+z_{\alpha/2}$ *oikealle* puolelle jää $\alpha/2$ % massasta.
  (3) Pisteiden $-z_{\alpha/2}$ ja $+z_{\alpha/2}$ *väliin* jää $(1 - \alpha)$ % massasta.
- Huomaa, että

$$\alpha/2 + \alpha/2 + (1 - \alpha) = 1$$



Luottamusväli Bernoulli-jakauman odotusarvolle

- 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

$\hat{p}$ = parametrin *p harhaton estimaattori*

n = havaintojen *lukumäärä*

$-z_{\alpha/2}\ ,\ +z_{\alpha/2}$

= luottamustasoon $(1-\alpha)$ liittyvät *luottamus-kertoimet normaalijakaumasta* N(0,1)

Luottamusväli ja sen pituus 1/2

- Bernoulli-jakauman odotusarvoparametrin p approksimatiivinen luottamusväli luottamustasolla $(1-\alpha)$ esitetään usein muodossa

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

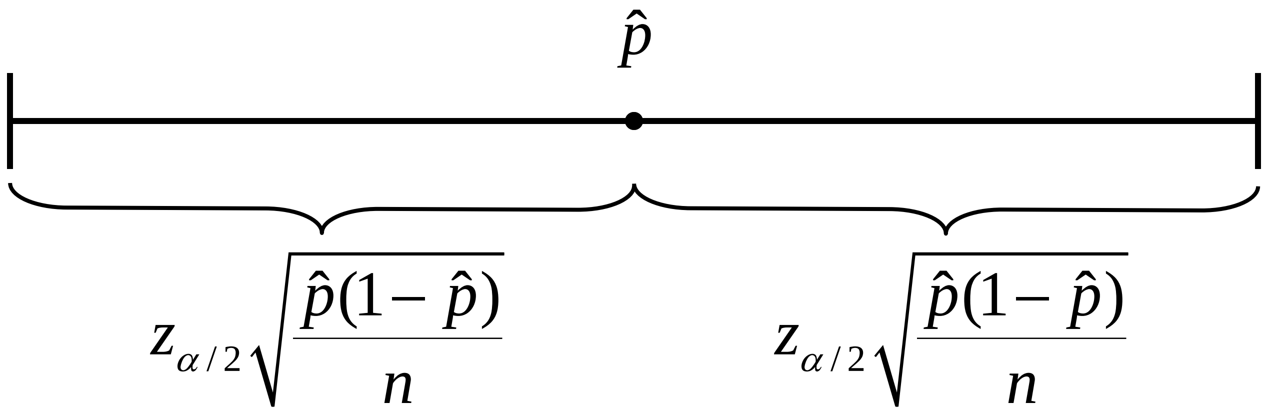
koska väli on symmetrinen keskipisteensä $\hat{p}$ suhteen.

- Bernoulli-jakauman parametrin p approksimatiivisen **luottamusvälin pituus** on

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

Luottamusväli ja sen pituus 2/2

- Bernoulli-jakauman odotusarvoparametrin p approksimatiivinen luottamusväli luottamustasolla $(1 - \alpha)$:



Luottamusvälin johto 1/5

- Olkoon

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

  yksinkertainen satunnaisotos Bernoulli-jakaumasta

  $$\mathrm{B}(p)$$

- Olkoon

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

  harhaton estimaattori parametrille p.

- Huomaa, että $\hat{p}$ on havaintojen $X_1 , X_2 , \dots , X_n$ *aritmeettinen keskiarvo*.

Luottamusvälin johto 2/5

- Määritellään satunnaismuuttuja

$$z = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}}$$

- Satunnaismuuttuja *z* noudattaa Bernoulli-jakautuneen otoksen *suhteellisen osuuden otosjakaumaa koskevan tuloksen perusteella suurissa otoksissa approksimatiivisesti standardoitua normaalijakaumaa* N(0,1) :

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

Luottamusvälin johto 3/5

- Määrätään normaalijakaumasta N(0,1) piste $+z_{\alpha/2}$ siten, että

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

jolloin (normaalijakauman symmetrian perusteella)

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

ja edelleen

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

Luottamusvälin johto 4/5

- Tarkastellaan epäyhtälöketjua

$$-z_{\alpha/2} \le z \le +z_{\alpha/2}$$

- Sijoittamalla tähän epäyhtälöketjuun satunnaismuuttujan z lauseke, saadaan epäyhtälöketju

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

- Tästä epäyhtälöketjusta saadaan sen kanssa *yhtäpitävä* epäyhtälöketju

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

Bernoulli-jakauman odotusarvon luottamusväli

Luottamusvälin johto 5/5

- Yhdistämällä saatu epäyhtälö siihen, että

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

saadaan vihdoin

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

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

Luottamusvälin ominaisuudet 1/2

- Bernoulli-jakauman parametrin p luottamusvälin *keskipiste* $\hat{p}$ vaihtelee otoksesta toiseen.
- Luottamusvälin *pituus*

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

  vaihtelee otoksesta toiseen.
- Luottamusvälin *pituus* riippuu valitusta luottamustasosta $(1 - \alpha)$, havaintojen lukumäärästä n ja estimaattorista $\hat{p}$.

Luottamusvälin ominaisuudet 2/2

- Luottamusväli *lyhenee* (*pitenee*), jos *luottamustasoa* $(1-\alpha)$ *pienennetään* (*kasvatetaan*).
- Luottamusväli *lyhenee*, jos *otoskokoa n kasvatetaan*.
- Luottamusväli on *lyhyimmillään*, kun

$$\hat{p} \approx 0 \text{ tai } \hat{p} \approx 1$$

- Luottamusväli on *pisimmillään*, kun

$$\hat{p} = \frac{1}{2}$$

Vaatimukset luottamusvälille

- Olisi toivottavaa pystyä konstruoimaan parametrille p mahdollisimman *lyhyt* luottamusväli, johon liittyvä luottamustaso olisi samanaikaisesti mahdollisimman *korkea*.
- 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 1/4

- Monissa tutkimustilanteissa toivotaan, että parametrille voitaisiin muodostaa *mahdollisimman lyhyt* luottamusväli.
- Bernoulli-jakauman odotusarvoparametrin p luottamus-välin pituus riippuu otoskoosta siten, että *väli lyhenee*, *jos havaintojen lukumäärää n kasvatetaan*.
- Tämä odotusarvoparametrin p luottamusvälin ominaisuus mahdollistaa tietyin ehdoin otoskoon valinnan niin, että *luottamusvälistä tulee* (suurin piirtein) *toivotun mittainen*.
- Oletetaan siis, että Bernoulli-jakauman odotusarvo-parametrille p halutaan konstruoida luottamusväli, jonka *toivottu pituus* on

$$2A$$

Otoskoon määrääminen 2/4

- *Bernoulli-jakauma* **odotusarvoparametrin** p **luottamusväli luottamustasolla** $(1-\alpha)$ on muotoa

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

jossa

$\hat{p}$ = parametrin p *harhaton estimaattori*

n = havaintojen *lukumäärä*

$-z_{\alpha/2}\,,\,+z_{\alpha/2}$

= luottamustasoon $(1-\alpha)$ liittyvät *luottamus-kertoimet normaalijakaumasta* N(0,1)

Bernoulli-jakauman odotusarvon luottamusväli

Otoskoon määrääminen 3/4

- Voidaan otoskoko *n* ratkaista yhtälöstä:

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

jossa

$z_{\alpha/2}$ = luottamustasoon $(1-\alpha)$ liittyvä *luottamus-kerroin normaalijakaumasta* N(0,1)

p = odotusarvon *oletettu arvo*

n = *otoskoko*

$2A$ = *toivottu pituus* luottamusvälille

Otoskoon määrääminen 4/4

- Siten *tarvittava otoskoko* on

$$n = \left(\frac{z_{\alpha/2}\sqrt{p(1-p)}}{A} \right)^2$$

- Huomaa, että tarvittava otoskoko *saavuttaa maksiminsa*

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

kun $p = 1/2$

- **Huomautus:**

  Tarvittavan otoskoon kaavasta nähdään, että mitä lyhyempää luottamusväliä toivotaan, sitä suurempi otos on poimittava:

  Jos esimerkiksi luottamusvälin pituus halutaan puolittaa, pitää havaintoja kerätä 4 kertaa enemmän.