3 Otannasta

3.1 Otanta ja tilastollinen päättely

Jos tutkitaan koko perusjoukko, tehdään kokonaistutkimus.

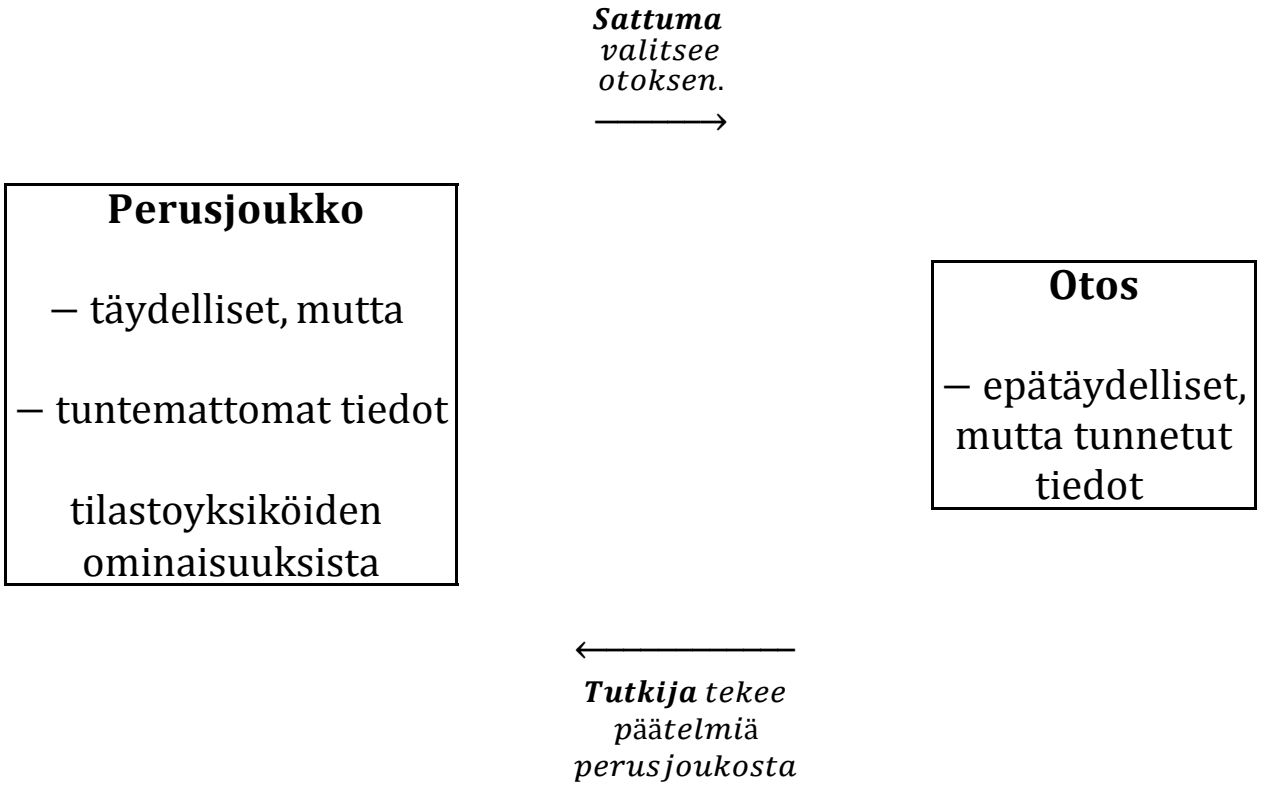
Tällainen on esimerkiksi poliittisen kannan jakauman selvittämiseksi pidettävät vaalit.

Kun perusjoukko on laaja, ei voida (kannata) tutkia kaikkia tilastoyksiköitä.

Silloin ainoa vaihtoehto on otantatutkimus, jonka vaiheet ovat

- otoksen poimiminen perusjoukosta ja

- perusjoukkoa koskevien päätelmien tekeminen otoksen informaation perusteella.



Otantatutkimuksessa **Sattumaa** tietoisesti hyväksi käyttäen valitaan tutkittavaksi perusjoukon osajoukko.

- **Kokonaistutkimuksesta** saatava tieto kuvaa periaatteessa täysin tarkasti, minkälaisessa tilassa perusjoukko on.

- **Otantatutkimuksen** tulokset ovat vain arvioita perusjoukosta.

Usein otantatutkimus on kuitenkin käytännössä ainoa ja parempikin vaihtoehto:

- Jos perusjoukko on ääretön, kokonaistutkimusta ei voi tehdä.

- Kokonaistutkimus vaatii usein liikaa resursseja.

- Jos mittauksia tehtäessä tilastoyksikkö tuhoutuu, tutkittavan määrän on oltava pieni.

- Kun perusjoukko on suuri, resurssit voidaan käyttää monipuolisemmin. Tiedon laatu korvaa määrän.

Eduista huolimatta vastattavaksi jää:

- Otos on yleensä hyvin suppea osa perusjoukosta.

Esimerkiksi (vain/jopa) noin 2000 suuruinen otos kaikista suomalaisista kuluttajista markkinatutkimuksessa.

Voidaanko otoksesta havaittava ”tilanne” ylipäänsä yleistää perusjoukkoon?

Ainakin tämä asettaa hyvin suuret laatuvaatimukset havaintojen valinnalle perusjoukosta. Otantateoria tutkii näitä kysymyksiä.

- Jos voidaan, kuinka **tarkkoja** ja **luotettavia** nämä yleistykset ovat?

Tilastollisen päättelyn avulla vastataan tällaisiin

- perusjoukkoa koskeviin kysymyksiin

- otoksesta tiivistetyn informaation avulla.

Sattuma ”valitsee” otoksen sisällön ja vaikuttaa sitä kautta perusjoukon ominaisuuksista tehtäviin päätelmiin.

Kuitenkin tilastollisen päättelyssä on tavoitteena, että päätelmien

- **luotettavuuden** ↔ kuinka suurella varmuudella (kääntäen riskillä) arviot ovat tosia (vääriä)
- ja **tarkkuuden** ↔ ”virhemarginaali”

suuruus voidaan esittää lukuina.

Tätä varten on hallittava lainalaisuudet, jotka ohjaavat sattuman käyttäytymistä otantatilanteessa.

”Hyviä” otantamenetelmiä on useita, mutta …

Oikea mielikuva: ”rehelliset arpajaiset”

Tällöin

- otos on **edustava osa** perusjoukosta, ”perusjoukko pienoiskoossa” ja

- tulokset voidaan **yleistää** perusjoukkoon.

Yleistysten **tarkkuus** ja **luotettavuus** voidaan (usein) selvittää ja esittää numeerisesti.

Esim. Poimittavasta otoksesta saatava

puolueen X kannattajien, hyödykkeen Y käyttäjien, alkoholin mainoskiellon kannattajien yms.

- suhteellinen osuus ”pyörii” oikean arvon tuntumassa.

Lisäksi halutaan tietää ja voidaan selvittää

- kuinka suuri on (esim. mainoskiellon kannattajien) suhteellisesta osuudesta tehdyn arvion **virhemarginaali** (esim.) **95 % varmuudella**? (95 %:n luottamusväli?)

Otantateoria tutkii, mitkä lainalaisuudet säätelevät informaatiovirtaa, jonka sattuma arpoo otokseen.

Yksinkertaisin otantamenetelmä on **yksinkertainen satunnaisotanta** (YSO):

YSO:ssa kaikilla perusjoukon tilastoyksiköillä on yhtä suuri todennäköisyys tulla valituksi otokseen.

Periaatteessa:

- Jokaista perusjoukon tilastoyksikköä vastaava ”nimilappu pannaan hattuun” ja sieltä poimitaan umpimähkään otos. Tämä on hankalaa ja

käytännössä:

- perusjoukon tilastoyksiköistä tehdään **kehikko** eli luettelo, jossa kaikki tilastoyksiköt numeroidaan 1:stä alkaen.

- Otokseen tulevat tilastoyksiköt arvotaan kehikosta **satunnaislukujen** avulla. Tietokoneen avulla tehtävät (tai valmiiksi taulukkoihin arvotut) satunnaisluvut **simuloivat** eli jäljittelevät lappujen arpomista ”hatusta”.

YSO on yleensä osana myös muissa(?) otantamenetelmissä.

Oleellista on, että hallitaan sattuman käyttäytymistä otantatilanteessa säätelevä ”mekanismi”. Silloin menetelmä on **todennäköisyysotantaa**.

Tämä takaa **edustavuuden**, ja arvioiden **tarkkuus** ja **luotettavuus** voidaan selvittää.

Jos todennäköisyysotannan sääntöjä ei noudateta, poimittua perusjoukon osajoukkoa sanotaan **näytteeksi**, ja siitä saadut tulokset voivat olla pahasti harhaisia.

”Edustava” informaatio voi syntyä myös toisella tavalla:

Kokeellinen tutkimus

Kokeellisessa tutkimuksessa tutkitaan jonkin ilmiön lainalaisuuksia.

Esim. Miten kehiteltävä lääke vaikuttaa verenpaineeseen?

Tutkija asettaa tutkimustilanteen ja kontrolloi sitä aktiivisesti.

Koeyksiköt joutuvat jonkin **käsittelyn** kohteeksi ja niiden **reaktio** mitataan.

Päämäärä on sama kuin otannassa:

Halutaan **edustavaa tietoa** ilmiöstä.

Tätä varten tutkijalla ovat välineinä **satunnaistaminen** ja **toistaminen**.

Esim. Uuden verenpainelääkkeen vaikutusta tutkittaessa

Koehenkilöt arvotaan **koe-** ja **vertailuryhmään**.

(Ei esim. naiset toiseen ja miehet toiseen ryhmään)

Tällä (**satunnaistamisella**) pyritään systemaattisten virheiden eliminoimiseen.

Toistamalla koe ”riittävän(?) monelle” koehenkilölle pienennetään satunnaisten virheiden vaikutusta.

Vaikka kokeellisessa tutkimuksessa informaatio syntyy eri tavalla kuin otannassa, koejärjestelyillä pyritään saamaan ”edustavaa tietoa” ilmiöstä.

Tällöin voidaan käyttää samoja analyysimenetelmiä kuin otantatutkimuksessa.

Seuraavassa käsitellään ”mekanismia”, joka säätelee yksinkertaisessa satunnaisotannassa

- minkälaisia arvoja otoksesta laskettavat tunnusluvut (otoskeskiarvo $\bar{x}$, jonkin ominaisuuden suhteellinen osuus $\hat{p}$ jne.) tulevat saamaan ja millä todennäköisyyksillä,

kun otokseen (joskus tulevaisuudessa) osuvaa informaatiota tiivistetään.

3.2 Otantajakaumista

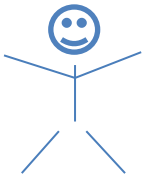
Otantatutkimus voidaan ajatella pelinä, jossa ovat vastakkain

- **tutkija**, joka yrittää paljastaa perusjoukon ”salaisuudet”, ja
- **Sattuma**, joka ”tietää perusjoukosta kaiken”.

Sattuma säätelee informaatiovirtaa perusjoukosta otokseen päin.

Tutkija taas yrittää ”seurata sattuman jälkiä toiseen suuntaan” ja tekee päätelmiä otoksesta perusjoukkoon päin.

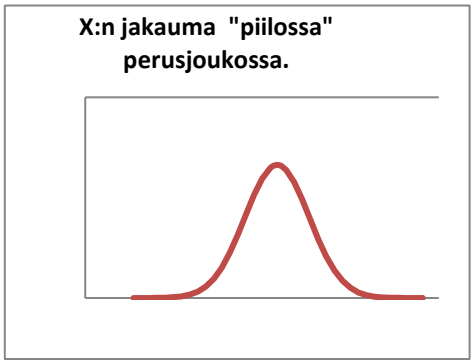
Perusjoukko



Sattuma on kunkku täällä.

”Sattuma tietää koko perusjoukon sisällön.”

μ = EX = todellinen keskiarvo
σ = DX = todellinen hajonta
….
jne. kaikista muuttujista



Sattuma *määrää otoksen sisällön.*

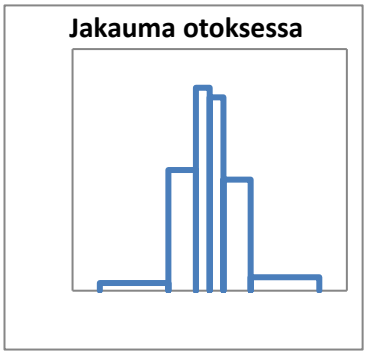
Otos ?

Tutkija

Sattuma arpoo tutkijalle muuttujan X jakaumasta arvot $x_1, x_2, \ldots, x_n$.

→ $\bar{x}$, s, …

”Rosoinen” kuva oikeasta jakaumasta:



↑

Satunnaiskoe ε = ”Perusjoukosta arvotaan tilastoyksikkö”.

Satunnaismuuttuja

X = tutkittavan muuttujan arvo valittavassa tilastoyksikössä

välittää informaation perusjoukosta otokseen.

Arvot valikoituvat vallitsevan jakauman mukaisesti

- sellaisia arvoja, joita on paljon perusjoukossa tulee paljon myös otokseen ja

- perusjoukossa harvinaisia arvoja tulee otokseen vähän.

Näin tapahtuu suurella **varmuudella**. Toisaalta on pieni **riski**, että näin ei käykään. ”Vain ihan sattumalta” otokseen voi osua aineisto, joka poikkeaa paljon siitä, miten keskimäärin pitäisi olla.

Todennäköisyyslaskennan tehtävä on juuri sen selvittäminen,

kuinka suurella varmuudella ja kuinka suurella riskillä?

Sattuman suhde tutkijaan on kuitenkin asiallisen neutraali. Se ei pyri johtamaan harhaan, mutta ei myöskään pyri erityisesti suosimaan tutkijaa.

Se noudattaa todennäköisyyslaskennan teoriassa kirjattuja sääntöjään.

Jos otantamenetelmä ottaa huomioon nämä lainalaisuudet, voidaan luottaa, että otoksessa nähdään ”perusjoukko pienoiskoossa”.

Tutkija saa (tutkimusresurssiensa rajoitusten puitteissa) otokseen niukasti mutta edustavaa tietoa perusjoukosta. Informaatio tiivistetään erilaisiksi tunnusluvuiksi (otossuureiksi).

Tiivistettyä informaatiota hyväksi käyttäen tilastollisen päättelyn menetelmien avulla päätellään, millaisia perusjoukon tilastoyksiköiden ominaisuudet ”keskimäärin” ovat.

Jotta todella realisoituneen otoksen perusteella voidaan tehdä johtopäätöksiä taustalla olevasta perusjoukosta, on tunnettava ”mekanismi”, jonka mukaan sattuma ”yleisesti ottaen” määrää otoksen sisällön.

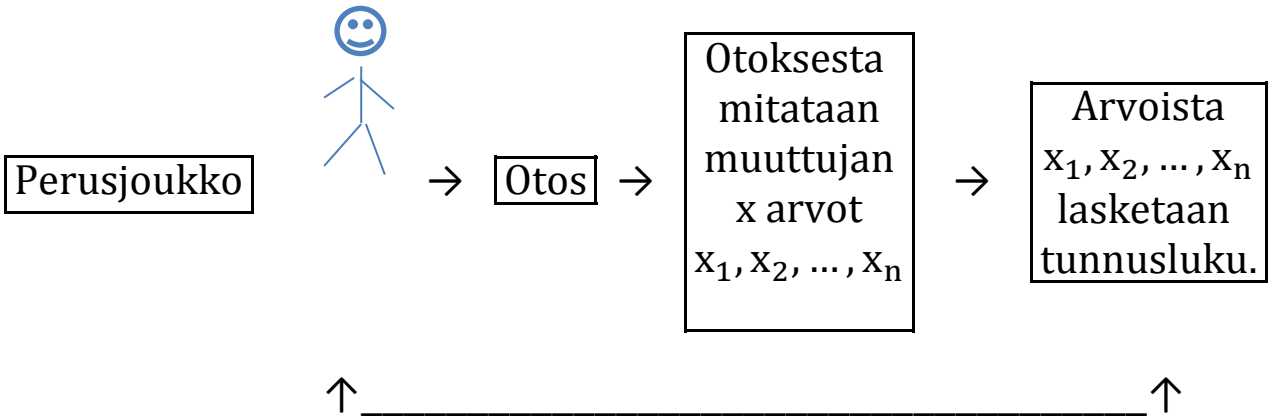
Otokseen osuvia muuttujan x arvoja ei tarkastella vain yksitellen, vaan

arvojen sisältämä informaatio tullaan tiivistämään tunnusluvuiksi, kun otos (joskus tulevaisuudessa) poimitaan.

- Sattuma määrää muuttujan arvot.

- Silloin sattuma määrää myös otoksesta laskettavien tunnuslukujen (otoskeskiarvo $\bar{x}$, keskihajonta s, suhteellinen osuus $\hat{p}$ jne.) arvot,

- joten **otoksesta laskettavat tunnusluvut ovat satunnaismuuttujia**.



Näkökulma on siis tässä spekulatiivinen:

”Jos otos joskus tulevaisuudessa tullaan poimimaan, niin kuinkahan suuri esimerkiksi otoskeskiarvo $\bar{x}$ tulee olemaan?”

- Perusjoukossa vallitseva jakauma, otantatapa ja satunnaismuuttujien yleiset ominaisuudet määräävä tunnusluvun **otantajakauman**.

Tässä käsitellään tarkemmin otoksesta laskettavan keskiarvon $\bar{X}$ ja suhteellisen osuuden $\hat{P}$ otantajakaumia:

Keskimääräisen suuruuden otantajakauma

Esim. (jatkoa) Tehtaan tuottamien suklaalevyjen (perus-)joukossa

levyn paino $X \sim N(100\text{ g}, (4\text{ g})^2)$.

Tuotannosta aiotaan poimia 20 suuruinen otos.

Sattuma tulee määräämään jokaisen poimittavan levyn painon perusjoukossa vallitsevan normaalisen jakauman mukaisesti.

Silloin

1. otokseen osuvan levyn paino $X_1 \sim N(100\text{ g}, (4\text{ g})^2)$,

2. otokseen osuvan levyn paino $X_2 \sim N(100\text{ g}, (4\text{ g})^2)$,

…

20. otokseen osuvan levyn paino $X_{20} \sim N(100\text{ g}, (4\text{ g})^2)$.

Levyt arvotaan otokseen YSO:aa käyttäen, jolloin painoja kuvaavia satunnaismuuttujia X_i, i = 1, 2,…, 20, voidaan pitää riippumattomia.

Otokseen tulevan suklaan kokonaismäärän $T = X_1 + X_2 + \ldots + X_{20}$ jakauma on riippumattomien normaalisten muuttujien yhteenlaskuominaisuuden mukaan

↙ 20 kpl↘ ↙ 20 kpl↘

$T = X_1 + X_2 + \dots + X_{20} \sim N(100\text{ g}+100\text{ g}+\dots+100\text{ g},\ (4\text{ g})^2+(4\text{ g})^2+\dots+(4\text{g})^2)$

↑ ↑ ↑

3) 1) 2)

$= \mathbf{N(20\cdot 100\text{ g},\ 20\cdot(4\text{ g})^2)}$

$(= N(2000\text{ g},\ (17.9\text{ g})^2).)$

Keskimääräinen levyn paino $\bar{X}$ on kokonaismäärä/otoskoko.

painon hajonta σ perusjoukossa

$\bar{X} = \frac{X_1 + X_2 + \dots + X_{20}}{20} = \frac{T}{20}$ ↓

$= \mathbf{\frac{1}{20}} \cdot T \sim N(\mathbf{\frac{1}{20}} \cdot 20 \cdot 100\text{ g},\ \mathbf{(\frac{1}{20})^2} \cdot 20 \cdot (4\text{ g})^2) = N(\boxed{100\text{ g}},\ \frac{(\boxed{4\text{ g}})^2}{\boxed{20}})$

↗ ↑

keskipaino μ perusjoukossa otoskoko n

$= N(100\text{ g},\ 0.8\text{ g}^2) = N(100\text{ g},\ (0.8944\text{ g})^2)$

Tässä yhdistettiin spekulatiivisesti ennen otoksen poimimista otokseen (tulevaisuudessa) osuvaa informaatiota ja saatiin selville otokseen osuvien suklaalevyjen keskipainon $\bar{X}$ otantajakauma.

Tämän ”ohjesäännön” mukaan sattuma määrää otoksen sisällön:

- Sattuma ”pyrkii” asettamaan otokseen tulevan keskipainon $\bar{x}$ todellisen keskipainon μ = 100 g kohdalle.

- Hajonnan avulla mitattuna ”sattuman pelivara” otoskeskiarvon $\bar{x}$ suuruuden ”heiluttelemiseen” oikean keskipainon μ ympärillä on keskimäärin (vain) noin 0.9 g.

- Tilanteeseen liittyvien todennäköisyyksien laskemisessa voidaan käyttää mallina normaalijakaumaa.

Esimerkiksi, kuinka todennäköistä on, että keskipaino otoksessa tulee olemaan alle 1 g:n päässä oikeasta keskipainosta 100 g?

$$P(99 < \bar{X} < 101) = P\left(\frac{99-100}{0.8944} < \frac{X-100}{0.8944} < \frac{101-100}{0.8944}\right)$$

$$= P(-1.12 < Z < 1.12) = \phi(1.12) - \phi(-1.12) = = \phi(1.12) – (1 - \phi(1.12))$$

$$= 2\,\phi(1.12) - 1 = 2 \cdot 0.8686 - 1 = 0.7374.$$

- Siis noin 74 % varmuudella on odotettavissa, että (vain) 20 levyn otoksessa keskipaino $\bar{X}$ tulee olemaan näin lähellä todellista keskipainoa.

Esimerkissä erikoistapauksena johdettu tulos on voimassa yleisesti:

Jos tutkittavan muuttujan jakauma on perusjoukossa $X \sim N(\mu, \sigma^2)$,

niin n:n suuruisesta otoksesta saatavan keskiarvon $\bar{X}$ jakauma on

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

- otos poimitaan palauttaen

- tai perusjoukko on ääretön (käytännössä ”hyvin suuri”) tai sen kokoa ei tiedetä.

Jos otos poimitaan palauttamatta N:n suuruisesta perusjoukosta, niin jakauma on muuten sama, mutta mukaan tulee äärellisen perusjoukon korjaustekijä varianssia pienentämään:

Voidaan osoittaa(?), että silloin

$\boldsymbol{\bar{X} \sim N(\mu, \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1})}$.

Informaatiota otoksesta tiivistettäessä tunnuslukujen (otossuureiden) jakaumien muodostumisessa vaikuttaa voimakas ”pyrkimys normaalisuuteen”:

Vaikka edellisessä tilanteessa jakauma ei olisikaan normaalinen perusjoukossa,

otoskeskiarvon $\bar{X}$ otantajakaumaa koskevat edelliset tulokset ovat likimain voimassa, kun otoskoko n on ”suuri”.

Tällainen ”pyrkimys normaalisuuteen” esitetään

keskeisessä raja-arvolauseessa,

jonka oleellinen sisältö on:

Jos satunnaismuuttujat $X_1, X_2, \ldots, X_n$ ovat samoin jakautuneita ja riippumattomia, niin

satunnaismuuttujan $T = X_1 + X_2 + \ldots + X_n$ jakauma

”lähestyy”(?) normaalijakaumaa, kun n kasvaa.(?)

Tämä konvergenssi toteutuu ”hyvin yleisesti(?) voimassa olevilla ehdoilla”.

(**Ei kuitenkaan**: ”Satunnaismuuttujan X jakauma riippuu useista osatekijöistä, joten se on normaalinen.)

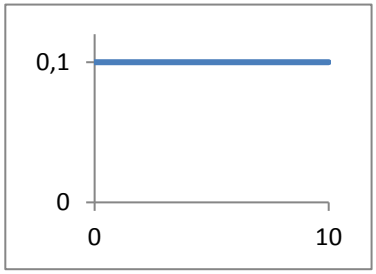
Esim. jatkoa) Linja-auton vuoroväli on 10 minuuttia.

Satunnaiskoe ε = ”Menet aikataulusta tietämättä pysäkille.” ja

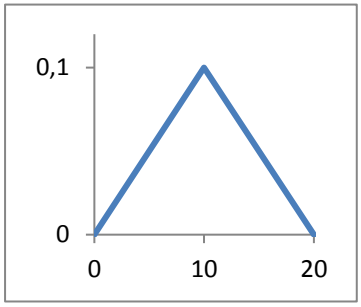
X_1 = odotusaika 1. kerralla, X_2 = odotusaika 2. kerralla, …

Kokonaisodotusajan $T = X_1 + X_2 + \ldots$ jakauma muuttuu kertojen kasvaessa:

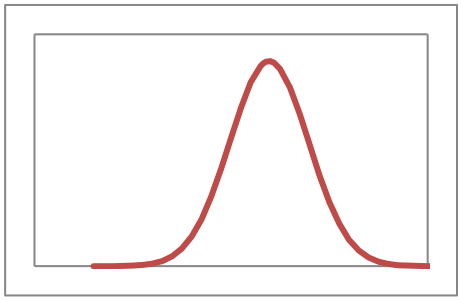
$X_1 \sim Tas(0,10)$ $\qquad\qquad$ $X_1 + X_2$



→



Hyvin nopeasti $T = X_1 + X_2 + \ldots$ →



- Käytännössä otoskokoa n pidetään ”riittävän suurena”, kun **n** $>$ **30**.

- Näin siis hallitaan ”mekanismi”, jonka mukaan sattuma määrää otoksessa realisoituvan keskiarvon $\bar{X}$ suuruuden.

- Ainoastaan ei-normaalinen muuttuja hyvin pienessä otoksessa jää käsittelemättä.

Sattuma toimii ohjesääntönsä mukaisesti.

Perusjoukko

muuttujan **X** todellinen keskiarvo $\boldsymbol{\mu}$ ja hajonta $\boldsymbol{\sigma}$

→

Otos

otoskeskiarvo $\bar{x}$

↑

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

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}\right).$$

Huom.

1) $\bar{X}$:n otantajakaumasta näkyy, että $E\bar{X} = \mu$,

joten otoskeskiarvo $\bar{X}$ ”pyrkii asettumaan” todellisen keskiarvon μ kohdalle.

Tästä seuraa, että tarkastelun suunta voidaan kääntää:

Otos todella poimitaan ja siitä laskettu keskiarvo $\bar{x}$ on ”keskimäärin” oikea arvio (estimaatti) tutkittavan muuttujan arvojen todelliselle (tuntemattomalle) keskimääräiselle suuruudelle μ perusjoukossa.

2) Otantajakauman varianssi

$$\frac{\sigma^2}{n} \text{ tai } \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

ja hajonta eli **keskiarvon keskivirhe** $\sigma_{\bar{X}}$ (←merkintä)

$$\frac{\sigma}{\sqrt{n}} \text{ tai } \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

mittaavat, kuinka paljon eri n:n suuruisista otoksista saatavat $\bar{X}$:n arvot voivat ”keskimäärin” vaihdella μ:n ympärillä.

- Odotettavissa oleva vaihtelu on sitä suurempaa, mitä suurempi hajonta σ on.

Tämä on **empiirisesti järkevää**:

Mitä erilaisempia muuttujan x arvot ovat perusjoukossa, sitä enemmän sattumalla on ”pelivaraa” tuottaa erisuuruisia keskiarvoja otokseen.

- Otoskoko n on keskivirheen $\boldsymbol{\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}}$ nimittäjässä, joten

suuresta otoksesta laskettava $\bar{X}$:n arvo poikkeaa todellisesta keskiarvosta μ ”keskimäärin” vähemmän kuin pienestä otoksesta laskettu.

Myös tämä on **empiirisesti järkevää**:

Jos keskiarvon laskemiseen on käytettävissä paljon informaatiota perusjoukosta, $\bar{x}$ ”vastaa paremmin” todellista keskiarvoa.

- Kun otos poimitaan palauttamatta äärellisestä perusjoukosta, $\bar{X}$:n otantajakauman vaihtelu on pienempää, minkä kuvaa

äärellisen perusjoukon korjaustekijä $\boldsymbol{\frac{N-n}{N-1}} < 1$.

Tämäkin on **empiirisesti järkevää**:

Otannan edistyessä muuttujan arvojen vaihtelu perusjoukossa pienenee koko ajan, kun muuttujan arvojen määrä vähenee. Silloin sattumalla on vähemmän pelivaraa otoskeskiarvon $\bar{X}$ ”heiluttelemiseen” todellisen keskiarvon μ ympärillä.

Siis otanta palauttamatta on (hieman) edullisempi kuin otoksen poimiminen palauttaen.

Vastaavat ominaisuudet ovat voimassa yleisemminkin otoksesta laskettaville tunnusluvuille.

Esim. (jatkoa) Tehtaan tuottamien suklaalevyjen (perus-)joukossa

levyn paino $X \sim N(100\ \text{g}, (4\ \text{g})^2)$.

Tuotannosta aiotaan poimia 80 suuruinen otos.

Perusjoukon kokoa ei tunneta, jolloin joudutaan toimimaan, kuin se olisi ääretön. Äärellisyyskorjausta ei voida käyttää hyväksi (!) otantajakauman varianssissa.

$$\bar{X} \sim N(100\ \text{g}, \frac{(4\ \text{g})^2}{80}) = N(100\ \text{g}, 0.2\ \text{g}^2) = N(100\ \text{g}, (0.4472\ \text{g})^2)$$

↑

Otoskoko on kasvatettava 4-kertaiseksi, jotta keskivirhe (sattuman ”pelivara”) pienenee puoleen! (Ks. aikaisemmasta, kun n = 20.)

Nyt todennäköisyys, että keskipaino otoksessa tulee olemaan alle 1 g:n päässä oikeasta keskipainosta 100 g, on

$$P(99 < \bar{X} < 101) = P\left(\frac{99-100}{0.4472} < \frac{\bar{X}-100}{0.4472} < \frac{101-100}{0.4472}\right)$$

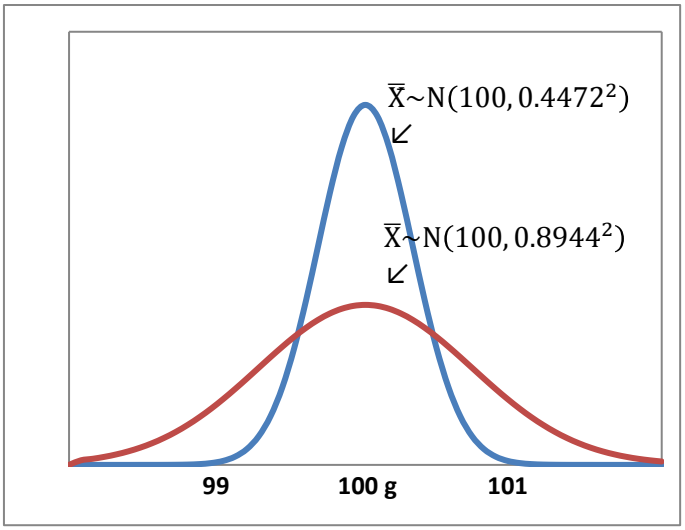
$$= P(-2.24 < Z < 2.24) = \phi(2.24) - \phi(-2.24) = \phi(2.24) - (1 - \phi(2.24))$$

$$= 2\,\phi(2.24) - 1 = 2 \cdot 0.9875 - 1 = 0.975$$

- Siis otoskeskiarvo $\bar{X}$ ”asettuu” peräti 97.5 % varmuudella näin lähelle todellista keskipainoa 100 g, jos otoskoko n = 80.

(Vrt. ”vain” 74 % varmuudella, jos otoksesta saatava informaatiomäärä on neljäsosa n = 20 tästä.)

Kuvassa on otoskeskiarvon otantajakauma, kun otoskoko n=20 ja n=80:



Esim. (jatkoa) Suklaatehdas väittää, että

levyn keskipaino μ = 100 g ja painon hajonta σ = 4 g.

50 suuruisessa otoksessa saatiin keskipainoksi vain $\bar{x}$ = 98.5 g.

Uskallatko väittää suklaatehtailijaa tämän perusteella huijariksi?

- Jos tehtaan väite ((?)nollahypoteesi H_0) μ = 100 g

- olisi tosi,

- niin otoksesta laskettavan keskiarvon otantajakauma, jonka mukaan sattuma tällaisessa otantatilanteessa silloin toimisi, olisi ollut

$\bar{X} \sim N(100\text{ g}, \frac{(4\text{ g})^2}{50}) = N(100\text{ g}, 0.32\text{ g}^2) = N(100\text{ g}, (0.5657\text{ g})^2)$.

- Kuinka todennäköistä on, että tällainen sattuman toimintaa säätelevä mekanismi tuottaisi ”vain sattumalta”

otoksesta havaitun 1.5 g tai vielä enemmän todellisesta keskipainosta poikkeavan otoskeskiarvon?

- Suklaatehtaan tuotteiden laatua ei ole koskaan epäilty.

Nyt kuitenkin joudutaan pohtimaan, onko otoksessa havaittu 1.5 gramman keskimääräinen alipaino vain sattuman leikkiä vai ei.

Silloin on kohtuullista pitää tarkastelussa mukana mahdollisuus, että yhtä hyvin ”vain sattumalta” keskipaino $\bar{X}$ olisi voinut ”heilahtaa” yhtä paljon suurempaan suuntaan.

Silloin todennäköisyys, että vain sattumalta otokseen osuu tehtaan väitteen

(H_0:) $\mu = 100$ g

epäilyksen alaiseksi saattava havaitun suuruinen poikkeama 1.5 g keskipainossa,

on (ehdollinen todennäköisyys)

Spekuloidaan hypoteesilla, että tehtaan väite olisi totta

↓

$p = P(\bar{X} \leq 98.5 \text{ g tai } \bar{X} \geq 101.5 \text{ g} \mid \mu = 100 \text{ g})$

↑ ↑

Näin kävi otoksessa.

Näin olisi voinut yhtä hyvin käydä.

$= P(\frac{\bar{X}-100}{0.5657} \leq \frac{98.5-100}{0.5657}$ tai $\frac{\bar{X}-100}{0.5657} \geq \frac{101.5-100}{0.5657})$

$= P(Z \leq -2.65$ tai $Z \geq 2.65) = \phi(-2.65) + 1 - P(Z < 2.65)$

$= 1- \phi(2.65) + 1- \phi(2.65)$

$= 2\cdot(1 - 0.9960)$

$= 0.008$

Siis keskimäärin vain 8 kertaa 1000:sta näin käy vain sattumalta.

- Varsin hyvin perustein voidaan päätellä, että tehtaan väite ei ole totta.

- Kuitenkin silloin hyväksytään 0.8 %:n riski sille, että

- keskipaino onkin väitetty 100 g ja

- tehdas pystyy sen todistamaan (vaikkapa punnitsemalla 10 000 … levyä)

- ja sattuma on ”vain sattumalta” aivan otantajakauman puitteissa generoinut tällaisen otoksen

- ja tehdas haastaa herjaajan oikeuteen ja vaatii korvausta aiheutetusta vahingosta …

Siis päätöksenteko:

uskalletaanko lähtökohtana olevasta (nolla-)hypoteesista luopua vai ei,

on suhteutettava siihen, kuinka vakavia käytännön seurauksia tällaisesta ((?)hylkäämis-)virheestä seuraa.

Tässä tarkasteltiin tilannetta 2-suuntaisesti. Ajateltiin, että poikkeama väitettyyn 100 gramman keskipainoon voisi tulla sattumalta yhtä hyvin ylöspäin.

Jos tässä voitaisiin jostain syystä ennen otoksen poimimista olla ”täysin varmoja”,

”että keskipaino ei nyt ainakaan 100 grammaa suurempi voi olla”,

niin (hylkäämisvirheen) riski on puolta pienempi

$p = P(\bar{X} \leq 98.5 \text{ g } \cancel{\text{tai } \bar{X} \geq 101.5 \text{ g}} \mid \mu = 100 \text{ g}) = 0.004.$

Tällaiseen 1-suuntaiseen päättelyyn (testaamiseen (?)) on oltava todella vankat otoksen ulkopuolelta tulevat perusteet.

Tätä aihetta jatketaan myöhemmin.

Suhteellisen osuuden otantajakauma

seuraa varsin suoraan binomi- ja hypergeometrisen jakauman normaaliapproksimaatiosta:

Esim. (jatkoa) Aiotaan tehdä markkinatutkimus, jossa

- eräs taustatieto on, että tässä perusjoukossa on 45 % naisia.

- Otoskoko on n = 1000.

- Etukäteen halutaan tarkistaa, että ”riittävän suurella varmuudella” (mm.) naisten osuus tulee olemaan otoksessa ”lähellä” perusjoukossa olevaa 45 prosenttia.

Esimerkiksi

”kuinka varmasti ” eli kuinka suurella todennäköisyydellä

- naisten suhteellinen osuus poimittavassa otoksessa tulee poikkeamaan naisten todellisesta 45 %:n osuudesta perusjoukossa

korkeintaan 2 % - yksikköä eli on välillä [0.43, 0,47]?

Perusjoukko on suuri, jolloin naisten lukumäärä otoksessa

$X \sim \text{Bin}(1000, 0.45)$

ainakin likimain poimittiinpa otos palauttaen tai palauttamatta.

Jakaumasta voidaan laskea joko (Excelillä) suoraan tai normaaliapproksimaation avulla $P(430 \leq X \leq 470)$.

Toisena vaihtoehtona on, että siirrytään suoraan suhteellisiin osuuksiin:

$np = 1000 \cdot 0.45 = 450 > 5$ ja $n(1-p) = 1000 \cdot (1 - 0.45) = 550 > 5$, joten

X:n jakaumaa voidaan approksimoida normaalisella apumuuttujalla

$$Y \sim N(1000 \cdot 0.45, 1000 \cdot 0.45 \cdot (1 - 0.45)) = N(450, 247.5) = N(450, 15.732^2).$$

$$\hat{P} = \text{naisten suhteellinen osuus} = \frac{\text{naisten lkm otoksessa}}{\text{otoskoko}} = \frac{X}{1000}.$$

Koska normaaliapproksimaatio käy, on $\hat{P}$:n jakauma likimain

$$\hat{P} = \frac{X}{1000} \approx \frac{Y}{1000} \sim N\left(\frac{1}{1000} \cdot 450, \left(\frac{1}{1000}\right)^2 \cdot 247.5\right) = N(0.45, 0.0002475)$$

$$= N(0.45, (0.015732)^2).$$

$P(0.43 \leq \hat{P} \leq 0.47) = P\left(\frac{0.43-0.45}{0.015732} \leq \frac{\hat{P}-0.45}{0.015732} \leq \frac{0.47-0.45}{0.015732}\right)$

$= P(-1.27 \leq Z \leq 1.27) = \phi(1.27) - \phi(-1.27) = \phi(1.27) - (1 - \phi(1.27))$

$= 2\,\phi(1.27) - 1 = 2 \cdot 0.8980 - 1$

$= 0.796.$

- Siis noin 80 % varmuudella naisten osuus otoksessa tulee osumaan alle 2 % - yksikön päähän oikeasta arvosta 45 %.

- Toisaalta on noin 20 % suuruinen riski, että naisten osuus sattuukin olemaan kauempana 45 prosentista. Jos tällainen riski on liian suuri, sitä voi pienentää otoskokoa kasvattamalla. Silloin sattuman ”pelivara” pienenee.

Tämä näkyy selvästi(?) $\hat{P}$:n otantajakauman varianssista.

- Oikeastaan laskussa olisi pitänyt käyttää jatkuvuuskorjausta, mutta sen merkitys on tässä mitätön:

$P(430 \leq X \leq 470) \approx P(430 - ½ \leq Y \leq 470 + ½) = P(429.5 \leq Y \leq 470.5)$

$= P\left(\frac{429.5}{1000} \leq \hat{P} \leq \frac{470.5}{1000}\right) = P(0.4295 \leq \hat{P} \leq 0.4705) \approx P(0.43 \leq \hat{P} \leq 0.47)$

Edellä laskettiin

$$EX = np = 1000 \cdot 0.45 \quad Var(X) = np(1-p) = 1000 \cdot 0.45 \cdot (1-0.45)$$

$$\downarrow \qquad \downarrow$$

$$\hat{P} = \frac{X}{1000} \approx \frac{Y}{1000} \sim N\left(\frac{1}{1000} \cdot \mathbf{450}, \left(\frac{1}{1000}\right)^2 \cdot \mathbf{247.5}\right)$$

$$= N\left(\frac{1}{1000} \cdot 1000 \cdot 0.45, \left(\frac{1}{1000}\right)^2 \cdot 1000 \cdot 0.45 \cdot (1-0.45)\right)$$

naisten suhteellinen osuus perusjoukossa p

$1 - p$

$$\downarrow \quad \searrow \qquad \swarrow$$

$$= N\left(\boxed{\mathbf{0.45}}, \frac{\boxed{\mathbf{0.45}} \cdot \boxed{(\mathbf{1-0.45})}}{\boxed{\mathbf{1000}}}\right)$$

↑ otoskoko n

Suhteellisen osuuden otantajakaumalla on sama rakenne myös yleisesti:

Oletetaan, että niiden tilastoyksiköiden suhteellinen osuus, joilla on ominaisuus A, on perusjoukossa p = P(A).

Jos perusjoukosta poimittavan otoksen koko n on ”suuri”

eli $np > 5$ ja $n(1-p) > 5$,

niin otoksesta laskettavan suhteellisen osuuden $\hat{P}$ otantajakauma on likimain normaalinen.

- Jos otos poimitaan palauttaen tai

palauttamatta äärettömän (käytännössä hyvin) suuresta perusjoukosta tai perusjoukon kokoa ei tiedetä, on

suhteellisen osuuden $\hat{P}$ otantajakauma

$$\hat{P} \sim N\left(\mathbf{p}, \frac{\mathbf{p(1-p)}}{\mathbf{n}}\right)$$

- Jos otos poimitaan palauttamatta N:n suuruisesta perusjoukosta,

niin

$$\hat{P} \sim N\left(\mathbf{p}, \frac{\mathbf{p(1-p)}}{\mathbf{n}} \cdot \frac{\mathbf{N-n}}{\mathbf{N-1}}\right).$$

Rakenne on hyvin samanlainen kuin otoskeskiarvon jakaumalla.

Sattuma toimii ohjesääntönsä mukaisesti.

Perusjoukko		**Otos**
Ominaisuuden A suhteellinen osuus on p	→	A: n suhteellinen osuus on $\hat{p}$

↑

$$\hat{P} \sim N\left(p, \frac{p(1-p)}{n}\right) \text{ tai}$$

2) ↓ ↙ 3)

$$\hat{P} \sim N\left(p, \frac{p(1-p)}{n} \cdot \frac{N-n}{N-1}\right)$$

1) ↗

1) Otoksesta saatava suhteellinen osuus $\hat{P}$ ”pyrkii” asettumaan todellisen suhteellisen osuuden p kohdalle. (Vrt. $\bar{X} \leftrightarrow \mu$)

2) Otantajakauman varianssi (, josta käytetään merkintää) $\sigma^2_{\hat{P}}$ ja keskivirhe $\sigma_{\hat{P}}$ mittaavat, kuinka suuri on sattuman ”pelivara” otoksesta saatavan suhteellisen osuuden $\hat{P}$ ”heiluttelemiseen” oikean suhteellisen osuuden p ympärillä.

Ne rakentuvat kahdesta osasta:

X:n hajonta DX

↙

$$\sigma^2_{\hat{P}} = \frac{p(1-p)}{n} \text{ ja } \sigma_{\hat{P}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

↖ otoskoko

- Mitä lähempänä p on arvoa 0.5, sitä suurempi on(?) ”sattuman pelivara”. (Piirrä f(p) = p(1-p):n kuvaaja.)

- Jos otoskokoa n kasvatetaan, keskivirhe pienenee.

3) Otannassa palauttamatta N:n kokoisesta perusjoukosta saadaan mukaan äärellisen perusjoukon korjaustekijä.

Vaihtelu pienenee otannan edistyessä ja tämä pienentää keskivirhettä. Otantajakauma on tiiviimmin todellisen suhteellisen osuuden p ympärillä.

Esim. Kunnassa M puoluetta Ö kannatti edellisissä vaaleissa 35 % äänestäjistä.

Puolueen epäillään sotkeutuneen törkyiseen lahjusskandaaliin. Uudet vaalit ovat tulossa, ja paikallislehti teki otantatutkimuksen. Siinä selvitettiin (mm.), onko puolueen Ö kannatus pienentynyt.

Lehden kesätoimittaja päättää etukäteen väittää artikkelissaan lehdelle tärkeän Ö:n kannatuksen merkitsevästi (**) pienentyneen, jos otoksesta havaittava vaaleja pienempi kannatus voi tulla vain sattumalta alle 1 % todennäköisyydellä.

Kunnan 40000 äänestysikäisestä poimittiin 1500 suuruinen otos palauttamatta.

Otoksessa Ö:tä kannatti 32 % vastaajista.

On ”täysin selvää”, että Ö:N kannatus ei ainakaan ole kasvanut edellisistä vaaleista. Silloin toimittajan riski:n suuruus paikkansa pitämättömän väitteen esittämiseen on

1-suuntaisen satunnaisen kannatuksen ”heilahduksen” todennäköisyys 32 prosenttiin, vaikka kannatus edelleen olisikin 35 %.

Jos ((?)nollahypoteesi H_0:) kannattajien osuus olisi edelleen p = 0.35,

niin sattuma olisi generoinut otoksen otantajakauman

(Korjaustekijä pienentää varianssia noin 3.7 %.) ↓

$$\hat{P} \sim N\left(0.35, \frac{0.35(1-0.35)}{1500} \cdot \frac{40000-1500}{40000-1}\right) = N(0.35, 0.0001516 \cdot \mathbf{0.9625})$$

$= N(0.35, 0.0121^2)$ mukaan.

Silloin on todennäköisyys, että otoksessa ”vain sattumalta” korkeintaan 32 % kannattaa Ö:tä, vaikka todellisuudessa kannatus koko äänestäjäkunnassa olisikin 35 %:

$p = P(\hat{P} \leq 0.32) = P(\frac{\hat{P} - 0.35}{0.0121} \leq \frac{0.32 - 0.35}{0.0121})$

$= P(Z \leq -2.48) = \phi(-2.48) = 1 - \phi(2.48) = 1 - 0.9934$

$= 0.0066$

$=$ **0.66 % $<$ 1 %,**

joten päätössääntönsä mukaan toimittaja käy jutun tekoon.

Tässäkin on jo otantajakaumien tutkimisen ohella samalla ennakolta käsitelty hypoteesien testaamista. Aiheeseen syvennytään tarkemmin myöhemmin, kun ensin on käsitelty toista tilastollisen päättelyn osa-aluetta estimointiteoriaa.

Muiden otoksesta laskettavien tunnuslukujen (mediaani Md, otosvarianssi s^2, korrelaatiokerroin r, jne.) otantajakaumien tarkka käsittely sivuutetaan tässä.

4 Estimoinnista

Sattuma ”toimii” perusjoukosta otokseen päin ja tutkija päinvastaiseen suuntaan.

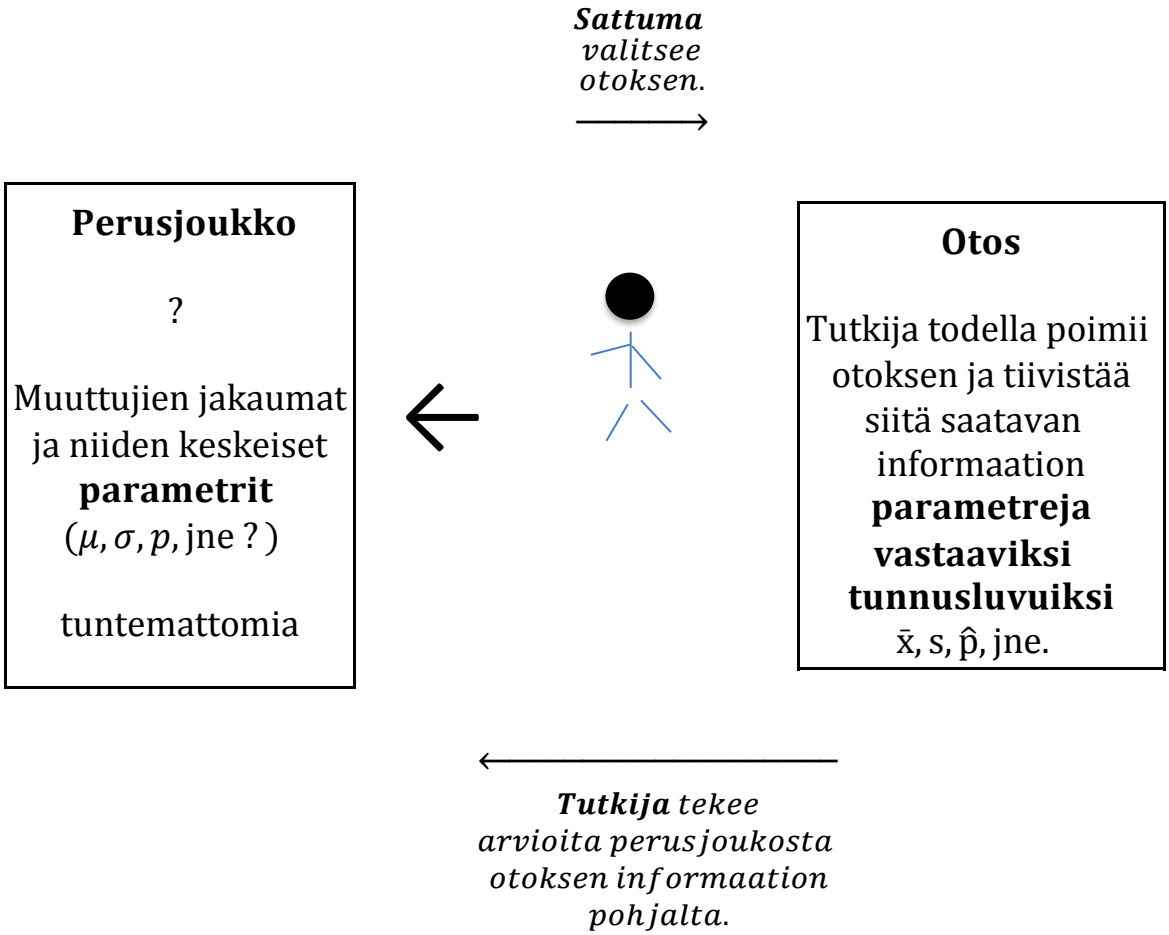
- Edellä on selvitetty, minkälaisten sääntöjen mukaan **informaatiovirta perusjoukosta otokseen** muodostuu.

Nyt käytössä ovat riittävät ”työkalut”, joiden avulla **tarkastelun suunta voidaan kääntää toisin päin** normaaliksi tutkimustilanteeksi keskimääräisen suuruuden μ ja suhteellisen osuuden p osalta:

- Todellinen keskiarvo μ (tai suhteellinen osuus p) on tuntematon.

- Otos todella poimitaan ja siihen osunut informaatio tiivistetään tunnusluvuiksi $\bar{x}$, s, jne. (tai $\hat{p}$).

- Tiivistetyn informaation avulla arvioidaan ”otantajakauman jälkiä takaisin perusjoukkoon päin seuraten”, mikä on μ:n (p:n) suuruus.



- Tutkija yleistää otoksesta havaitsemansa tulokset perusjoukkoon eli tekee **tilastollista päättelyä**, jossa välineinä ovat

- otoksesta tunnuslukuihin tiivistetty informaatio ja

- sen syntymekanismista tiedetyt säännöt (otantajakaumat).

Estimointiteoria on tilastollisen päättelyn tärkein osa-alue,

jonka menetelmien avulla voidaan tehdä

tarkkoja ja **luotettavia** arvioita perusjoukon keskeisistä ominaisuuksista

otokseen sisältyvän tiedon avulla.

Otoksessa on vain suppea osa perusjoukosta, joten perusjoukon tilastoyksiköiden ominaisuuksien arviointi eli **estimointi** ei voi kohdistua muuttujien yksittäisiin arvoihin vaan

niiden tilaa ”yleisesti ottaen” kuvaaviin suureisiin eli **parametreihin**.

Esim. Markkinatutkimuksessa hyödykkeen H käytöstä kotitalouksissa

tutkitaan:

- muuttujaa ”hyödykkeeseen H viikoittain käytetty rahamäärä”

Sitä kuvaavat varsin hyvin parametrit

μ = keskimäärin käytetty rahamäärä ja σ = käytetyn rahamäärän hajonta.

- H:n käytön yleisyyttä (käyttää/ei käytä) kuvaa parametri

p = H:n käyttäjien suhteellinen osuus.

- Minkälainen on H:n yksikköhinnan x_1 (selittävä muuttuja) ja kulutuksen määrän y (selitettävä muuttuja) yhteyttä kuvaava kysyntäfunktio?

Sopiiko malliksi edes likimain lineaarinen funktio $y = \beta_0 + \beta_1 x_1$?

Jos sopii, mitkä ovat parametrien β_0 ja β_1 arvot?

Paraneeko mallin selitysaste, jos malliin otetaan toiseksi selittäväksi muuttujaksi x_2 = vuositulo ja malliksi hahmotetaan

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$?

Mitkä ovat nyt β_0, β_1 ja β_2?

Otoksesta tiivistetyn informaation avulla tutkija **estimoi** (arvioi) parametrien arvot.

Estimoinnissa käytetään välineinä otoksesta lasket**tavia** tunnuslukuja eli **estimaattoreita**.

Otoksesta laske**tut** estimaattoreiden arvot ovat **estimaatteja**.

Jotta tunnusluku kelpaa estimaattoriksi,

- sen otantajakauman on keskityttävä estimoitavan parametri ympärille (tai edes lähelle). Silloin otoksesta saatava estimaatti ”pyörii” estimoitavan parametrin tuntumassa. Lisäksi

- otantajakauma on tunnettava, jolloin arvion **tarkkuus** voidaan selvittää.

Esim. Otoksesta laskettavan keskiarvon $\bar{X}$ otantajakauman perusteella tiedetään, että

- hyödykkeeseen H käytetyn rahamäärän todellisen keskiarvon μ estimaattorilla on ominaisuus $\mathbf{E\bar{X} = \mu}$
- eli estimaattori $\bar{X}$ ”tähtää suoraan kohti maalia”
- eli otoksesta saadaan ”keskimäärin” juuri oikea keskiarvo.

Estimaattoria, jonka odotusarvo on estimoitava parametri, sanotaan **harhattomaksi**.

Voidaan myös osoittaa(?), että (rahamäärän) todellisen varianssin σ^2 estimaattori $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$ on myös(?) harhaton eli $\mathbf{Es^2 = \sigma^2}$ (, mikä perustelee jakajassa olevan -1:n).

Otantajakaumasta tiedetään kuitenkin paljon enemmänkin:

Edellä nähtiin, että normaalijakauma käy ”melkein aina” malliksi ja

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ tai $\bar{X} \sim N(\mu, \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1})$ ainakin, jos otoskoko n on ”suuri”.

Perusjoukossa H:ta käyttävien todellisen suhteellisen osuuden p estimaattori $\hat{P}$ on myös harhaton ja sen otantajakauma on (likimain)

$\hat{P} \sim N(p, \frac{p(1-p)}{n})$ tai $\hat{P} \sim N(p, \frac{p(1-p)}{n} \cdot \frac{N-n}{N-1})$

Perusjoukossa (piilossa) olevan

todellisen keskimääräisen suuruuden μ, varianssin σ^2 ja suhteellisen osuuden p

estimoinnissa on estimaattorin valinta intuitiivisesti varsin selvää (paitsi ehkä n-1 s^2:n nimittäjässä?):

Perusjoukossa olevan parametrin suuruus arvioidaan vastaavalla otoksesta lasketulla arvolla. Yleisesti ottaen tilanne ei kuitenkaan ole aina yhtä ilmeisen selvä.

- On olemassa useita yleisiä periaatteita(?), joiden mukaan voidaan määritellä järkeviä estimaattoreita perusjoukon parametreille. Niiden selvittely sivuutetaan tässä.

- Usein parametria estimoitaessa on monia vaihtoehtoisia(?) estimaattoreita. Vaihtoehtojen paremmuutta voidaan vertailla estimaattoreiden yleisten ominaisuuksien(?) avulla, joista eräs on edellä määritelty harhattomuus. Myös tämä voidaan sivuuttaa tässä.

- Kun tilastotieteen alkeita käytetään, ei soveltajan tarvitse pohtia estimaattorien ominaisuuksia, vaan ne ovat suuntaviivoja uutta teoriaa kehitettäessä.

Edellä esimerkissä otoksesta laskettu

otoskeskiarvo $\bar{\mathbf{X}}$ on todellisen keskiarvon $\boldsymbol{\mu}$,

otosvarianssi $\mathbf{s^2}$ on todellisen varianssin $\boldsymbol{\sigma^2}$ ja

otoksesta havaittu suhteellinen osuus $\hat{\mathbf{p}}$ on todellisen suhteellisen osuuden **p**

yksittäinen arvio eli **piste-estimaatti**.

Jos tunnetaan estimaattorin otantajakauma, saadaan estimoitavasta parametrista paljon enemmän tietoa. Silloin voidaan ”mitata” tehdyn arvion **tarkkuus** ja **luotettavuus**

ja selvitetään,

millä välillä estimoitavan parametrin arvo on jollain etukäteen valittavissa olevalla varmuudella eli luottamustasolla.

Esim. (jatkoa edelliseen) Kun tutkitaan hyödykkeen H käyttöä, tiedetään otantajakaumien perusteella:

- Otoskeskiarvo $\bar{X}$ on todellisen keskiarvon μ, otosvarianssi s^2 on(?) todellisen varianssin σ^2 ja otoksesta havaittu suhteellinen osuus $\hat{p}$ on todellisen suhteellisen osuuden p paras piste-estimaatti.

- Lisäksi tiedetään, että estimaatit ovat

sitä suuremmalla **varmuudella lähellä** estimoitavan parametrin arvoa, mitä suurempi otoskoko on.

Mutta **kuinka lähellä** ja **kuinka suurella varmuudella?**

Tätä voidaan tutkia periaatteessa kahdella tavalla:

Esimerkiksi todellista (H:n kulutukseen käytetyn rahamäärän) keskiarvoa μ estimoitaessa

- voidaan laskea, kuinka suurella todennäköisyydellä otoksesta laskettu keskiarvo $\bar{x}$ ei poikkea todellisesta keskiarvosta vaikkapa yli 100 €

tai

- selvitetään, minkälainen väli sisältää todellisen keskiarvon vaikkapa 95 % varmuudella.

- Ensimmäisessä vaihtoehdossa arvion (näennäinen) **tarkkuus on vakio** ja tutkitaan, kuinka luotettava arvio silloin on.

- Jälkimmäisessä taas **vakiona pidetään varmuuden tasoa**, jolla päätelmät tehdään. Sen pohjalta tutkitaan, kuinka tarkka arvio eli kuinka laaja väli saadaan.

Arvioiden ja niistä seuraavien johtopäätösten tekemisessä järkevän varovaisuusperiaatteen vuoksi jälkimmäinen lähtökohta valitaan estimoinnin tarkastelunäkökulmaksi.

Tällaista

- otoksen informaation ja estimaattorin otantajakauman perusteella määrättävää

- väliä, joka ”peittää” estimoitavan parametrin arvon jollakin ennalta

valittavissa olevalla (suurella, esim. 95 %) varmuudella

sanotaan luottamus-(konfidenssi)väliksi.

Väliä määrättäessä käytettävää ”varmuuden astetta” sanotaan luottamustasoksi.

Luottamusvälin määräämistä sanotaan väli- tai intervalliestimoinniksi.

Otoksesta laskettavan keskiarvon $\bar{X}$ otantajakauman avulla saadaan selville

Todellisen keskimääräisen suuruuden μ luottamusväli

Otoksesta laskettavan keskiarvon otantajakaumana käy malliksi

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ tai $\bar{X} \sim N(\mu, \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1})$

ainakin likimain paitsi silloin, kun tutkittavan muuttujan jakauma perusjoukossa ei ole normaalinen ja otoskoko n on ”pieni” (< 30).

Koska

-otantajakauma on symmetrinen ja

- otoksesta saatava keskiarvo x̄ on paras piste-estimaatti μ:lle,

se on luonnollisesti määrättävän välin keskipiste.

Näin tutkitaan

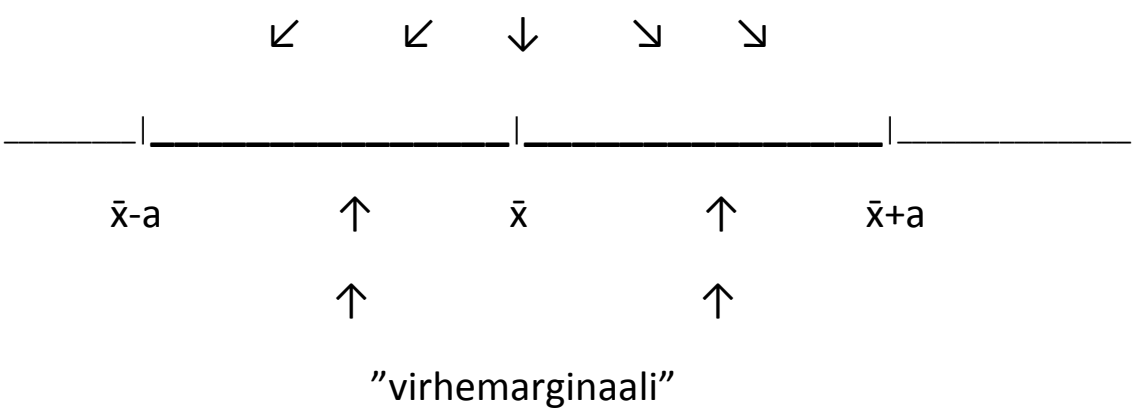
Kuinka suureksi on **”virhemarginaali”** a valittava,

jotta todellinen keskiarvo μ on välillä (x̄-a, x̄+a)

c:n (esimerkiksi 95 %) suuruisella **varmuudella** eli luottamustasolla?

Oikea μ:n arvo on jossain tällä välillä

c:n suuruisella varmuudella.



Keskimääräisen suuruuden μ luottamusvälin perusrakenne on kaikissa tapauksissa sama, mutta yksityiskohdat poikkeavat hieman.

Luottamusväliin vaikuttavat ainakin

- minkälainen tutkittavan muuttujan jakauma on perusjoukossa, erityisesti onko se normaalinen vai ei,

- kuinka suuri on otoskoko n,

- poimitaanko otos palauttaen vai palauttamatta, ja

- kuinka suuri on tutkittavan ominaisuuden hajonta σ perusjoukossa ja erityisesti, tiedetäänkö σ:n suuruutta edes.

- Luonnollisesti myös hajonta σ on yleensä aina tuntematon,

- kun estimoidaan keskimääräistä suuruutta μ, ja

- σ korvataan otoksesta laskettavalla estimaatillaan s.

Seuraavassa tutkitaan näitä eri tapauksia.

Jotta luottamusvälin rakenteen määräytymisen perusajatus saadaan esiin ottamatta vielä käyttöön lisää todennäköisyysjakaumia (t-jakauma(?)),

1.a) tässä aloitetaan tilanteesta, jossa (epärealistisesti) oletetaan, että muuttujan **todellisen hajonnan suuruus σ tiedettäisiin**!

- Osoittautuu, että lähtökohdan epärealistisuudesta huolimatta ”suuri” havaintojen määrä n tekee saatavasta tuloksesta käyttökelpoisen:

Esim. On havaittu, että tuotantoprosessin muutoksista huolimatta

- tuotteen kestoikä on likimain normaalinen

- ja hajonta $\sigma \approx 200$ h pysyy samana (!?)

- vaikka keskimääräinen kestoikä μ muuttuisikin.

Poimittiin n = 100 tuotteen otos, jossa keskimääräinen kestoikä oli $\bar{x} = 1500$ h (ja ”jostain syystä” ei laskettu keskihajontaa!?)

Mikä on todellisen keskimääräisen kestoiän μ 95 %:n luottamusväli?

Siis

Kuinka suureksi on ”virhemarginaali” a valittava, että todellinen

keskiarvo μ on välillä $(\bar{x}-a, \bar{x}+a) = (1500 \text{ h} - a, 1500 \text{ h} + a)$

95 %:n suuruisella varmuudella?

Jotta a:n suuruus voidaan laskea, on ”palattava alkuun takaisin” ja selvitettävä, minkä ”mekanismin” mukaisesti sattuma on tuottanut otoksessa havaitun tilanteen:

Jos voidaan luottaa siihen, että jakauma on normaalinen ja hajonta $\sigma \approx 200$ h,

niin otantajakauma, joka säätelee sattuman toimintaa tällaisessa otantatilanteessa, on

1) 2)

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) = N(\mu, \frac{(200\text{h})^2}{100}) = N(\mu, 400\ \text{h}^2) = \mathbf{N}(\boldsymbol{\mu}, (\mathbf{20\ h})^2).$$

3)

1) $E\bar{X} = \mu$, siis sattuma ”tähtää” oikean keskiarvon kohdalle.

2) Keskivirhe $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{200\ \text{h}}{\sqrt{100}} = 20$ h eli

tällaisessa 200 suuruisessa otoksessa ”sattuma voi keskimäärin heilutella” otoskeskiarvoa $\bar{X}$ noin 20 h todellisen keskiarvon μ ympärillä.

3) Todennäköisyyksien laskemisessa mallina voidaan käyttää normaalijakaumaa.

$\bar{X}$:n otantajakauman perusteella:

$0.95 = P(\bar{X} - a < \mu < \bar{X} + a)$ ← Tämä vaaditaan.

$= P(-a < \mu - \bar{X} < a)$ ← epäyhtälöiden

$= P(a > \bar{X} - \mu > -a)$ ← käsittelyä,

$= P(-a < \bar{X} - \mu < a)$ ← joka tähtää

$= P(\frac{-a}{\sigma/\sqrt{n}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{a}{\sigma/\sqrt{n}})$ ← standardointiin.

$= \phi(\frac{a}{\sigma/\sqrt{n}}) - \phi(\frac{-a}{\sigma/\sqrt{n}})$

$= \phi(\frac{a}{\sigma/\sqrt{n}}) - (1 - \phi(\frac{a}{\sigma/\sqrt{n}}))$

$= 2\,\phi(\frac{a}{\sigma/\sqrt{n}}) - 1$

Yhtälöstä

$2\,\phi(\frac{a}{\sigma/\sqrt{n}}) - 1 = 0.95$ ratkaistaan samalla tavalla kuin edellä

$\phi(\frac{a}{\sigma/\sqrt{n}}) = \frac{0.95+1}{2} = 0.975$

Toisaalta (Exelistä tai) taulukkoa ”toisin päin” käyttämällä nähdään

$\phi(1.96) = 0.975$, joten

$\frac{a}{\sigma/\sqrt{n}} = 1.96$, josta saadaan

”virhemarginaalin” suuruus a = 1.96 · $\frac{\sigma}{\sqrt{n}}$ ja

95 % luottamusväli on

$(\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}})$

Sillä on siis vaadittu ominaisuus

$P(\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}) = 0.95$

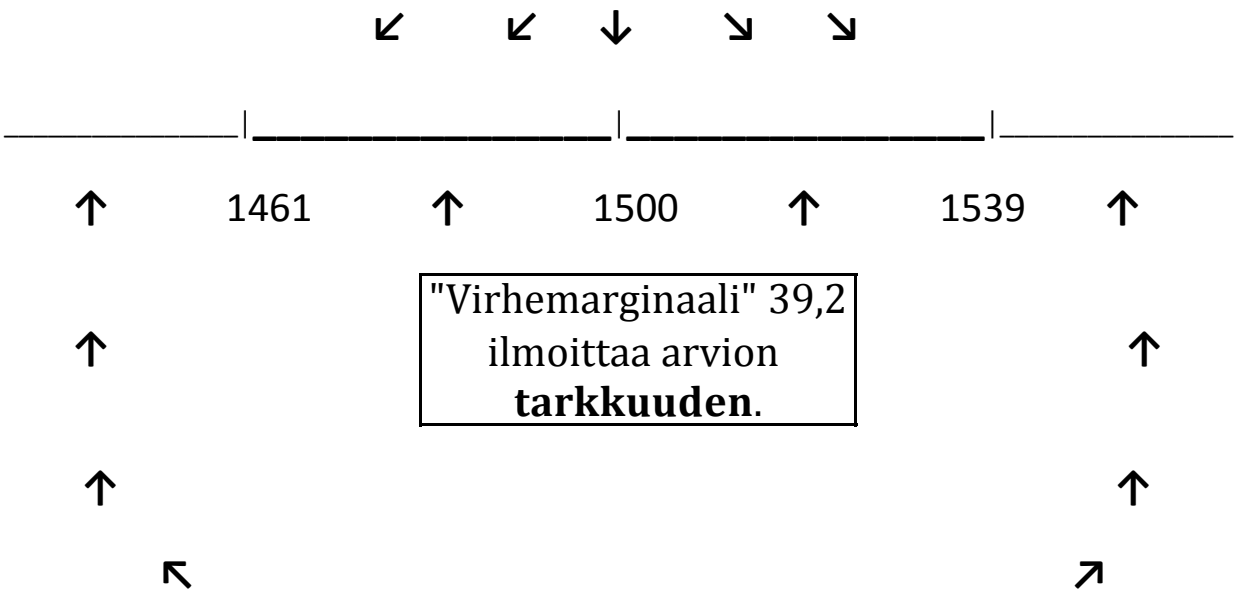
eli estimoitava kestoiän todellinen keskimääräinen suuruus μ on tällä välillä 95 % varmuudella.

100 suuruisesta otoksesta saatiin keskiarvoksi $\bar{x} = 1500$ h, ja tässä 95 % luottamusväli on

$(1500 - 1.96 \cdot \frac{200}{\sqrt{100}} < \mu < 1500 + 1.96 \cdot \frac{200}{\sqrt{100}}) = (1500 - 39.2, 1500 + 39.2)$

$\approx$ (1461 h, 1539 h) ($\approx$ (1460 h, 1540 h)).

Oikea todellinen keskimääräinen kestoikä μ
on jossain tällä valillä
95 %: n suuruisella **varmuudella**.



"Virhemarginaali" 39,2
ilmoittaa arvion
tarkkuuden.

Arviota tehtäessä hyväksytään
5 %: n **riski**,
että oikea μ: n arvo onkin välin ulkopuolella
"ihan vain sattumalta".

Esimerkissä havaitun mukaan yleisin merkinnöin (sääntöjen täsmällinen matemaattinen johtaminen sivuutetaan):

Oletetaan, että

- perusjoukko E on ääretön (käytännössä hyvin suuri) tai
- E:n kokoa ei tunneta (äärellisyyttä ei pystytä hyödyntämään) tai
- N:n kokoisesta perusjoukosta otos poimitaan palauttaen

ja tutkittava ominaisuus $\mathbf{X \sim N(\mu, \sigma^2)}$ **perusjoukossa** ja $\boldsymbol{\sigma}$ **tunnetaan**(!?).

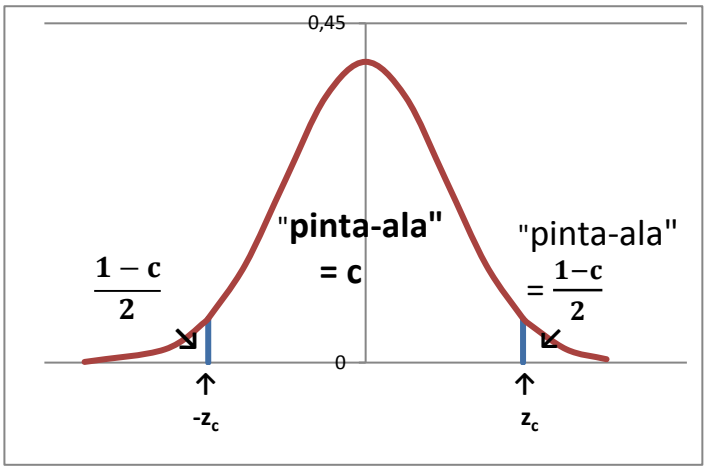
- Perusjoukosta poimitaan n:n suuruinen otos, josta lasketaan otoskeskiarvo $\bar{x}$ (ei keskihajontaa s (?!)).

Todellisen perusjoukossa olevan keskimääräisen suuruuden μ luottamusvälin luottamustasolla c päätepisteet ovat

$\bar{\mathbf{x}} \pm \mathbf{z_c} \frac{\boldsymbol{\sigma}}{\sqrt{\mathbf{n}}}$, missä

luottamuskerroin z_c saadaan standardoidusta normaalijakaumasta ehdosta:

$\mathbf{P(-z_c < Z < z_c) = c} \quad \leftrightarrow$



Kun luottamuskertoimen määräävästä ehdosta jatketaan laskua, saadaan kerroin aivan samalla tavalla kuin edellä esimerkissä.

Esim. (jatkoa)

Jos 95 % **varmuus** ei riitä, pitää luottamustasoa kasvattaa.

Sillä on kuitenkin **kova hinta**:

Kun luottamustasoksi valitaan 99 %, niin luottamuskerroin saadaan ehdosta

$0.99 = P(-z_{0.99} < Z < z_{0.99})$ Piirrä myös kuva!

$= \phi(z_{0.99}) - \phi(-z_{0.99})$

$= \phi(z_{0.99}) - (1 - \phi(z_{0.99}))$

$= 2\,\phi(z_{0.99}) - 1,$

josta saadaan kuten edellä

$\phi(z_{0.99}) = \frac{0.99+1}{2} = 0.955$

ja taulukosta ”toisin päin” $z_{0.99} \approx 2.58$ (Excelistä $z_{0.99} = 2{,}575829$).

Luottamusvälin päätepisteet ovat nyt

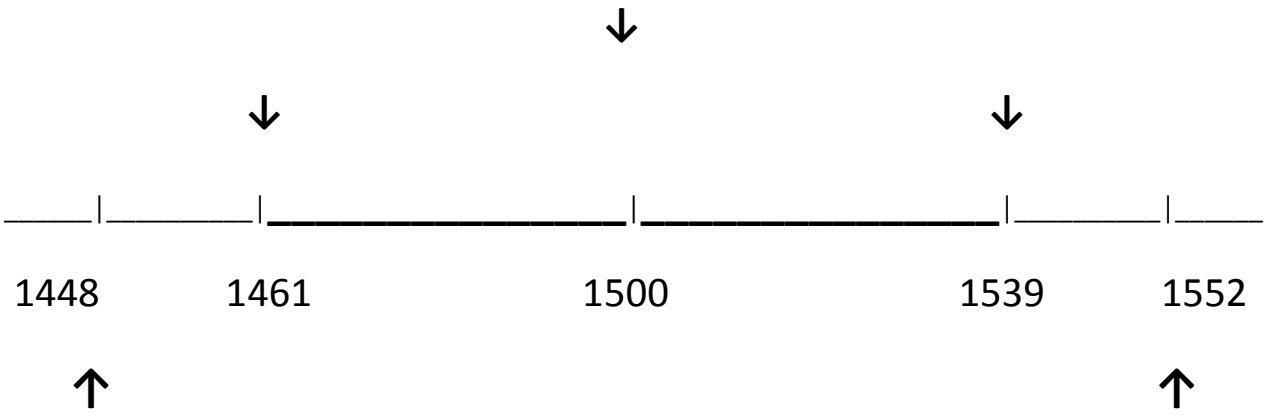
$1500 \pm 2.576 \cdot \frac{200\ \text{h}}{\sqrt{100}}$ = 1500 h $\pm$ 51.52 h ja

99 % luottamusväli todelliselle keskimääräiselle kestoiälle μ on

(1500 h - 51.52 h, 1500 h + 51.52 h) $\approx$ (1448 h, 1552 h).

Kun verrataan aikaisempaan, näkyy:

Oikea todellinen keskimääräinen kestoikä μ
on jossain tällä valillä "vain"
95 %: n suuruisella **varmuudella**.



Oikea todellinen keskimääräinen kestoikä μ
on tällä välillä "peräti"
99 %: n suuruisella **varmuudella**.

Siis

- luottamusväli levenee eli arvion **tarkkuus heikkenee**,

- kun luottamustasoa suurennetaan eli arvion **luotettavuus paranee**.

Nämä hyvät ominaisuudet ovat toisiinsa kytkeytyneitä ja tällä tavalla ”vaihdannaisia”.

Rajallisesta informaatiomäärästä ei voi ”puristaa” kuin rajallisen määrän ”hyvää”.

Se voidaan ”suunnata” joko arvion luotettavuuteen tai tarkkuuteen, mutta toisen parantaminen maksetaan toisen heikkenemisellä.

Kaikkialla tilastotieteessä vaikuttaa periaate ”Mitään ei saa ilmaiseksi.” Jos halutaan lisää varmuutta (pienentää riskiä), se on maksettava (näennäisen) tarkkuuden menettämisellä.

Jos ei kuitenkaan haluta tyytyä tähän, vaan

halutaan sekä hyvä luotettavuus että tarkka arvio,

nämä molemmat hyvät ominaisuudet ”voi ostaa”:

Esim. (jatkoa edelliseen)

Mikä on todellisen keskimääräisen kestoiän μ 99 %:n luottamusväli, jos

- tilanne olisi muuten sama kuin edellä eli perusjoukossa tuotteen kestoikä on normaalinen ja hajonta $\sigma \approx 200$ h tunnetaan (?!)

- mutta olisikin poimittu 400 suuruinen otos, jossa olisi saatu keskiarvoksi (vaikkapa) $\bar{x} = 1510$ h.

Nyt informaatiota on 4-kertainen määrä ja siitä saadaan enemmän ”hyvää”:

Olkoon luottamustaso c = 0.99 kuten edellä, jolloin luottamuskerroin $z_{0.99} = 2.576$ (≈ 2.58) tässäkin.

Luottamusvälin päätepisteet ovat

$$1510 \pm 2.576 \cdot \frac{200\ \text{h}}{\sqrt{\mathbf{400}}} = 1510\ \text{h} \pm 25.76\ \text{h} \cong 1510 \pm \mathbf{26\ h}$$

↑

99 % luottamustasosta tinkimättä ”virhemarginaali” on puolet siitä, mitä se oli 100 suuruisessa otoksessa. **Siis arvion tarkkuus on 2-kertainen!**

Kun otoskoko n kasvaa, keskivirhe $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, ”sattuman pelivara” pienenee.

Hinta on kuitenkin kova:

Informaation määrä otoksessa on (tässä) 4-kertaistettava, jotta tarkkuus 2-kertaistuu, kun luotettavuudesta ei tingitä.

Tutkimusresursseja (aikaa, vaivaa, rahaa) kuluu enemmän.

1.b) Jos **N:n suuruisessa perusjoukossa** tutkittavan ominaisuuden jakauman malliksi käy (Arvoja äärellinen määrä!)

$X \sim N(\mu, \sigma^2)$ ja **σ tunnetaan**(!?)

- ja perusjoukosta poimitaan n:n suuruinen otos **palauttamatta**, ja siitä lasketaan otoskeskiarvo $\bar{x}$ (ei keskihajontaa s (?!)).

Nyt otoskeskiarvon generoiva otantajakauma on

$\bar{X} \sim N(\mu, \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1})$.

Todellisen perusjoukossa olevan keskimääräisen suuruuden μ

luottamusvälin rakenne on samanlainen kuin edellä. Nyt vain keskivirheeseen saadaan mukaan äärellisen perusjoukon korjaustekijä tarkentamaan arviota:

Luottamustasolla c luottamusvälin päätepisteet ovat

$\bar{x} \pm z_c \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$, missä

luottamuskerroin z_c saadaan samasta ehdosta kuin edellä:

$P(-z_c < Z < z_c) = c$.

Korjaustekijä $\sqrt{\frac{N-n}{N-1}} < 1$ ja se pienentää sattuman ”pelivaraa” otoskeskiarvon $\bar{x}$ määräämisessä. **Arvio on tarkempi kuin otannassa palauttaen.**

1.c) Edellä oli vaatimuksena, että tutkittava muuttuja on normaalinen perusjoukossa. Kaikki muuttujat **eivät** kuitenkaan ole **normaalisia**, mutta tässäkin **keskeinen raja-arvolause** tulee apuun:

Edellä todettiin (todistamatta), että otoskeskiarvon otantajakauma on (hyvin erikoisia erikoistapauksia lukuun ottamatta(?)) ainakin likimain

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ tai } \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}\right)$$

vaikka tutkittava muuttuja ei olisikaan normaalinen perusjoukossa,

jos otoskoko n on ”suuri”, mille käytännössä rajana pidetään

(varsin pientä otoskokoa) $\mathbf{n > 30}$.

Luottamusvälin rakenne seuraa otantajakaumasta, joten edellä olevia menetelmiä saa käyttää (riittävän tarkkoina approksimaatioina), vaikka muuttuja ei olisikaan normaalinen.

2. a) Rasitteena on edelleen epärealistinen oletus, että tutkittavan muuttujan X todellinen hajonta σ tiedettäisiin.

- Tällainen oletus on kyllä hyödyllinen esimerkiksi

optimaalista otoskokoa(?) arvioitaessa, jos käytettävissä on edes jonkinlainen järkevä ”arvaus” σ:n koosta. Tätä käsitellään myöhemmin.

- Jos kuitenkin estimoidaan todellisen keskiarvon μ suuruutta, ei **todellista hajonnan σ suuruutta tietenkään tiedetä**.

Silloin **otoksesta lasketaan** μ:n estimaatin $\bar{x}$ lisäksi myös

σ:n paras piste-estimaatti otoskeskihajonta $s = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$.

Sattuma määrää myös s:n arvon, kun otos poimitaan.

Silloin sattuman ”pelivara” kasvaa, kun **todellisen hajonnan σ korvikkeena käytetään sen estimaattia s** luottamusväliä määrättäessä.

Kuitenkin,

kun otoskoko n on ”suuri”, minkä rajana tässäkin käytännössä $\mathbf{n > 30}$,

otoskeskihajonta s vastaa riittävän hyvin todellista hajontaa σ ja luottamusväli voidaan määrätä kuten edellä.

↓ luottamuskerroin ↓

$$\bar{x} \pm z_c \frac{s}{\sqrt{n}} \qquad \text{tai} \qquad \bar{x} \pm z_c \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

↗ ↖ ↗ ↖

μ:n estimaatti keskivirhe μ:n estimaatti keskivirhe

Siis päätepisteet saadaan edellä kuvatuilla edellytyksillä samoista säännöistä kuin edellä. Tuntematon **todellinen hajonta σ vain korvataan otoksesta lasketulla estimaatillaan s.**

Menetelmä ei ota huomioon ”lisääntynyttä sattuman vaikutusta” ja todellisuudessa arvion luotettavuus on vähän(?) pienempi kuin käytetty luottamustaso.

2.b) Jos otoskoko n on pienempi (**n** $\leq$ **30**), ”sattuma ei aseta” otoshajontaa s riittävän varmasti riittävän lähelle todellista hajontaa σ. Silloin luottamusväliä ei saa laskea edellisellä tavalla.

σ tunnettu	tuntematon
Otosta poimittaessa	
Sattuma on vaikuttanut vain keskiarvon suuruuteen.	**Sattuma** on päässyt vaikuttamaan sekä keskiarvoon että hajontaan.
↘	↘ ↓
$\bar{\mathbf{x}} \pm z_c \frac{\sigma}{\sqrt{n}}$	$\bar{\mathbf{x}} \pm z_c \frac{\mathbf{s}}{\sqrt{n}}$

- Kun σ korvataan keskihajonnalla s, joudutaan tämä ”sattuman vaikutuksen lisääntymisestä aiheutuva epävarmuus” maksamaan joko arvion luotettavuuden tai tarkkuuden vähenemisellä.

- Jos **luotettavuudesta** c ei haluta tinkiä, on tingittävä arvion **tarkkuudesta** suurentamalla $\pm$ -osaa ”sopivalla” tavalla:

Voidaan osoittaa, että(?))

luottamuskerroin on silloin määrättävä **t-jakaumasta**,

mutta muuten luottamusvälin **rakenne säilyy** samanlaisena kuin edellä.

t-jakauma on normaalijakaumasta johdettu jakauma, jonka

määritelmä on:

Jos $X, Z_1, Z_2, \ldots, Z_n \sim N(0,1)$ ja ovat riippumattomia,

niin sanotaan, että satunnaismuuttuja

$$t(n) = \frac{X}{\sqrt{\frac{1}{n}\sum Z_i{}^2}}$$ (Huom. osoittaja ja nimittäjä riippumattomia!)

noudattaa t-jakaumaa vapausastein n. (?)

Käytännössä

- edellistä teoriakysymysten tarkasteluissa tärkeää (tässä väistämättä varsin hämäräksi jäävää)

- määritelmää ei tarvita jatkossa sovelluksissa.

- Tiheysfunktion lauseke on erittäin hankala, ja lisäksi myöskään tässä tapauksessa kertymäfunktiota ei pystytä määräämään integroimalla tavalliseen tapaan.

- Sen sijaan tässäkin ”sopivien” polynomien(?) avulla saadaan kertymäfunktion arvoille hyvät approksimaatiot.

- Kertymäfunktion $F_{t(n)}$ arvot saa (mm.) Excelistä.

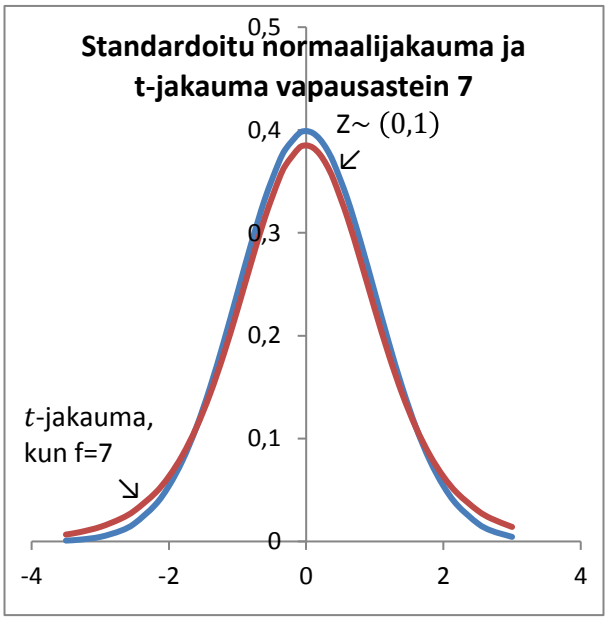
(Formulas → More Functions → Statistical → T.DIST)

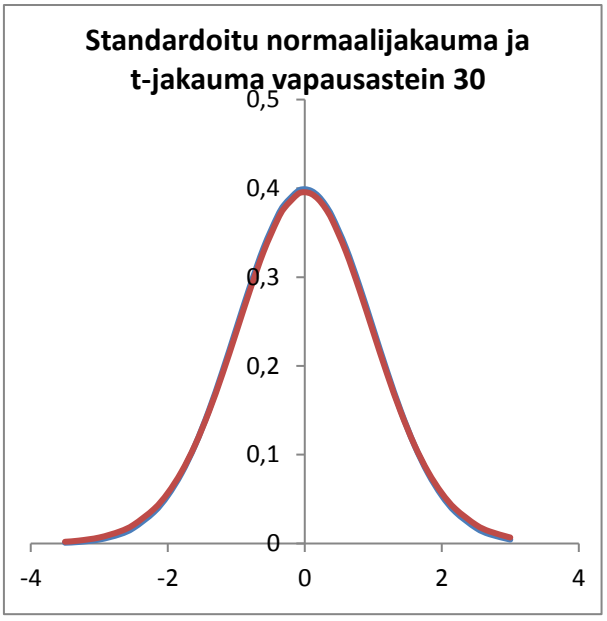
Esimerkiksi $P(t(10) \leq 2.228) = F_{t(10)}(2.228) = 0.974994 \approx 0.975$.

t-jakaumaan liittyy vapausasteluku f(?), joka täsmentää käsiteltävän jakauman.

t-jakauman muoto muistuttaa normaalijakaumaa, mutta se on ”laakeampi”:

Kuviossa on standardoidu normaalijakauman ja t-jakauman vapausastein f = 7 ja f = 30 tiheysfunktioiden kuvaajat:





Kun f = 30, tiheysfunktioiden ero on hyvin pieni

Voidaan osoittaa(?), että t-jakauma ”lähestyy”(?) normaalijakaumaa, kun vapausasteluku f kasvaa kohti ääretöntä.

- Luottamusväleihin ja myöhemmin käsiteltävään testaamiseen t-jakauma saadaan mukaan (mm.) otoskeskiarvon $\bar{X}$ otantajakauman kautta:

Jos perusjoukossa on $X \sim N(\mu, \sigma^2)$,

- niin n:n suuruisessa otoksessa on $\boldsymbol{\bar{X} \sim N(\mu, \frac{\sigma^2}{n})}$,

- jolloin $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

- Edellä luottamusvälin päätepisteiden määrääminen perustui tähän standardoituun muuttujaan.

- Käytännössä σ on lähes aina tuntematon ja se joudutaan korvaamaan otoskeskihajonnalla **s.**

Silloin näin määriteltävän satunnaismuuttujassa

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

sattuma määrää sekä otoskeskiarvon $\bar{X}$ että hajonnan s arvon.

Voidaan osoittaa, että(?) tämä testisuureeksi nimitettävä satunnaismuuttuja

$\mathbf{t = \frac{\bar{X} - \mu}{s/\sqrt{n}}}$ noudattaa t-jakaumaa vapausastein **f = n-1**.(?)

Huom. t-jakautuneen satunnaismuuttujan

määrittelyssä $t(n) = \frac{X}{\sqrt{\frac{1}{n}\sum Z_i^{\,2}}}$ osoittaja ja nimittäjä ovat **riippumattomia**.

Voidaan osoittaa, että edellisessä

informaatio **keskimäräisestä** ↙ suuruudesta

sovelluksessa $\mathbf{t = \frac{\overset{\swarrow}{\bar{X} - \mu}}{\frac{s}{\sqrt{n}}}}$ on samalla tavalla!

↖ informaatio **vaihtelun** suuruudesta

Siis normaalisti jakautuneissa muuttujan arvoissa $x_1, x_2, \ldots, x_n$ informaatio on ”niin hyvin järjestäytynyttä”, että

sattuma generoi otokseen keskimääräistä suuruutta ja vaihtelun suuruutta kuvaavan informaation toisistaan riippumatta,

vaikka otoskeskiarvo $\bar{x}$ ja keskihajonta s tullaan laskemaan samoista arvoista!

- Tämä perustavalaatuinen teoreettinen ihmeellisyys ei tässä vaikuta käytännön laskemiseen, vaan laskut sujuvat hyvin samalla tavalla kuin edellä.

Voidaan osoittaa(?), samoin kuin edellä tehtiin, että todellisen keskimääräisen suuruuden μ luottamusväli määrätään seuraavalla tavalla:

Oletetaan, että

- perusjoukko E on ääretön (käytännössä hyvin suuri) tai
- E:n kokoa ei tunneta (äärellisyyttä ei pystytä hyödyntämään) tai
- N:n kokoisesta perusjoukosta otos poimitaan palauttaen

ja tutkittava ominaisuus $\mathbf{X \sim N(\mu, \sigma^2)}$ **perusjoukossa**.

- Perusjoukosta poimitaan n:n suuruinen otos, josta lasketaan

sekä μ:n estimaatti otoskeskiarvo x̄ **että** σ:n estimaatti keskihajonta s.

Todellisen perusjoukossa olevan keskimääräisen suuruuden μ luottamusvälin päätepisteet luottamustasolla c ovat

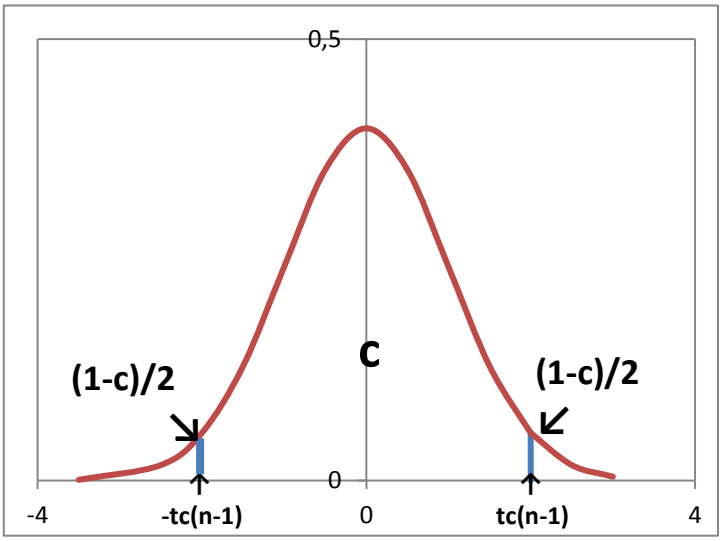
$\bar{x} \pm t_c(n-1)\frac{s}{\sqrt{n}}$, missä

luottamuskerroin $t_c(n-1)$

määrätään t-jakaumasta

ehdosta: ↔

$P(-t_c(n-1) < t(n-1) < t_c(n-1)) = c$



Jos n:n suuruinen otos poimitaan palauttamatta N:n suuruisesta perusjoukosta, jossa tutkittavan ominaisuuden jakauma on $X \sim N(\mu, \sigma^2)$, ovat

luottamusvälin päätepisteet luottamustasolla c ovat

$\bar{x} \pm t_c(n-1) \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$, missä

luottamuskerroin $t_c(n-1)$ saadaan samasta ehdosta kuin edellä.

Rakenne on aivan sama kuin aikaisemmassa tapauksessa.

Luottamuskerroin saadaan helposti Excelistä:

Esim. Jos otoskoko on vaikkapa n = 25, niin vapausasteluku f = 25 - 1 = 24.

Jos luottamustaso on c = 0.95,

t-jakauman kertymäfunktion arvoa (ks. edellinen kuvio)

0.95 + (1-0.95)/2 = 0.975 vastaava luottamuskertoimen arvo on

$t_{0.95}(24) \approx 2.064$

(Formulas → More Functions → Statistical → T.INV)

Jos Exceliä tai vastaavaa välinettä ei ole, saadaan kertoimet taulukoista:

merkitsevyystaso α 1-suuntaisessa testissä

	0.05	0.025	0.01	0.005	0.001	0.0005
	merkitsevyystaso α 2-suuntaisessa testissä = 1- luottamustaso c					
α=1-c →	0.10	0.05	0.02	0.01	0.002	0.001
f ↓						
1	6,314	12,706	31,821	63,657	318,309	636,619
2	2,920	4,303	6,965	9,925	22,327	31,599
3	2,353	3,182	4,541	5,841	10,215	12,924
4	2,132	2,776	3,747	4,604	7,173	8,610
5	2,015	2,571	3,365	4,032	5,893	6,869
6	1,943	2,447	3,143	3,707	5,208	5,959
7	1,895	2,365	2,998	3,499	4,785	5,408
8	1,860	2,306	2,896	3,355	4,501	5,041
9	1,833	2,262	2,821	3,250	4,297	4,781
10	1,812	2,228	2,764	3,169	4,144	4,587
11	1,796	2,201	2,718	3,106	4,025	4,437
12	1,782	2,179	2,681	3,055	3,930	4,318
13	1,771	2,160	2,650	3,012	3,852	4,221
14	1,761	2,145	2,624	2,977	3,787	4,140
15	1,753	2,131	2,602	2,947	3,733	4,073
16	1,746	2,120	2,583	2,921	3,686	4,015
17	1,740	2,110	2,567	2,898	3,646	3,965
18	1,734	2,101	2,552	2,878	3,610	3,922
19	1,729	2,093	2,539	2,861	3,579	3,883
20	1,725	2,086	2,528	2,845	3,552	3,850
21	1,721	2,080	2,518	2,831	3,527	3,819
22	1,717	2,074	2,508	2,819	3,505	3,792
23	1,714	2,069	2,500	2,807	3,485	3,768
→24	1,711	2,064	2,492	2,797	3,467	3,745
25	1,708	2,060	2,485	2,787	3,450	3,725
26	1,706	2,056	2,479	2,779	3,435	3,707
27	1,703	2,052	2,473	2,771	3,421	3,690
28	1,701	2,048	2,467	2,763	3,408	3,674
29	1,699	2,045	2,462	2,756	3,396	3,659
30	1,697	2,042	2,457	2,750	3,385	3,646
35	1,690	2,030	2,438	2,724	3,340	3,591
40	1,684	2,021	2,423	2,704	3,307	3,551
50	1,676	2,009	2,403	2,678	3,261	3,496
60	1,671	2,000	2,390	2,660	3,232	3,460
80	1,664	1,990	2,374	2,639	3,195	3,416
100	1,660	1,984	2,364	2,626	3,174	3,390
200	1,653	1,972	2,345	2,601	3,131	3,340
500	1,648	1,965	2,334	2,586	3,107	3,310
N(0,1) ∞	1,645	1,960	2,326	2,576	3,090	3,291

Taulukko on kirjoitettu testaamisen näkökulmasta(?) ja siinä oleva

$\alpha = 1 - c$ on päättelyssä ”suurin siedettävissä oleva erehtymisen riski”(?).

Tätä käsitellään tarkemmin vähän myöhemmin.

Käytännössä luottamuskerroin määrätään taulukosta:

Vapausasteluku $f = n-1 = 25 - 1 = 24$ määrää rivin.

Merkitsevyystaso 2- suuntaisessa testissä(?)

$\alpha = 1 - c = 1 - 0.95 = 0.05$ määrää sarakkeen.

Keskeltä saadaan $t_{0.95}(24) = 2.064$

Esim. Uuden lääkkeen kehittelyssä tutkittiin koe-eläinten avulla seerumin S tasoa (erittäin kalliin analyysimenetelmän avulla).

25 suuruisessa havaintoaineistossa oli

$\bar{x} = 7.46$ (mMol/l) ja $s = 2.86$ (mMol/l) ja jakauma näytti normaaliselta.

95 % luottamusväli todelliselle keskimääräiselle S:n määrälle:

f = n-1 = 25 – 1 = 24 ja c = 0.95 ja $t_{0.95}(24) = 2.064$ saatiin jo edellä.

$\bar{x} \pm t_c(n-1) \frac{s}{\sqrt{n}} = 7.46 \pm 2.064 \frac{2.86}{\sqrt{25}} = 7.46 \pm 1.18$

ja 95 % luottamusväli μ:lle on

(7.46 - 1.18, 7.46 + 1.18) = (6.28, 8.64).

Esim. Markkinatutkimuksessa kuntosaliketjun jäsenrekisteristä poimitusta 200 suuruisesta otoksesta laskettiin haastateltujen ilmoittamista arvoista (mm.) (muihin kuin liikunta-)

kulttuuripalveluihin kuukausittain käytetyn rahamäärän

keskiarvo $\bar{x}$ = 308 € ja hajonta s = 280 €.

Mikä on jäsenten käyttämän rahamäärän μ 95 %:n luottamusväli?

Vaikka otoksesta saatujen arvojen jakauma ei näyttänyt sitä tutkittaessa aivan normaaliselta, otoskoko on niin ”suuri”, että luottamusväli voidaan määrätä t-jakauman tai normaalijakauman avulla:

Hajonta on laskettu otoksesta, mikä viittaa t-jakauman käyttöön. Muuttujan normaalisuus on tässä kuitenkin oleellinen edellytys(?).

t-jakaumaa voidaan kuitenkin ”huonolla omalla tunnolla” käyttää, koska otos on näin ”suuri”.

Luottamustaso c = 0.95 ja virhearvion riski $\alpha = 1 - c = 0.05$,

vapausasteluku $f = 200 - 1 \approx 200$

ja t-jakauman taulukosta (tai Excelistä) saadaan $t_{0.95}(199) \approx 1.972$

(Vrt. $z_{0.95} = 1.96$, ero on hyvin pieni.)

Päätepisteet ovat

$\bar{x} \pm t_c(n-1) \frac{s}{\sqrt{n}}$

$= 308 \pm 1.972 \cdot \frac{280}{\sqrt{200}} = 308 \pm 39.04$ (38.8 normaalijakaumalla)

95 %:n luottamusväli todelliselle keskimääräiselle kulttuuripalveluihin käytetylle rahamäärälle μ on

(308 – 39, 308 + 39) = (269, 347) €.

Perusjoukon koko N = 5002 tiedetään, ja otos poimittiin palauttamatta, joten äärellisen perusjoukon korjaustekijä saadaan vielä mukaan:

$\bar{x} \pm t_c(n-1) \frac{s}{\sqrt{n}} \sqrt{\frac{\mathbf{N-n}}{\mathbf{N-1}}}$

$= 308 \pm 1.972 \cdot \frac{280}{\sqrt{200}} \sqrt{\frac{\mathbf{5002-200}}{\mathbf{5002-1}}} = 308 \pm 39.04 \cdot \mathbf{0.9799}$

Tarkkuus paranee ↑ noin 2 %.

$= 308 \pm 38.3$

ja 95 %:n luottamusväli (308 – 38, 308 + 38) = (270, 346) €

on vähän kapeampi.

Esim. (jatkoa)

Keskiarvo = $\frac{\text{kokonaismäärä}}{\text{arvojen lkm}}$, jolloin

kokonaismäärä = (arvojen lkm)·(keskiarvo)

Edellisen mukaan 95 %:n varmuudella

5002 jäsenen keskimäärin käyttämä rahamäärä μ on 270:n ja 346 €:n välillä.

Silloin jäsenten käyttämä myös 95 % varmuudella

kokonaisrahamäärä 5002· μ on välillä

(5002·270, 5002·346) = (1350540, 1730692) ≈ (1 350 500, 1 730 700) €.

Siis

Ominaisuuden X perusjoukossa olevan kokonaismäärän luottamusväli saadaan kertomalla keskimääräisen suuruuden μ luottamusvälin päätepisteet perusjoukon koolla N.

Edellä olevissa monissa tapauksissa oli yhteinen rakenne:

- Tutkitaan perusjoukon E tilastoyksiköiden ominaisuutta x, jonka

todellinen keskimääräinen suuruus on μ ja hajonta σ.

(Siis ominaisuuden EX = μ ja Var(X) = σ^2, kun X on ominaisuuden x arvo umpimähkään arvottavassa tilastoyksikössä.)

- Poimitaan n:n suuruinen otos.

Todellisen keskimääräisen suuruuden μ luottamusväli luottamustasolla c:

Päätepisteet ovat

1) 2) 3)

$\boxed{\bar{x}} \pm \boxed{\text{luottamuskerroin}} \cdot \boxed{\text{keskivirhe}}$

1) $\bar{x}$ on μ:n paras piste-estimaatti.

2) $\bar{x}$:n keskivirhe on

a) $\frac{\mathbf{s}}{\sqrt{\mathbf{n}}}$, jos σ on tuntematon, kuten se yleensä on, ja se korvataan otoshajonnalla s.

b) $\frac{\boldsymbol{\sigma}}{\sqrt{\mathbf{n}}}$, jos σ (jostain kummallisesta syystä) tiedetään.

c) Perään tulee tekijäksi $\sqrt{\frac{\mathbf{N-n}}{\mathbf{N-1}}}$, jos otos poimitaan palauttamatta N:n suuruisesta perusjoukosta.

3) Luottamuskerroin katsotaan

a) t-jakaumasta, jos σ on tuntematon ja se korvataan otoksesta lasketulla hajonnalla s.

Myös normaalijakaumaa saa käyttää, jos n on ”suuri” (> 30).

b) normaalijakaumasta, jos σ tunnetaan.

4) Menetelmää saa käyttää,

a) kun ominaisuus $X \sim N(\mu, \sigma^2)$ perusjoukossa,

b) ja silloin, kun otoskoko $n > 30$, vaikka X ei olisikaan normaalinen.

Suhteellisen osuuden p luottamusväli

seuraa otoksesta laskettavan suhteellisen osuuden $\hat{P}$ otantajakaumasta samalla tavalla kuin edellä estimoitiin todellista keskiarvoa μ.

Esim. Otantatutkimuksen avulla tutkittiin (mm.) hyödykkeen R käytön yleisyyttä alueen kotitalouksissa.

500 suuruisessa otoksessa 32 % kotitalouksista käytti R:ää.

Minkä rajojen sisällä R:ää käyttävien kotitalouksien todellinen suhteellinen osuus p on 95 % varmuudella?

- Otoksesta saatu suhteellinen osuus $\hat{p} = 0.32$ on p:n paras piste-estimaatti.

- Sattuma on generoinut otoksen sisällön suhteellisen osuuden otantajakauman määrittelemien sääntöjen mukaisesti:

- Jos perusjoukossa R:n käyttäjien suhteellinen osuus on p,

- niin n:n suuruisessa otoksessa käyttäjien suhteellisen osuuden $\hat{P}$ otantajakauma on

kaikissa muissa tapauksissa: paitsi

kun otos poimitaan palauttamatta N:n suuruisesta perusjoukosta:

$\hat{p} = 0.32$ estimoi ”hyvin” tuntematonta p:n arvoa.

↙ ↘

$$\hat{P} \sim N\left(p, \frac{p(1-p)}{n}\right) \qquad \hat{P} \sim N\left(p, \frac{p(1-p)}{n} \cdot \frac{N-n}{N-1}\right)$$

↕ Varianssista saadaan keskivirhe. ↕

$$\sigma_{\hat{P}} = \sqrt{\frac{p(1-p)}{n}} \qquad \sigma_{\hat{P}} = \sqrt{\frac{p(1-p)}{n} \cdot \frac{N-n}{N-1}}$$

↑ ↑

Myös keskivirheessä on juuri estimoitavana oleva p korvattava otoksesta lasketulla estimaatillaan $\hat{p}$.

Otantajakaumat perustuvat normaaliapproksimaatioon, jossa vaaditaan,

että $np > 5$ ja $n(1-p) > 5$.

Näin on oltava myös luottamusväliä määrättäessä. (p korvataan $\hat{p}$:lla.)

Samoin kuin keskimääräistä suuruutta estimoitaessa luottamusväli siis perustuu (symmetriseen) normaalijakaumaan.

Silloin ”on luonnollista”(?), että välin päätepisteet määrätään samalla tavalla:

”piste-estimaatti” $\pm$ **”luottamuskerroin”** x **”keskivirhe”**

tähän ↑ tähän ↑ tähän ↑

paras yksittäinen arvio p:stä | tieto vaaditusta varmuuden asteesta | sattuman ”pelivara”

Siis tässä:

Perusjoukossa tilastoyksiköiden ominaisuuden A (esim. R:n käyttö) todellinen suhteellinen osuus **p on tuntematon**.

Poimitusta n:n suuruisesta **otoksesta** saadaan suhteellinen osuus $\hat{\mathbf{p}}$.

Luottamustasolla c todellisen suhteellisen osuuden p luottamusvälin päätepisteet ovat(?), kun

- otos poimitaan **palauttaen** tai

- perusjoukko on **ääretön** tai sen kokoa ei tunneta, jolloin otos voidaan poimia myös palauttamatta

$$\hat{\mathbf{p}} \pm \mathbf{t_c(n\text{-}1)} \sqrt{\frac{\hat{\mathbf{p}}(\mathbf{1}-\hat{\mathbf{p}})}{\mathbf{n}}} \quad \text{ja}$$

↗ ↑ ↖

p:n paras piste-estimaatti ↘ | luottamus-kerroin ↓ | keskivirhe ↙

ja $$\hat{\mathbf{p}} \pm \mathbf{t_c(n\text{-}1)} \sqrt{\frac{\hat{\mathbf{p}}(\mathbf{1}-\hat{\mathbf{p}})}{\mathbf{n}} \cdot \frac{\mathbf{N}-\mathbf{n}}{\mathbf{N}-\mathbf{1}}},$$

jos otos poimitaan **palauttamatta N:n suuruisesta perusjoukosta**.

Tässäkin saa (Ei tehdä suurta virhettä!) korvata luottamuskertoimen normaalijakaumasta saatavalla z_c:llä, kun otoskoko $n > 30$.

Esim. (jatkoa) Markkinatutkimuksessa 200 suuruisessa otoksessa 32.5 % alueen kotitalouksista käytti R:ää.

- $np = 200 \cdot 0.325 = 65$ (käyttäjien määrä) > 5 ja

$n(1-p) = 200 \cdot (1-0.325) = 135$ (ei-käyttäjien määrä) > 5,

joten taustalla olevan otantajakauman normaalisuus on kunnossa.

- Otos on niin suuri, että luottamuskerroin voidaan määrätä normaalijakaumasta ja edellä saatiin
95% luottamustasoa vastaa vastaava kerroin $z_{0.95} = 1.96$.

$$0.325 \pm 1.96 \cdot \sqrt{\frac{0.325 \cdot (1-0.325)}{200}} \approx 0.325 \pm 1.96 \cdot \mathbf{0.0331} = 0.325 \pm 0.065$$

↑

Näin pienessä otoksessa sattumalla on 3.31 %-yksikön ”pelivara”.

ja 95 % luottamusväli R:n käyttäjien todelliselle suhteelliselle osuudelle p alueen kotitalouksissa on

(0.325 – 0.065, 0.325 + 0.065) = (0.260, 0.390).

95 % varmuudella R:n käyttäjien osuus on 26.0 ja 39.0 % välillä.

Arvion tarkkuus paranee vähän, kun ”muistetaan”, että alueella on yhteensä 45 000 kotitaloutta ja otos poimitaan palauttamatta.

Silloin voidaan hyödyntää korjaustekijän sattuman ”pelivaraa pienentävä vaikutus”:

Päätepisteet ovat

$$0.325 \pm 1.96 \cdot \sqrt{\frac{0.325 \cdot (1-0.325)}{200} \cdot \frac{45000-200}{45000-1}}$$

$$\approx 0.325 \pm 1.96 \cdot \mathbf{0.0330} = 0.325 \pm 0.065$$

”Virhemarginaali” pienenee vain aavistuksen verran, mutta edelleen arvio on käytännössä yhtä epätarkka.

<u>Samasta otoksesta</u> saatiin hyödykettä Q käyttävien osuudeksi 9.5 %.

95 % luottamusväli Q:n käyttäjien todelliselle suhteelliselle osuudelle alueen kotitalouksissa on

$$0.095 \pm 1.96 \cdot \sqrt{\frac{0.095 \cdot (1-0.095)}{200} \cdot \frac{45000-200}{45000-1}}$$

$$\approx 0.095 \pm 1.96 \cdot \mathbf{0.0207} = 0.095 \pm 0.041$$

↗ ↖

Sattuman ”pelivara” ja siten myös ”virhemarginaali” ovat pienempiä kuin edellä.

95 % luottamusväli (0.095 – 0.0401, 0.095 + 0.0401) = (0.054, 0.136) arvioi tarkemmin Q:n käyttäjien suhteellisen osuuden.

- Kun otoskoko on n, niin suhteellisen osuuden otantajakauman keskivirheen $\boldsymbol{\sigma_{\hat{P}}} = \sqrt{\frac{\mathbf{p(1-p)}}{\mathbf{n}}}$ suuruuden määrää **p(1-p)**, joka saa maksimiarvon, kun p = 0.5.

Tällöin myös luottamusväli on levein, kun $\hat{p} = 0.5$, ja taas arvio on sitä tarkempi, mitä lähempänä $\hat{p}$ on nollaa tai ykköstä.

Tässä Q:n käytölle saadaan tarkempi arvio, mutta molemmat tulokset

ovat varsin epätarkkoja, kun arvioiden luotettavuuden on oltava 95 %:n suuruinen.

R:n käyttäjien **kokomaismäärän luottamusväli** saadaan vastaavalla tavalla kuin edellä:

- Otoksessa R:ää käytti 32.5 % haastatelluista,

- Silloin 45000 suuruisessa perusjoukossa R:n käyttäjille

paras piste-estimaatti on $45000 \cdot 0.325 = 14625$.

- Kun todellinen suhteellinen osuus on **95 %:n varmuudella** välillä (0.260, 0.390), niin rajoja vastaava

- **kokonaismäärä** on välillä

$(45000 \cdot 0.260, 45000 \cdot 0.390) = (11700, 17550)$.

Siis perusjoukkoon saadaan **luottamusväli** niiden tilastoyksiköiden **kokonaismäärälle,** joilla on ominaisuus A, **kertomalla vastaavan suhteellisen osuuden luottamusvälin päätepisteet perusjoukon koolla N.**

Edellisessä esimerkissä saadut tulokset ovat hyvin epätarkkoja.

- Sattuman ”pelivaraa” ja sitä kautta ”virhemarginaalia” voidaan kyllä pienentää ”maksamalla” siitä otoskoon n kasvattamisella.

- Toisaalta suuren otoksen poimiminen vaatii paljon resursseja.

Miten saadaan selville sekä tutkimuksen laatuvaatimusten että resurssien säästämisen kannalta **optimaalinen otoskoko**?

Otoskoon suuruuden arvioimisesta

Esim. (jatkoa) Markkinatutkimuksessa 200 suuruisessa otoksessa 32.5 % alueen kotitalouksista käytti R:ää.

95 % luottamusväliksi R:n käyttäjien todelliselle suhteelliselle osuudelle p alueen kotitalouksissa laskettiin

$0.325 \pm 1.96 \cdot \sqrt{\frac{0.325 \cdot (1-0.325)}{200}} \approx 0.325 \pm 0.065$ ja

luottamusväli on

(0.325 – 0.065, 0.325 + 0.065) = (0.260, 0.390).

Arvion tarkkuus paranee vain vähän, kun hyödynnetään äärellisen perusjoukon korjaustekijä.

Tulos, jonka mukaan

95 % varmuudella R:n käyttäjien osuus on 26.0 ja 39.0 % välillä,

ei kelpaa tilaajalle ja hän vaatii parantamaan tutkimusta (samaan hintaan?) niin, että

- **luotettavuuden** on oltava edelleen 95 % tasolla, mutta

- **tarkkuuden**, jonka $\pm$ - osa, ”virhemarginaali”, esittää on oltava noin 2 %-yksikköä.

Siis uusi otantatutkimus on suunniteltava niin, että

95 % varmuudella todellinen R:n käyttäjien suhteellinen osuus p

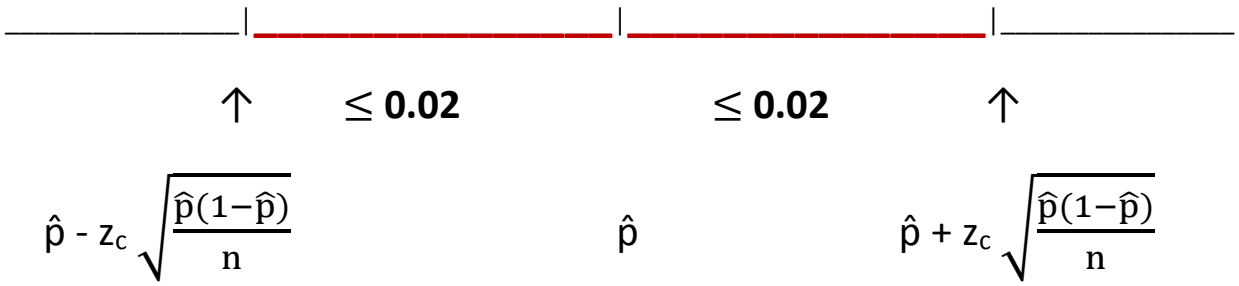
ei poikkea yli 2 % -yksikköä otoksesta saatavasta arviosta $\hat{p}$?

Oikea todellinen R:n käyttäjien suhteellinen osuus p on jossain tällä välillä 95 %:n suuruisella varmuudella.

↓

↓ ↓

”virhemarginaali” ”virhemarginaali”



Luottamuskerroin on normaalijakaumasta saatava $z_{0.95} = 1.96$, mutta kaikki muu onkin tuntematonta.

Keskivirheessä $\boldsymbol{\sigma_{\hat{P}} = \sqrt{\frac{p(1-p)}{n}}}$ tarvittaisiin oikea p:n arvo, mutta edes sen estimaattia ei vielä ole, kun vasta suunnitellaan tutkimusta.

Kuitenkin

aikaisemmasta (tosin liian epätarkkoja tuloksi antaneesta) tutkimuksesta

saadaan jonkinlainen suuruusluokka-arvio p:n suuruudesta.

Sitä käytetään (paremman puutteessa) **alkuarvona** otoskoon suuruutta arvioitaessa.

Siis on laskettava, kuinka suuri otoskoko **n** tarvitaan, että

$$\boxed{\begin{array}{c}\text{alkuarvoksi}\\ \hat{p} \approx 0.325\\ \downarrow\end{array}}$$

$$|\hat{p} - p| \le z_c \sqrt{\frac{\hat{p}(1-\hat{p})}{\mathbf{n}?}} \le 0.02 \qquad \boxed{\leftarrow \textbf{tarkkuusvaatimus}}$$

$$\boxed{\begin{array}{c}\nearrow\\ 95\ \%\ \textbf{varmuutta}\\ \text{vastaava}\\ z_{0.95} = 1.96\end{array}} \quad \text{ja} \quad 1.96\sqrt{\frac{0.325(1-0.325)}{\mathbf{n}}} \le 0.02,$$

josta saadaan $$1.96^2 \cdot \frac{0.325(1-0.325)}{\mathbf{n}} \le 0.02^2$$

ja $$\mathbf{n} \ge 1.96^2 \cdot \frac{0.325(1-0.325)}{0.02^2} \approx 2106.9 \cong 2100.$$

Otoskoko on noin 10-kertaistava, jotta saadaan vaadittu tarkkuus luotettavuudesta tinkimättä!

Jos arviolta vaadittaisiin myös parempaa, esim. 99 %:n tasoista luotettavuutta,

edellisessä laskussa vain 1.96:n tilalle tulee $z_{0.99} \approx 2.58$ ja

$$n \geq \mathbf{2.58}^2 \cdot \frac{0.325(1-0.325)}{0.02^2} \approx 3650.6 \text{ !}$$

Sattuma ei paljasta perusjoukon salaisuuksia ilmaiseksi. Jos halutaan ”hyviä” arvioita eli **korkeaa luotettavuutta** ja **suurta tarkkuutta**, tarvitaan paljon informaatiota otokseen, josta tätä ”hyvää” voidaan informaatiosta jalostaa.

Jos edellisessä tilanteessa ei ole mitään käsitystä p:n suuruusluokasta, alkuarvona voitaan käyttää p:n arvoa 0.5. Se maksimoi otantajakauman keskivirheen, ja tuloksena laskusta on aina riittävän (ehkä liiankin) suuri n.

Todellista keskimääräistä suuruutta

tutkittaessa ei tällaista aina riittävää alkuarvoa ole, vaan tarvitaan jonkinlainen järkevä ”arvaus” muuttujan hajonnasta σ:

Esim. Yritys muuttaa tuottamiensa vekottimien tuotantoprosessia ja haluaa tutkia tuotteen kestoikää otoksen avulla.

Ennen tutkimusta asetetaan vaatimuksiksi, että

95 %:n varmuudella (**luotettavuus**)

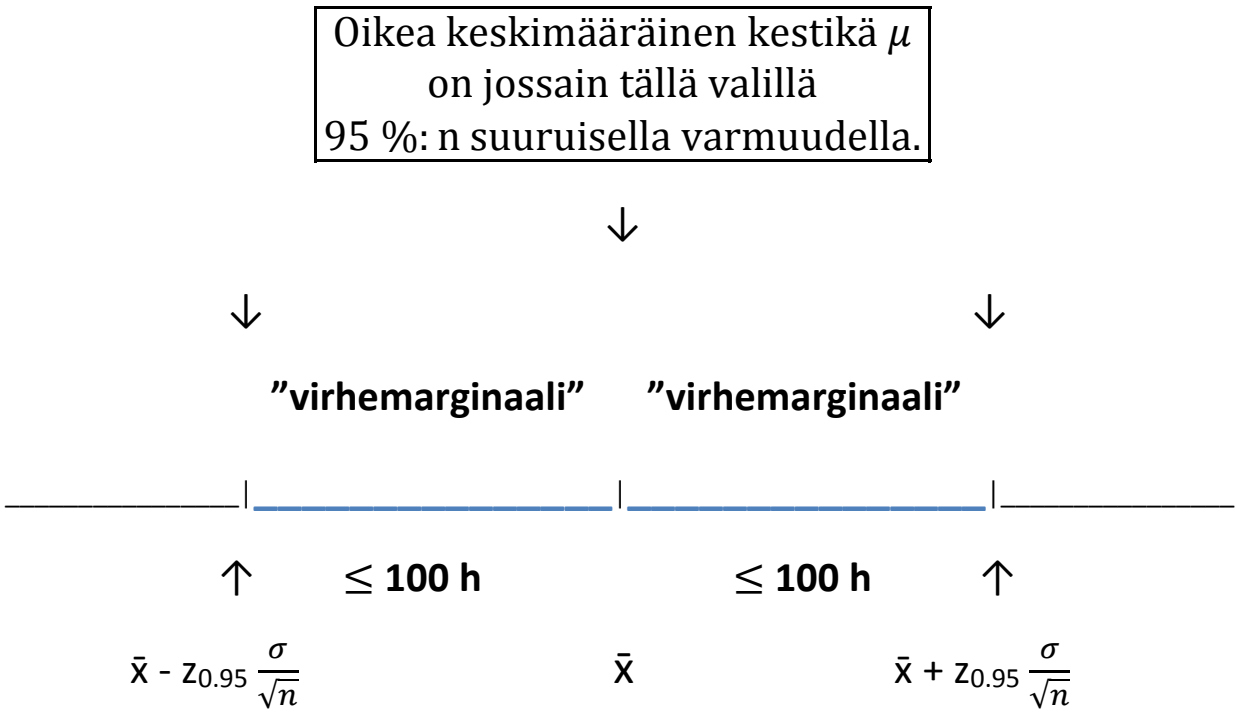
todellinen keskimääräinen kestoikä μ kaikkien vekottimien perusjoukossa ja otoksesta saatava kestoiän arvio $\bar{x}$

eivät saa poiketa toisistaan yli 100 h (**tarkkuus**).

Aikaisemmin vastaavanlaisten tuotteiden tutkimuksissa hajonta σ

ei ole ollut yli 500 h. Tästä saadaan (ylöspäin arvioitu) alkuarvo σ:lle.

Siis μ:n luottamusvälin on oltava



σ:n (ylöspäin tehtynä) arviona on 500 h

ja epäyhtälöstä $1.96 \cdot \frac{500}{\sqrt{\mathbf{n}}} \leq 100$

saadaan $n \geq (\frac{1.96 \cdot 500}{100})^2 = 96.04$.

Otoskoko n = 97 ≈ 100 riittää halutun

tarkkuuden (virhemarginaali ≤ 100 h) ja luotettavuuden (c = 0.95) saamiseen.

Jos vaadittaisiin 99 %:n varmuus tarvittaisiin ($z_{0.99} \approx 2.58 \rightarrow 1.96$)), saadaan samalla tavalla laskemalla

$$n \geq \left(\frac{\mathbf{2.58} \cdot 500}{100} \right)^2 = 166.41.$$

Voi olla yllättävää, että tarvitaan lisää noin 70 muuttujan arvon informaatio, jotta arvion **luotettavuus kasvaa** 95 %:sta 99 %:in.

Asiaa kannattaa kuitenkin tarkastella virheellisen arvion tekemisen riskin näkökulmasta.

Virhearvion **riski pienenee** 5 %:sta 1 %:in!