4 Estimoinnista

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

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

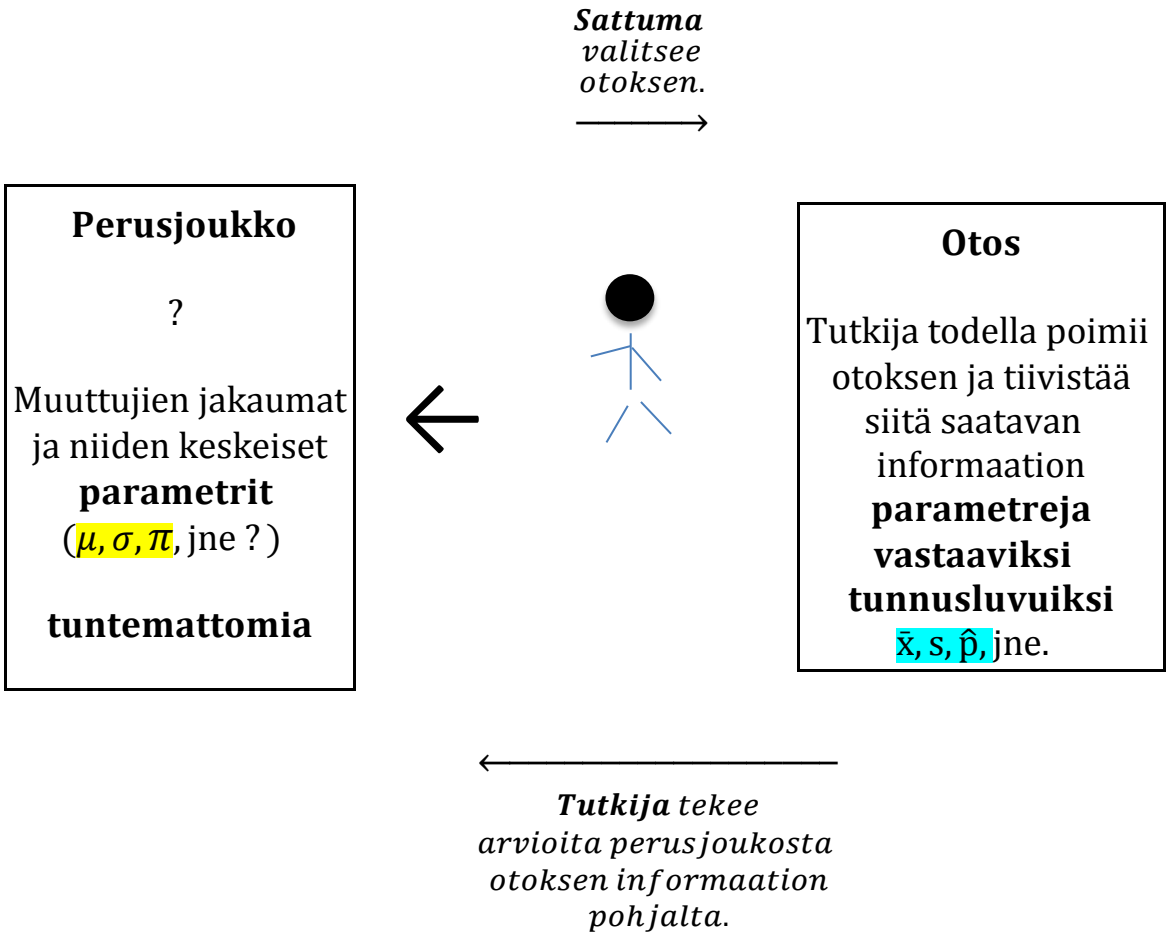
Nyt käytössä ovat riittävät ”työkalut”, joiden avulla **tarkastelun suunta voidaan kääntää toisin päin** normaaliksi tutkimustilanteeksi keskimääräisen suuruuden μ ja suhteellisen osuuden π osalta.

Huom. Tässä käytetään perusjoukossa olevasta jonkun ominaisuuden suhteellisesta osuudesta symbolia $\boldsymbol{\pi}$, jotta perusjoukon parametrit ja niille otoksesta laskettavat arviot (estimaatit) eivät sekoitu keskenään.

- Todellinen keskiarvo $\boldsymbol{\mu}$ (tai suhteellinen osuus $\boldsymbol{\pi}$) on **tuntematon**.

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

- Tiivistetyn informaation avulla arvioidaan ”otantajakauman jälkiä takaisin **perusjoukkoon päin** seuraten”, mikä on μ:n (π: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

π = 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 $\mathbf{y = \beta_0 + \beta_1 x_1}$?

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

$\mathbf{y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2}$?

Mitkä ovat nyt γ_0, γ_1 ja γ_2?

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

Estimoinnissa käytetään välineinä otoksesta laske**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.

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

- (esim.) 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.

Tällaista estimaattoria 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 π estimaattori $\hat{P}$ on myös harhaton ja sen otantajakauma on (likimain)

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

Perusjoukossa (piilossa) olevan

todellisen keskimääräisen suuruuden $\boldsymbol{\mu}$, varianssin $\boldsymbol{\sigma}^2$ ja suhteellisen osuuden $\boldsymbol{\pi}$

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.

Tässä käsiteltävien ***parametrien*** osalta on selvää, että

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 π

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*.

Otantajakaumien perusteella tiedetään, että estimaatit ovat

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

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

Arvioiden ja niistä seuraavien johtopäätösten tekemisessä järkevän **varovaisuusperiaatteen** vuoksi lähtökohtana **pidetään varmuuden (korkeaa) vakioitua tasoa**, jolla päätelmät tehdään.

Sen pohjalta tutkitaan, kuinka tarkka arvio eli kuinka laaja väli saadaan.

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

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

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 $\bar{x}$ on paras piste-estimaatti μ:lle,

se on luonnollisesti määrättävän välin keskipiste.

Näin tutkitaan,

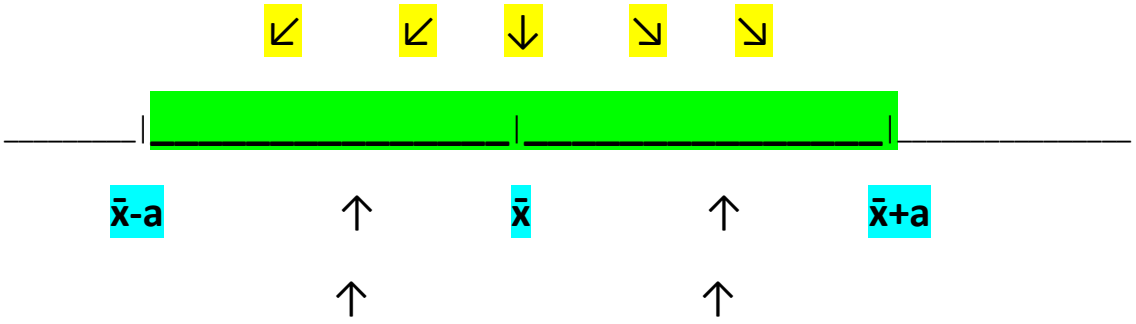
kuinka suureksi on **”virhemarginaali”** a valittava,

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

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

Oikea $\boldsymbol{\mu}$:n arvo on jossain tällä välillä

c:n suuruisella varmuudella.



”virhemarginaali”

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, että myös hajonta σ on yleensä aina tuntematon ja

- silloin luonnollisesti σ korvataan otoksesta laskettavalla estimaatillaan s.

Kontrollirajat ja luottamusväli

Edellä otantajakaumia käsiteltäessä tutkittiin **informaatiovirtaa perusjoukosta otokseen päin** (siis päinvastaiseen suuntaan kuin luottamusväliä määrättäessä).

Otoksesta laskettavan keskimääräisen suuruuden $\bar{X}$ otantajakauman käsittelyssä edellä oli esimerkki, jossa määrättiin ***kontrollirajat*** otoskeskiarvolle:

Siis selvitettiin, mille välille otoskeskiarvo $\bar{X}$ tulee osumaan, jos otos joskus poimitaan:

Esim. (jatkoa)

Elintarvikeannoksen lisäaineen E määrä $X \sim N(200 \text{ mg}, (15 \text{ mg})^2)$.

Annoksista **aiotaan poimia** 100 suuruinen otos.

On määrättävä a niin, että 95 %:n varmuudella otoskeskiarvo $\bar{x}$ **tulee poikkeamaan** todellisesta keskiarvosta μ = 200 mg korkeintaan a:n verran.

Sattuman käyttäytymistä säätelee otantajakauma

todellinen keskiarvo **perusjoukossa** μ ↓ todellinen hajonta **perusjoukossa** σ ↓

$$\bar{X} \sim N(200 \text{ mg}, \frac{(15 \text{ mg})^2}{100}) = N(\mathbf{200\ mg}, (\frac{\mathbf{15\ mg}}{\sqrt{\mathbf{100}}})^2)$$ (= N(100 g, (1,5 mg)^2).)

On oltava ↑ otoskoko **n**

0.95 = P(200-a $< \bar{X} <$ 200+a)

$$= P(\frac{200-a-200}{\frac{15}{\sqrt{100}}} < \frac{\bar{X}-200}{\frac{15}{\sqrt{100}}} < \frac{200+a-200}{\frac{15}{\sqrt{100}}}) = P(\frac{-a}{\frac{15}{\sqrt{100}}} < Z < \frac{a}{\frac{15}{\sqrt{100}}})$$

$$= \phi(\frac{a}{\frac{15}{\sqrt{100}}}) - \phi(\frac{-a}{\frac{15}{\sqrt{100}}}) = \phi(\frac{a}{\frac{15}{\sqrt{100}}}) - (1 - \phi(\frac{a}{\frac{15}{\sqrt{100}}}))$$

$$= 2\,\phi(\frac{a}{\frac{15}{\sqrt{100}}}) - 1,$$

josta $\phi\left(\frac{a}{\frac{15}{\sqrt{100}}}\right) = \frac{0.95+1}{2} = 0.975 = \phi(1.96)$ (← taulukosta).

Silloin $\frac{a}{\frac{15}{\sqrt{100}}} = 1.96$ ja $a = 1.96 \cdot \frac{15}{\sqrt{100}}$ (= 2.94 ≈ 3 mg)

ja

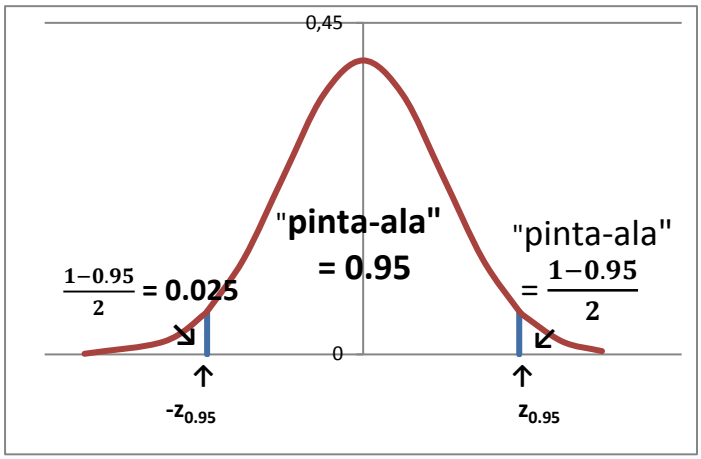
väli, jolle **otoskeskiarvo $\bar{X}$ tulee osumaan** 95 %:n varmuudella, jos otos tullaan joskus poimimaan, on

$$\left(200 - 1.96 \cdot \frac{15}{\sqrt{100}},\ 200 + 1.96 \cdot \frac{15}{\sqrt{100}}\right)$$

Siis, kun perusjoukossa todellinen keskiarvo on μ (=200) ja todellinen hajonta $\boldsymbol{\sigma}$(=15), niin n:n (=100) suuruisesta otoksesta laskettavasta otoskeskiarvosta $\bar{X}$ voidaan väittää, että se tulee osumaan c = 95 % varmuudella tälle välille eli

$$P\left(\mu - 1.96 \cdot \frac{\boldsymbol{\sigma}}{\sqrt{\mathbf{n}}} < \bar{X} < \mu + 1.96 \cdot \frac{\boldsymbol{\sigma}}{\sqrt{\mathbf{n}}}\right) = 0.95$$

Tavoiteltavaa 95 %:n varmuuden astetta vastaava normaalijakaumasta saatavan kertoimen (, josta käytetään merkintää) $z_{0.95}$ =1.96 määrääminen vastaa kuviona tilannetta:



Tästä saadaan ehto

$\phi(z_{0.95}) = 0.95 + 0.025 = 0.975,$

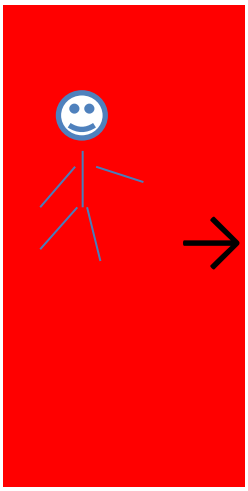
mihin edellä päädyttiin.

Kontrollirajat:

Sattuma

Perusjoukko

muuttujan **X** todellinen keskiarvo μ ja hajonta σ



Otos

otoskeskiarvo $\bar{x}$ osuu välille

$$\left(\mu - z_c \cdot \frac{\sigma}{\sqrt{n}}, \mu + z_c \cdot \frac{\sigma}{\sqrt{n}}\right)$$

c: n suuruisella varmuudella

↑

toimii ”ohjesääntönsä” $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

mukaisesti, kun se ”määrää” otoskeskiarvon suuruuden n:n suuruiseen otokseen.

Kun varmuuden aste, jonka mukaan rajat määrätään, on c (edellä oli c=0.95), kerroin z_c saadaan normaalijakaumasta vastaavalla tavalla kuin esimerkissä.

Huom. Tässä tarkastelunäkökulma on perusjoukosta otokseen päin ja tämä väli **ei ole luottamusväli,** kuten sitä lukiossa on saatettu virheellisesti nimittää.

Näin määrättäviä rajoja sanotaan ***kontrollirajoiksi.***

Tämän suuntainen perusjoukosta otokseen päin suuntautuva tarkastelu voi kuitenkin olla hyödyllistä mm. laadunvalvonnassa.

Jos otoksesta laskettu otoskeskiarvo $\bar{x}$ ei ole kontrollirajojen sisällä, kannattaa tarkistaa, onko tuotantoprosessissa jotain vikaa.

Kontrollirajat

$$\mu - z_c \cdot \frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + z_c \cdot \frac{\sigma}{\sqrt{n}}$$

esittävät, kuinka suuri otoskeskiarvon $\bar{X}$ etäisyys todellisesta (tunnetusta) keskiarvosta μ tulee enimmillään olemaan c:n suuruisella todennäköisyydellä, **jos otos joskus tullaan poimimaan.**

Siis $|\bar{X} - \mu| < z_c \cdot \frac{\sigma}{\sqrt{n}}$.

Jos otos on todella poimittu ja siitä on laskettu keskiarvo $\bar{x}$ (ja tietysti myös keskihajonta s), on järkevää päätellä symmetrisesti toisin päin,

että c:n suuruisella varmuudella todellisen (tuntemattoman) keskiarvon μ maksimietäisyys lasketusta keskiarvosta $\bar{x}$ on saman suuruinen.

Siis $\bar{x} - z_c \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_c \cdot \frac{\sigma}{\sqrt{n}}$ c:n suuruisella varmuudella.

Ongelmaksi jää kuitenkin, että todellinen hajonta σ on samoin kuin todellinen keskiarvo μ myös tuntematon.

Luonnollinen ratkaisu tähän on, että σ korvataan (parhaalla estimaatillaan) otoksesta lasketulla keskihajonnalla s.

Näin **tarkastelun suunta on käännetty otoksesta perusjoukkoon päin** ja

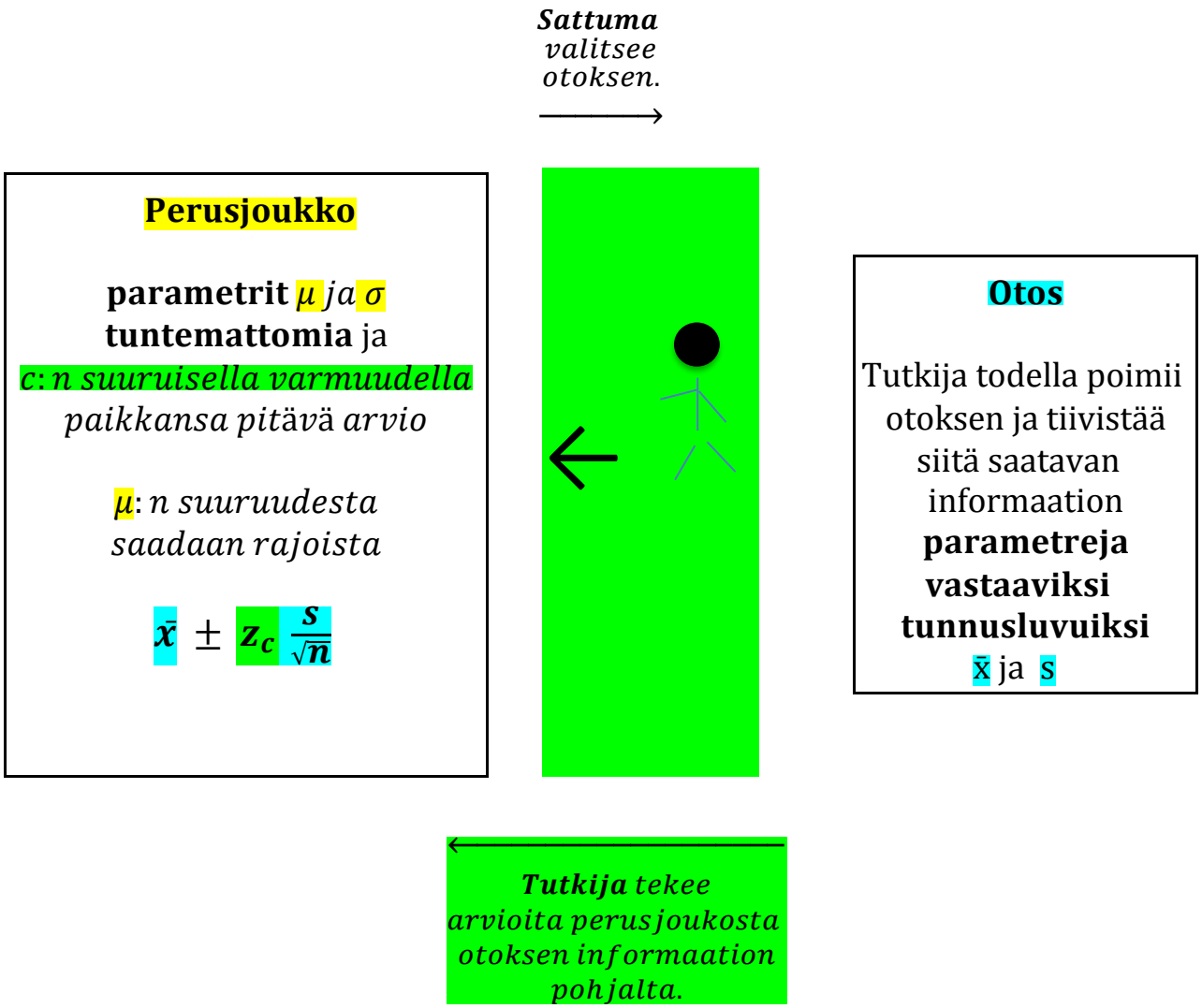
esimerkissä havaitun mukaan yleisin merkinnöin (sääntöjen täsmällinen matemaattinen johtaminen sivuutetaan):

Todellisen keskimääräisen suuruuden μ luottamusväli

kun

1. luottamuskerroin määrätään normaalijakaumasta

- ”suurten” otosten (n > 30) tapaus



1.a) 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 tutkittavasta ominaisuudesta voidaan olettaa, että **perusjoukossa** on

$\mathbf{X \sim N(\mu, \sigma^2)}$

- Perusjoukosta poimitaan n:n suuruinen otos, josta lasketaan otoskeskiarvo $\bar{x}$ ja keskihajonta s.

Jos otoskoko n on ”riittävän suuri”, otoskeskihajonta s on kelvollinen approksimaatio todelliselle hajonnalle σ.

Käytännössä otoskokoa n pidetään ”riittävän suurena”, kun $\mathbf{n > 30}$.

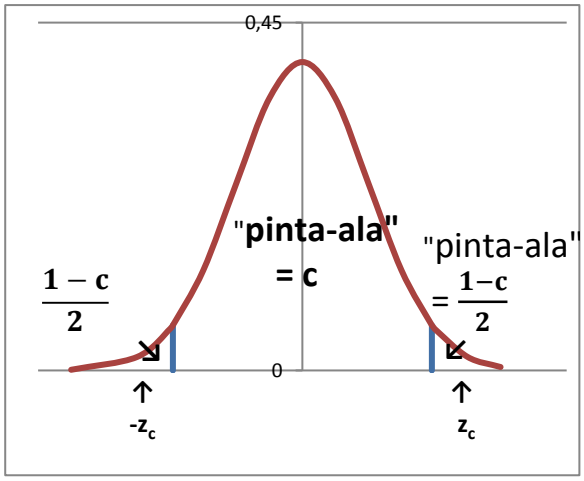
Silloin

todellisen perusjoukossa olevan keskimääräisen suuruuden μ
luottamusvälin *päätepisteet* ***luottamustasolla c*** ovat

$\bar{x} \pm z_c \frac{s}{\sqrt{n}}$, missä

luottamuskerroin z_c saadaan standardoidusta normaalijakaumasta ehdosta:

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



Tulos on kuitenkin vain approksimaatio. Esimerkin laskussa tarvittiin todellinen hajonta σ se joudutaan korvaamaan otoksesta lasketulla estimaatillaan **s**, mikä muuttaa asetelman rakennetta.

Jos otoskoko n on riittävän suuri, otoskeskihajonta **s** vastaa kuitenkin ”riittävän hyvin” todellista hajontaa σ ja approksimaatio on riittävän hyvä.

Tämän ongelman parempaan ratkaisuun palataan pian.

Esim. Tuotteen T **todellista** keskimääräistä kestoikää μ **ei tiedetä**.

On havaittu, että tuotantoprosessin muutoksista huolimatta tuotteen kestoikä on likimain normaalinen, mitä tuki myös

100 suuruisesta **otoksesta** havaittu jakauma, josta laskettu

keskimääräinen kestoikä oli $\bar{x}$ = 1500 h ja keskihajonta s = 200 h.

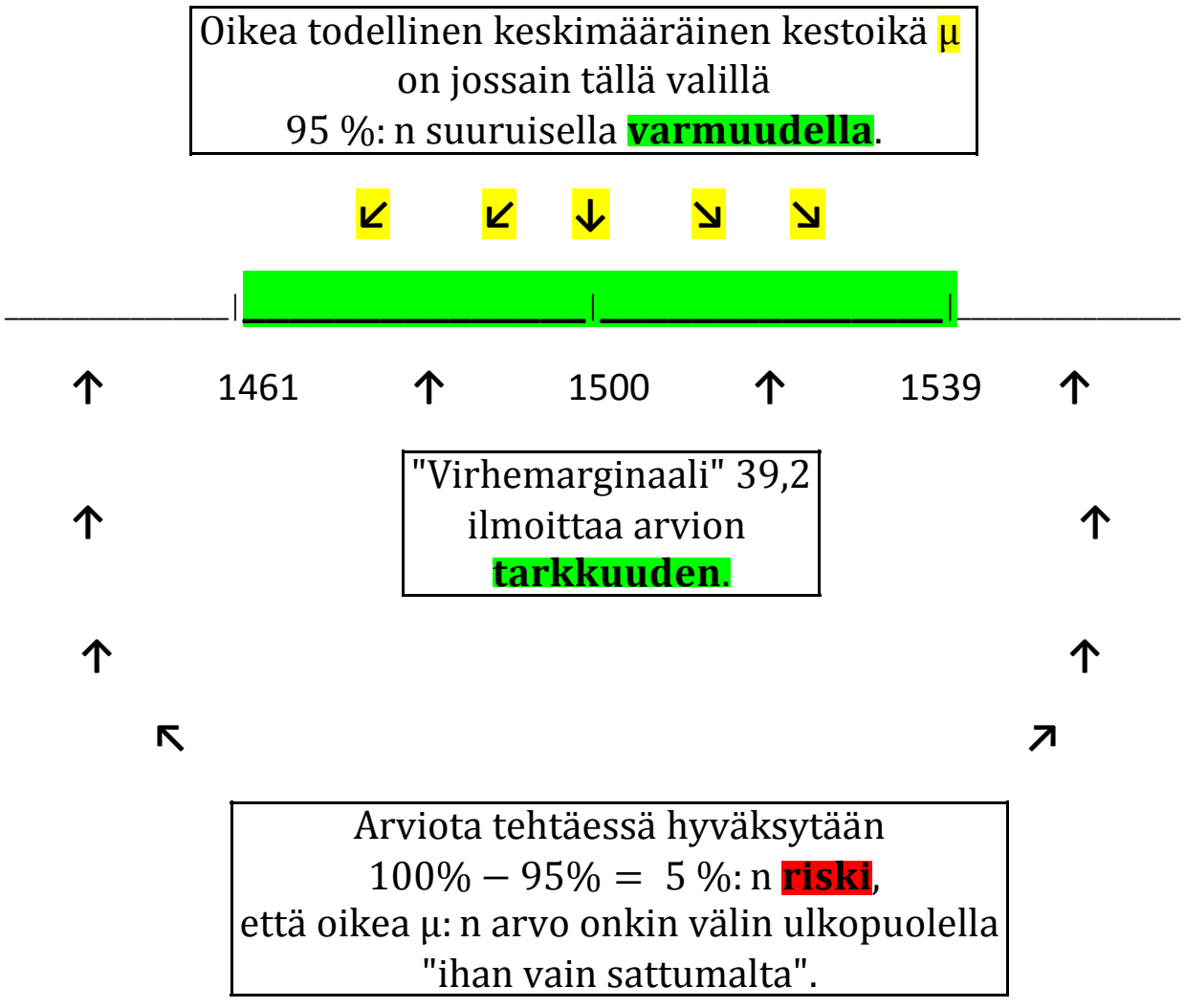
Mikä on todellisen keskimääräisen kestoiän μ 95 %:n luottamusväli?

Edellä määrättiin taulukosta (tai Excelistä yms.)

luottamuskerroin $z_{0.95}$ =1.96 ja 95 % luottamusväli tuotteen T todelliselle keskimääräiselle kestoiälle μ on

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

$$\approx (1461\text{ h},\ 1539\text{ h})$$



Esim. (jatkoa)

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

Sillä on kuitenkin **kova hinta**:

Kun luottamustasoksi valitaan 99 %, niin luottamuskerroin saadaan ehdosta

$0.99 = P(-z_{0.99} < Z < z_{0.99})$

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

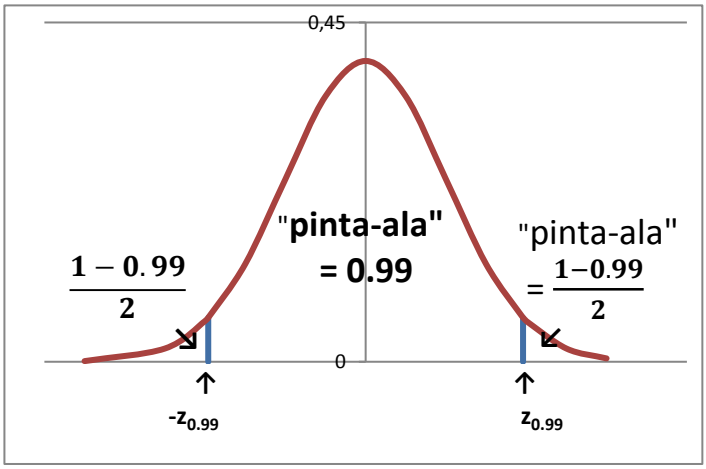
$= \phi(z_{0.99}) - (1 - \phi(z_{0.99}))$

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

josta saadaan

$$\phi(z_{0.99}) = \frac{0.99+1}{2} = 0.995$$

tai kuviosta



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

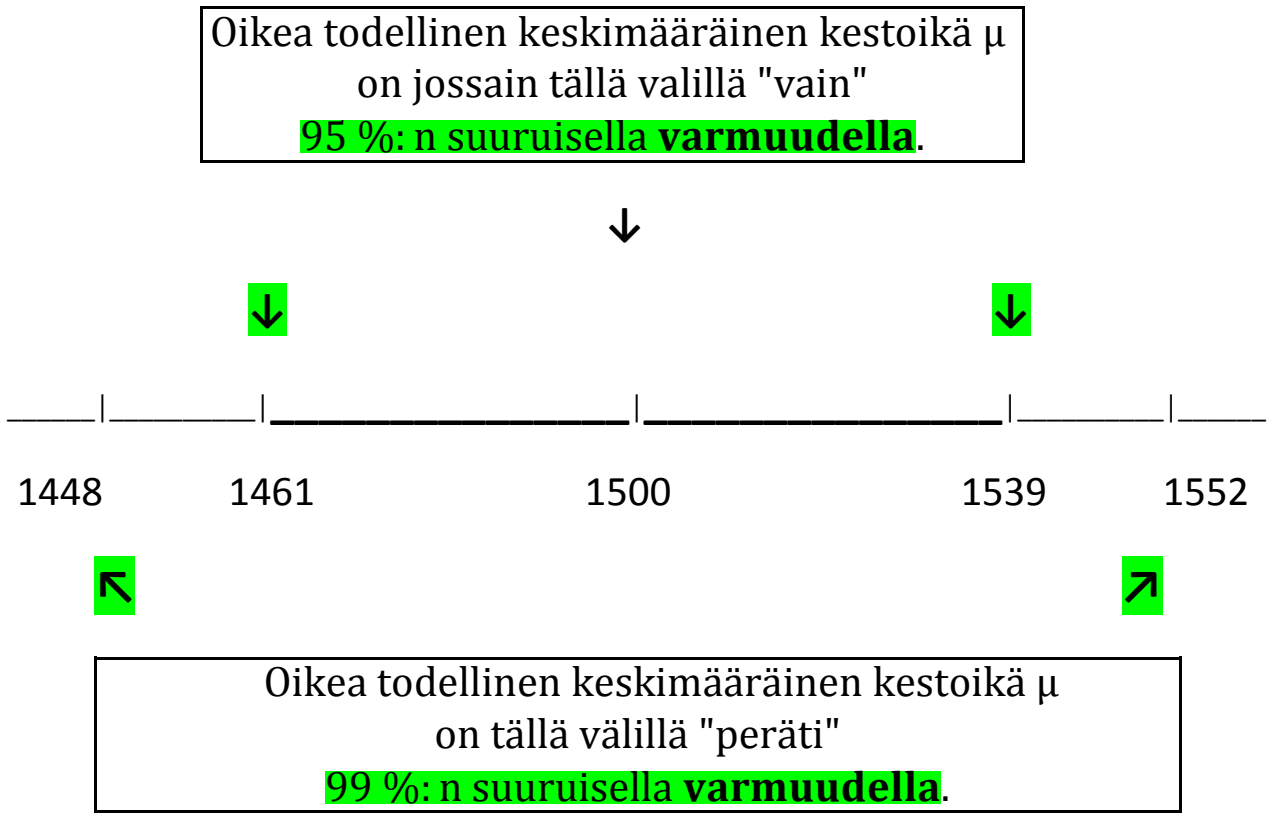
Luottamusvälin päätepisteet ovat nyt

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

99 % luottamusväli todelliselle keskimääräiselle kestoiälle μ on

(1500 h - 51.52 h, 1500 h + 51.52 h) ≈ (1448 h, 1552 h).

Kun verrataan aikaisempaan, näkyy:



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

- mutta olisikin poimittu 400 suuruinen otos, jossa olisi saatu (vaikkapa) keskiarvoksi $\bar{x}$ = 1510 h ja keskihajonnaksi s=205 h ($\approx$ 200 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

$$1510 \pm 2.576 \cdot \frac{205\ \text{h}}{\sqrt{\mathbf{400}}} = 1510\ \text{h} \pm 26.40\ \text{h} \cong 1510 \pm \mathbf{26\ h}$$

↑

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

Kun otoskoko n kasvaa, keskivirhe ”sattuman pelivara”

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{\mathbf{n}}}$ (ja sen estimaatti $\frac{s}{\sqrt{\mathbf{n}}}$) 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 $\mathbf{X \sim N(\mu, \sigma^2)}$ ja

ja perusjoukosta poimitaan n:n suuruinen otos **palauttamatta**, otoskeskiarvon generoiva otantajakauma on

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

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

$$\bar{x} \pm z_c \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$, 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.**

Tässäkin väli on vain likimäärin oikea luottamustasoa c vastaava luottamusväli, kun todellinen hajonta σ korvataan estimaatillaan s. Tulos on kuitenkin riittävän hyvä, kun otoskoko on ”suuri”, tässäkin n > 30.

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.

Esim. Yrityksen 700 työntekijästä poimittiin palauttamatta 60 suuruinen otos. Otokseen osuneiden viikoittain työaikana sähköpostiin käyttämän ajan keskiarvo oli $\bar{x}$ = 8.6 h ja hajonta s = 6.7 h. Määrää 95 %:n luottamusväli työntekijöiden sähköpostiin keskimäärin käyttämälle ajalle.

Vaikka jakauma perusjoukossa ei olisikaan normaalinen, otoskoko on tarpeeksi suuri ja luottamusväli saadaan riittävän oikein rajoista

$$8.6 \pm 1.96 \cdot \frac{6.7}{\sqrt{60}} \sqrt{\frac{700-60}{700-1}} = 8.6 \pm 0.83.$$

95 %:n varmuudella todellinen keskiarvo μ on välillä

(8.6 - 0.8, 8.6 + 0.8) = (7.6 h, 9.4 h).

2. luottamuskerroin määrätään t-jakaumasta

Kun 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.

Kun otoskoko n on ”suuri” ($n > 30$), otoskeskihajonta s vastaa riittävän hyvin todellista hajontaa σ ja luottamusväli voidaan määrätä kuten edellä.

Menetelmä ei kuitenkaan ota huomioon ”lisääntynyttä sattuman vaikutusta” verrattuna (käytännössä mahdottomaan tilanteeseen), missä todellinen hajonta σ tunnettaisiin ja todellisuudessa arvion luotettavuus on vähän pienempi kuin käytetty luottamustaso.

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.

jos σ olisi tunnettu **tuntematon**

Otosta poimittaessa

Sattuma on vaikuttanut vain keskiarvon suuruuteen.

Sattuma on päässyt vaikuttamaan sekä keskiarvoon että hajontaan.

↘ ↘ ↓

$$\bar{x} \pm z_c \frac{\sigma}{\sqrt{n}} \qquad \bar{x} \pm z_c \frac{s}{\sqrt{n}}$$

- Kun σ täytyy korvata otoskeskihajonnalla 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, …, $Z_n \sim N(0,1)$ ja ovat riippumattomia,

niin sanotaan, että satunnaismuuttuja

$$\mathbf{t(n) = \frac{X}{\sqrt{\frac{1}{n}\sum Z_i^2}}}$$ (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.

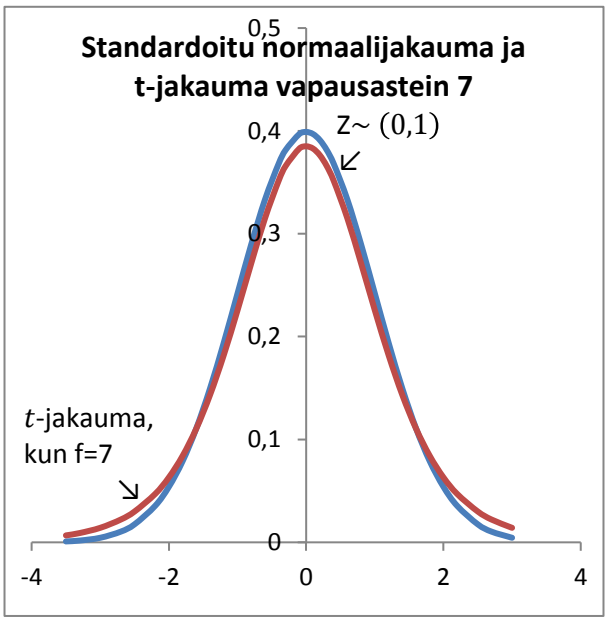
(Formulas → More Functions → Statistical → T.DIST)

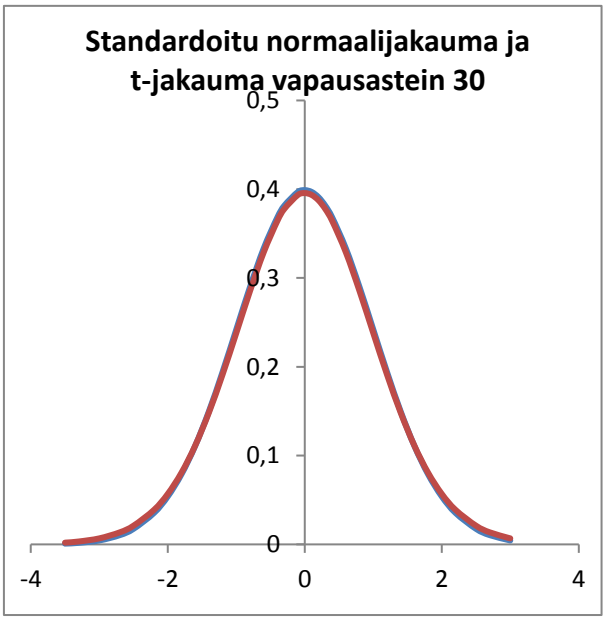
Esimerkiksi $P(t(10) \leq 2.228) = F_{t(10)}(2.228) = 0.974994 \approx 0.975$.

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(μ, σ^2) ,

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

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

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**.

<u>Huom.</u> 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 **keskimääräisestä** ↙ suuruudesta

sovelluksessa $\mathbf{t = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}}$ 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 **X** $\sim$**N(μ, σ^2) perusjoukossa**.

- Perusjoukosta poimitaan n:n suuruinen otos, josta lasketaan

sekä μ:n estimaatti otoskeskiarvo $\bar{x}$ **että** σ:n estimaatti keskihajonta s.

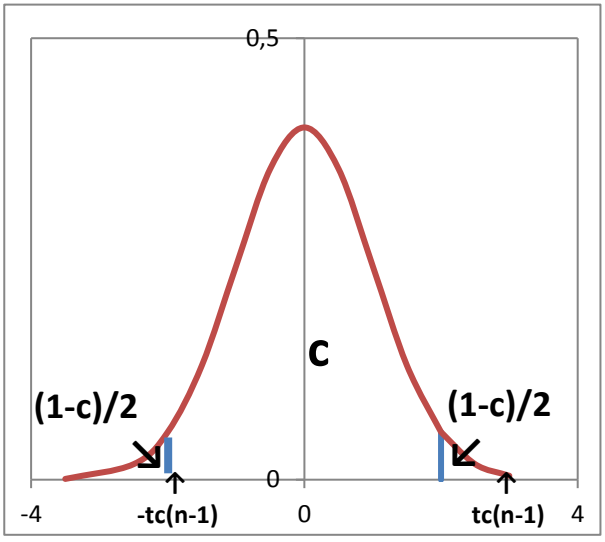
Todellisen perusjoukossa olevan keskimääräisen suuruuden μ **luottamusvälin päätepisteet luottamustasolla c** ovat

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

luottamuskerroin $t_c(n-1)$

määrätään t-jakaumasta ehdosta:

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



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 ”paremman puutteessa” 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 95 % varmuudella jäsenten käyttämä

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) 3) 2)

$\bar{x} \pm$ luottamuskerroin $\cdot$ 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 todella kummallisesta syystä) tiedetään.

Käytännössä tätä tapausta tarvitaan kuitenkin optimaalista otoskokoa arvioitaessa, missä σ:n suuruudesta tarvitaan ”valistunut arvaus”. Tätä käsitellään vähän myöhemmin.

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

3) Luottamuskerroin katsotaan

a) t-jakaumasta, jos (kun) 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 π 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 π alueen kotitalouksissa.

200 suuruisessa otoksessa 32.5 % kotitalouksista käytti R:ää.

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

- Otoksesta saatu suhteellinen osuus $\hat{p} = 0.325$ on :n paras piste-estimaatti.

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

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

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

<u>kaikissa muissa tapauksissa</u>: paitsi

<u>kun otos poimitaan palauttamatta N:n suuruisesta perusjoukosta</u>:

$\hat{p} = 0.325$ estimoi ”hyvin” tuntematonta π:n arvoa.

↙ ↘

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

↕ Varianssista saadaan keskivirhe. ↕

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

↑ ↑

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

Otantajakaumat perustuvat normaaliapproksimaatioon, jossa vaaditaan,

että $n\pi > 5$ ja $n(1-\pi) > 5$.

Näin on oltava myös luottamusväliä määrättäessä. (π 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 π:stä

tieto vaaditusta varmuuden asteesta

sattuman ”pelivara”

Siis tässä:

Perusjoukossa tilastoyksiköiden ominaisuuden A (esim. R:n käyttö) todellinen suhteellinen osuus $\boldsymbol{\pi}$ **on tuntematon**.

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

Luottamustasolla c todellisen suhteellisen osuuden $\boldsymbol{\pi}$ luottamusvälin päätepisteet ovat, kun

- otos poimitaan **palauttaen** tai

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

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

↗ ↑ ↖

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

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

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.

- n $\hat{p}$ = 200·0.325 = 65 (käyttäjien määrä) $>$ 5 ja

n(1- $\hat{p}$) = 200·(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 π 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.

Samasta otoksesta 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 $\boldsymbol{\sigma_{\hat{P}}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ suuruuden määrää **$\hat{p}$(1- $\hat{p}$)**, joka saa maksimiarvon, kun $\hat{p} = 0.5$.

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 **kokonaismäärän 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 **luottamusväli** niiden tilastoyksiköiden **kokonaismäärälle perusjoukossa,** joilla on ominaisuus A, saadaan

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 π 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

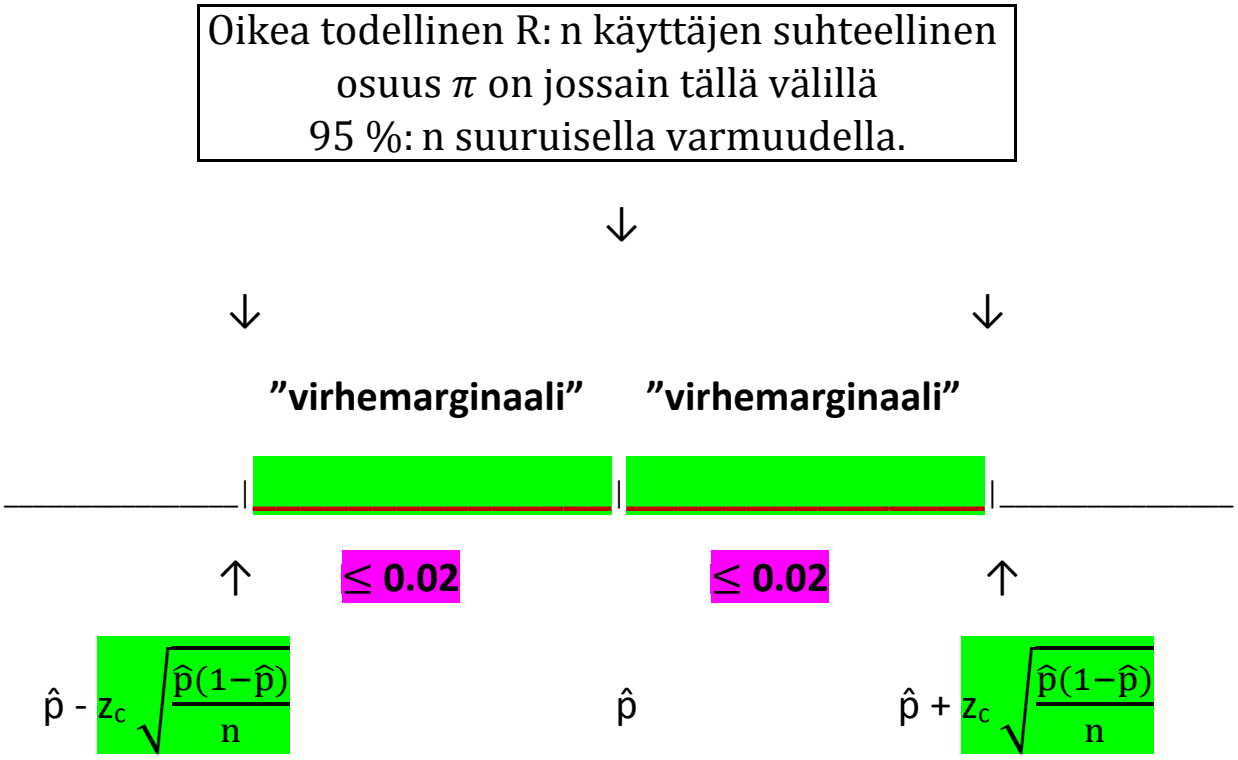
- **luotettavuuden** on oltava edelleen 95 % tasolla, mutta

- **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 π

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



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

Keskivirheessä $\boldsymbol{\sigma_{\hat{p}}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ tarvittaisiin oikea π: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** $\hat{p}$:n suuruudesta.

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

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

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

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

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

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 π: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**)

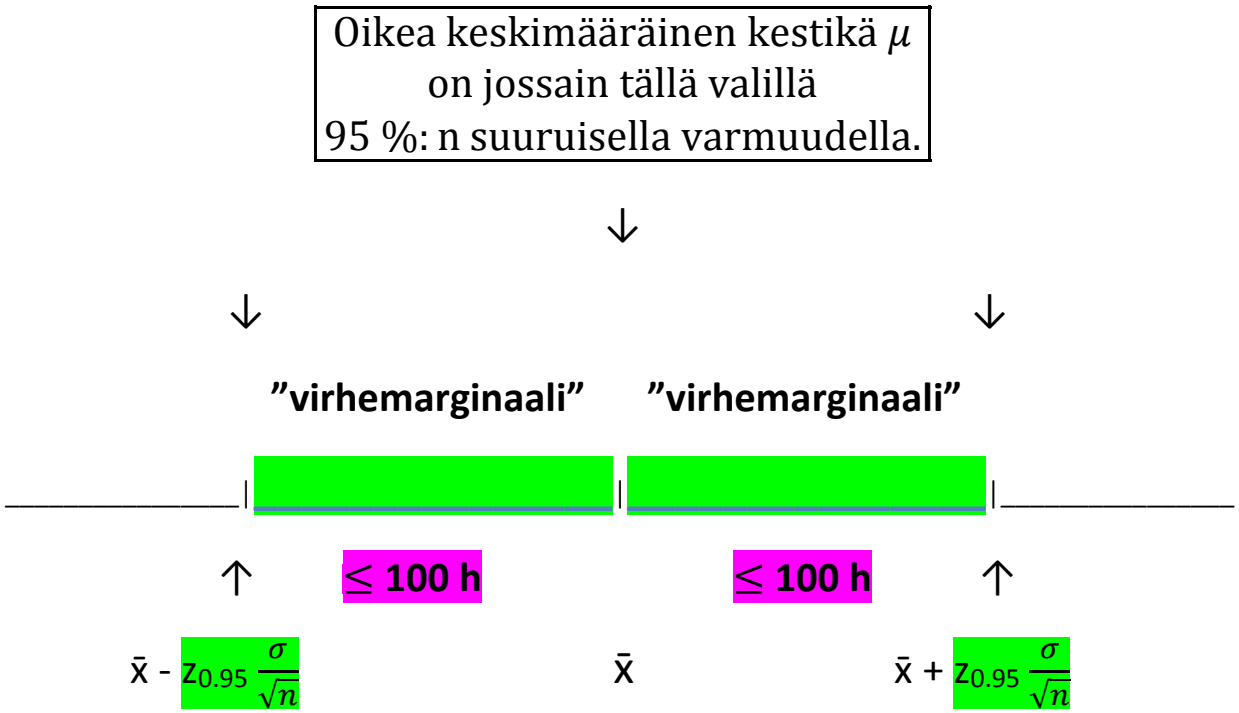
todellinen keskimääräinen kestoikä μ kaikkien vekottimien perusjoukossa ja otoksesta saatava kestoiän arvio $\bar{x}$

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!

5 Hypoteesien testaamisesta

Hypoteesien testaaminen on myös otoksesta perusjoukkoon päin tehtävää päättelyä.

Estimoinnissa yritetään selvittää parhaalla mahdollisella tavalla jonkin perusjoukossa ”piilossa” olevan parametrin suuruus.

Testaamisessa näkökulma on hieman erilainen ja myös rajoitetumpi:

- Lähtökohtana on jokin perusjoukkoa koskeva väite (**hypoteesi**).

- Otoksen perusteella tutkitaan, onko tämä hypoteesi **tosi vai ei**.

Esim. (jatkoa) Juomien makutestissä tulos määräytyy kokeella, jossa maistaja yrittää erottaa oman suosikkijuomamerkkinsä kolmen vastaavanlaisen juoman joukosta.

Maistaminen toistetaan 13 kertaa

ja tulos on oikein tunnistettujen kertojen määrä.

Arvostettu juoma-asiantuntija ekonomisti E osallistuu juomien makutestiin ja tunnistaa 13 yrityksestä 10 kertaa juoman oikein.

Saiko ekonomisti E tuloksen vai arvaamalla?

Jos tunnistaminen olisi arvaamista, niin ***nollahypoteesi***

H_0: π = P(Vastaus on oikein.) = 1/3

on tosi.

Jos ekonomisti E pystyy tunnistamaan juoman arvaamista paremmin, niin ***vastahypoteesi***

H_1: $\pi > 1/3$

on totta.

Tässä kokeen järjestäjät ovat ekonomisti E:n aikaisempien suoritusten perusteella ”täysin varmoja”, että kokelaan tunnistuskyky ei ainakaan ole arvaamista heikompi ($\pi < 1/3$).

Silloin ongelmaa voidaan tarkastella ***1-suuntaisesti*** arvausta parempaan suuntaan.

Arvosteluraadin on valittava:

- **Jääkö** maistajakokelaille asetettu lähtökohta $\mathbf{H_0}$ (arvaa vain) **voimaan**,

- vai **hylätäänkö** $\mathbf{H_0}$ ja uskotaan, että $\mathbf{H_1}$ (kyllä se tunnistaa) on totta.

Maistajaseuran raati luopuu nollahypoteesista H_0 vain, jos näyttö tätä arvaamis-oletusta vastaan on todella vahva.

Seuran jäsenyys on hyvin haluttu, ja pääsyehdoksi on asetettu **päätössääntö**:

Jos kokelas voisi saada pelkästään **arvaamalla** saavuttamansa tuloksen **alle** α = **5 % todennäköisyydellä**,

niin $\mathbf{H_0}$ **hylätään** ja kokelas otetaan jäseneksi.

On tapana sanoa, että tulos on ***melkein merkitsevä***, ja kokelas saa arvonimen MSM*.

Tätä päätössääntöä käyttäessään seura hyväksyy **riskin**, että 100:sta pelkästään arvaamalla maistavasta kokelaasta keskimäärin 5 onnekasta tulee **vain sattumalta** valituiksi,

eli ***H_0 hylätään, vaikka se on totta***. Tällaista valitettavaa, mutta sattumalta mahdollista, tapahtumaa sanotaan (H_0:n) ***hylkäämisvirheeksi*.**

Oleellista on, että seura pystyy sietämään $\alpha = 0.05$ suuruisen (H_0:n) ***hylkäämisvirheen*** tekemisen ***riskin*** MSM* arvonimeä antaessaan.

Jos H_0: $\pi = 1/3$, olisi tosi,

niin sattuma määrää kokeessa realisoituvan oikeiden arvausten lukumäärän binomijakauman $X \sim \text{Bin}(13, 1/3)$ mukaan.

Frekvenssifunktiosta saadaan todennäköisyys, että maistaja saavuttaa

vähintään **havaitun** tuloksen 10 oikein ↓

arvaamalla eli ehdolla, että **H_0 olisi tosi** ↙

$$\mathbf{p = P(X \geq 10 \mid \pi = 1/3)} = P(X = 10) + P(X = 11) + P(X = 12) + P(X = 13)$$

$$= \binom{13}{10} \cdot \left(\frac{1}{3}\right)^{10} \cdot \left(\frac{2}{3}\right)^{3} + \binom{13}{11} \cdot \left(\frac{1}{3}\right)^{11} \cdot \left(\frac{2}{3}\right)^{2}$$

$$+ \binom{13}{12} \cdot \left(\frac{1}{3}\right)^{12} \cdot \left(\frac{2}{3}\right)^{1} + \binom{13}{13} \cdot \left(\frac{1}{3}\right)^{13} \cdot \left(\frac{2}{3}\right)^{0}$$

$\approx 0.00165 = 0.165\ \%$.

Siis keskimäärin vain alle 2 kertaa tuhannesta näin hyvä tulos saavutetaan pelkästään arvaamalla.

Koska **p = 0.00165 $<$ 0.05 = α**,

asetetun päätössäännön mukaan **H_0 hylätään**.

Hylkäämisvirheen riski(, jota myös lyhyesti sanotaan ***p-arvoksi***)

p = 0.00165 alittaa jopa 1 %:n rajan.

Silloin sanotaan, että tulos on (seuran SM** arvonimeen oikeuttava)

(tilastollisesti) ***merkitsevä***.

Kun arvaamishypoteesin H_0 hylkäämisen päätössäännössä $\alpha = 0.01$, keskimäärin vain yksi ”Hannu Hanhi” sadasta saa arvaamalla SM** tittelin.

Jos H_0:sta luopumisen päätössäännöksi on asetettu ankara (ESM*** tasoon oikeuttava) arvo $\alpha = 0.001$,

p-arvo 0.00165 ylittää sen ja H_0 jää voimaan.

Tulos **ei ole *erittäin merkitsevä***.

Huom. Aikaisemmassa on, kuten on tavallista, binomijakauman toisen parametrin, ”onnistumistodennäköisyyden” symboli ollut p. Samaa merkintää käytetään myös usein jonkin ominaisuuden A suhteellisesta osuudesta perusjoukossa.

Sama merkintä **p on vakiintunut hylkäämisvirheen tekemisen riskin** symboliksi.

Jotta asiat eivät sekoitu keskenään, nyt käytetään tässä suhteellisesta osuudesta symbolia π.

Hypoteesien asettelusta

Ongelma esitetään kahden perusjoukkoa koskevan hypoteesin

- ***nollahypoteesin*** H_0 ja

- ***vastahypoteesin*** H_1

avulla.

Nollahypoteesi on usein ”varovainen perustilanne”:

Esim. a) Juomien tunnistustestissä

H_0: π = P(Tunnistaa juoman oikein.) = 1/3, siis pelkästään arvaa.

b) Sanomalehden suunnittelemassa äänestäjien poliittisen kannan tutkimuksessa (ks. edellä)

H_0: π = 0.35 eli puolueen Ö on pysynyt samana kuin edellisissä vaaleissa

(, vaikka ehkä vahvastikin epäiltäisiin sen pienentyneen).

c) Tutkitaan, onko suklaalevyn todellinen keskipaino μ tehtaan ilmoittama 100 g.

H_0: μ = 100 g eli lähtökohtana on väitetty arvo(, vaikka sitä epäiltäisiinkin).

d) Alueen kotitalouksista tehdyssä markkinatutkimuksessa

- 1500 suuruisessa otoksessa otokseen osuneista kotitalouksista 30 % käyttää hyödykettä Ö.

Mainoskampanjan jälkeen tehtiin uusi tutkimus, jossa

- 1400 suuruisessa otoksessa 34 % kotitalouksista käytti Ö:tä.

Voidaanko tämän perusteella väittää, että Ö:n käyttö on suurentunut kaikkien perusjoukon kotitalouksien joukossa?

Vaikka uskottaisiin, että 4 % -yksikön ero ei ole ”sattuman leikkiä”, niin kuitenkin nollahypoteesi asetetaan tylsän varovaisesti:

H_0: $\pi_1 = \pi_2$ eli käyttäjien suhteellinen osuus koko perusjoukossa

ennen (π_1) ja jälkeen (π_2) kampanjan on yhtä suuri.

e) Tutkittiin uuden kolesterolilääkkeen vaikutusta.

Kolmen kuukauden ajan lääkittiin ***koeryhmän*** 200 ja ***vertailuryhmän*** 150 korkeasta kolesterolista kärsivää koehenkilöä. Kokeen jälkeen olivat

koeryhmän 200:n lääkettä saaneen kolesteroliarvojen

keskiarvo $\bar{x}_1$ = 5.2 mmol/l ja keskihajonta s_1 = 1.2 mmol/l ja

vertailuryhmän 150:n lumelääkettä saaneen

keskiarvo $\bar{x}_2$ = 5.7 mmol/l ja keskihajonta s_2 = 1.6 mmol/l.

Vaikka lääkkeen kehittäjät tietenkin toivovat tuotteensa olevan hyvä, nollahypoteesi asetekaan kuitenkin varovaisesti

H_0: $\mu_1 = \mu_2$ eli todellinen keskimääräinen kolesterolimäärä on yhtä suuri.

Esim. (jatkoa) Yrityksen Y työntekijöistä poimittiin otos, jonka avulla selvitettiin suhtautumista tulospalkkauksen käyttöön ottoon yrityksessä. Saatiin tulokset:

o_{ij}	Kielteinen	Neutraali	Myönteinen	Yhteensä
Alle 40-v.	96	174	159	429
Yli 40-v.	117	155	122	394
Yhteensä	213	329	281	823

Voidaanko tämän perusteella päätellä, riippuuko suhtautuminen tulospalkkaukseen työntekijän iästä?

Nollahypoteesina on tässä (kuten aina riippuvuutta tutkittaessa)

H_0: ***Muuttujat*** ikä ja suhtautuminen tulospalkkaukseen ovat

***riippumattomia*.**

Vastahypoteesissa H_1 kerrotaan, missä tilassa perusjoukko on, jos H_0 ei ole tosi.

Jos H_1:ssä ei ilmoiteta (ei uskalleta, pystytä) ”poikkeaman suuntaa” H_0:n mukaiseen tilanteeseen verrattuna, on testaus ***2-suuntainen***.

Jos poikkeaman suunta ilmoitetaan, testi on ***1-suuntainen***.

1-suuntaisessa testauksessa tarvitaan **havaintoaineiston ulkopuolelta empiiristä lisätietoa**, joka oikeuttaa sulkemaan pois poikkeaman toiseen suuntaan H_0:n mukaisesta tilanteesta.

Esim. (jatkoa edelliseen) a) Juomien tunnistustestissä

vastahypoteesiksi on asetettava

H_1: $\pi \neq 1/3$ eli oikein osumisen todennäköisyys poikkeaa arvaamisesta,

jos kokelas on ”tavallinen tallaaja”, jonka kyvyistä ei ole mitään koetilanteen ulkopuolelta saatua lisätietoa.

Silloin on mahdollista, että ero H_0:n mukaiseen tilanteeseen on kahteen suuntaan:

- Maistaja voi tunnistaa juoman arvaamista paremmin,

- mutta mahdollista on myös, että hän tunnistaakin juoman arvaamista huonommin.

Ekonomisti E on yleisesti tunnettu hienona juoma-asiantuntijana, ja hänet on vihdoin saatu hakemaan juomaseuran jäsenyyttä.

Aikaisemman perusteella pidetään ”täysin varmana”, että tässä tapauksessa tunnistuskyky ”ei voi olla” arvaamista huonompi. Nyt vastahypoteesi voidaan tehdä 1-suuntaisesti.

H_1: π = P(Vastaa oikein.) $>$ 1/3 eli tunnistaa arvaamista paremmin.

b) Puolueen skandaaleihin sekaantumisesta johtuva ”poliittinen ilmasto” oikeuttanee sulkemaan pois mahdollisuuden, että puolueen kannatus olisi ainakaan suurentunut. Vastahypoteesiksi asetetaan 1-suuntaisesti

H_1: $\pi < 0.35$ eli kannatus on pienentynyt edellisistä vaaleista.

Vastahypoteesin H_0 asettelussa ajatuksena on:

- 1-suuntaista testiä **saa käyttää**, jos ”varmaa” lisätietoa on käytettävän havaintoaineisto ulkopuolelta.

- 2-suuntaista testiä **joudutaan käyttämään**, jos tällaista lisätietoa ei ole ja joudutaan olemaan varovaisempia.

c) Jos ilman mitään erityistä ennakkotietoa vain rutiininomaisesti tutkitaan suklaalevyjen painoa, on vastahypoteesi 2-suuntainen

H_1: $\mu \neq$ 100 g eli paino poikkeaa ilmoitetusta jompaankumpaan suuntaan.

Käytettävissä voi olla jotain lisätietoa koneiden virheellisestä toiminnasta tms., mikä selvästi voi vaikuttaa tuotteen laatuun. Silloin saattaa olla riittävästi perusteita 1-suuntaiseen testaamiseen.

d) Jos ollaan varovaisia, asetetaan 2-suuntaisesti

H_1: $\pi_1 \neq \pi_2$ eli käyttäjien suhteellinen osuus on muuttunut.

Jos vastaavanlainen kampanja on aikaisemmin toiminut hyvin eikä käyttäjien osuus ole koskaan ainakaan pienentynyt, voi olla riittävästi perusteita 1-suuntaiseenkin vastahypoteesiin

H_1: $\pi_1 < \pi_2$ eli käyttäjien osuus oli pienempi ennen kampanjaa.

e) Uutta lääkettä tutkittaessa on ehkä parasta olla varovainen ja tehdä vastahypoteesi 2-suuntaisesti

H_1: $\mu_1 \neq \mu_2$ eli keskimääräiset kolesteroliarvot ovat eri suuria

koe- ja vertailuryhmällä.

f) Otoksessa näyttäisi olevan poikkeamaa siihen suuntaan, että nuoremmat työntekijät olisivat myönteisempiä tulospalkkaukselle.

Vastahypoteesin asettelussa kuitenkaan **ei saa ”vilkuilla” otokseen** ja sulkea sen perusteella pois toista suuntaa. Jos poikkeama H_0:n mukaisesta tilanteesta on vain ”sattuman leikkiä”, se olisi voinut **yhtä hyvin** sattua toiseen suuntaan.

Tässä ilman muita lisätietoja oletetaan tylsästi

H_1: Suhtautuminen tulospalkkaukseen **riippuu** iästä.

Testaamisen lähtökohtana on nollahypoteesi H_0.

- H_0:aa pidetään totena (***H_0 hyväksytään***), jos otoksen ”perusjoukon tilasta” antama informaatio ei poikkea ”kohtuuttoman paljon” tilanteesta, joka otoksessa pitäisi keskimäärin olla H_0:n ollessa voimassa.

- H_0:aan ei uskota (***H_0 hylätään***), jos otoksessa havaittu tilanne on vahvasti ristiriidassa sen kanssa.

Siis H_0:an ei enää uskota,

- kun sattuman käyttäytymistä säätelevä ”mekanismi” eli otoksesta

laskettavien tunnuslukujen otantajakaumat

- eivät pysty tuottamaan otoksessa todella realisoitunutta tilannetta kohtuullisen suurella todennäköisyydellä,

- jos perusjoukko olisi H_0:n väittämässä tilassa.

Silloin päätellään, että otoksessa havaittu asetelma on lähtöisin toisenlaisesta, H_1:n mukaisessa tilassa olevasta perusjoukosta.

Otos on vain suppea osa perusjoukosta, ja nollahypoteesin hylkäämis- tai hyväksymispäätöstä tehtäessä ei koskaan tiedetä varmasti, onko tehty päätös oikea.

- Jos H_0 hylätään, vaikka se todellisuudessa onkin totta, tehdään ***hylkäämisvirhe*** **(1. lajin virhe)**.

- Jos H_0 hyväksytään, vaikka se todellisuudessa onkin väärä, tehdään ***hyväksymisvirhe*** **(2. lajin virhe)**.

Tilastollinen analyysi on **päättelyä epävarmuuden vallitessa**, ja nämä virheet ovat aina mahdollisia.

- Vaikka H_0 olisi tosi, ”vain sattumalta” otokseen voivat realisoitua sellaiset havainnot, jotka ”puhuvat H_0:a vastaan”.

- Samoin sattuma voi valita otokseen sellaiset havainnot, jotka tukevat H_0:a, vaikka tämä olisikin väärä.

Siis testaamisessa voidaan päätyä neljään tulokseen:

	Todellinen tilanne	perusjoukossa:
Päätelmä:	**H_0 on tosi.**	**H_0 on väärä.**
H_0 hyväksytään	Päätös on **oikea.**	Tehdään ***hyväksymisvirhe*.**
H_0 hylätään.	Tehdään ***hylkäämisvirhe***.	Päätös on **oikea.**

- Kun valitaan sopivaa testiä tutkittavana olevien hypoteesien testaamista varten, valinnassa on tärkeää ***testin voimakkuus***.

Tämä tarkoittaa (väljästi määriteltynä), kuinka hyvin testi pystyy hylkäämään väärän nollahypoteesin eli välttämään hyväksymisvirheen.

Alkeita opeteltaessa valittavia vaihtoehtoja on vähän ja ne ovat kaikki oikein hyviä.

Testien voimakkuuksien tutkimiseen ei syvennytä tässä tarkemmin.

- Kun testausmenetelmä on valittu jotakin yksittäistä testaustilannetta varten, keskitytään hylkäämisvirheen tekemisen todennäköisyyden laskemiseen.

Testaaminen on ”väittelyä” perusjoukon tilasta.

1. ”Oletetaan nyt sitten, että H_0 olisi voimassa, ja katsotaan … ”

2.”Katsominen”

- Sattuma määrää sääntöjensä (otantajakaumien) mukaan otoksen sisällön. Otoksessa havaittava tilanne voi olla (kuitenkin vain pienellä todennäköisyydellä) hyvinkin erikoinen H_0:n kannalta, vaikka se olisi tosi.

- Kuitenkin on loogista toimia niin, että H_0 hylätään, jos havaittu tilanne on hyvin epätodennäköinen H_0:n ollessa voimassa.

- Otantajakaumien avulla lasketaan (tai ainakin arvioidaan), kuinka suuri on

hylkäämisvirheen tekemisen riski, jota sanotaan myös ***merkitsevyystasoksi***. Se on ehdollinen todennäköisyys

$p = P(H_0$ hylätään. $|$ H_0 onkin tosi.)

↖ hylkäämisvirhe ↗

3. Jos hylkäämisvirheen riski, **p-arvo,** on ”pieni”, H_0 **uskalletaan** hylätä.

- Riippuu täysin tutkittavan ongelman luonteesta, kuinka ”pieni” merkitsevyystason p on oltava, ennen kuin H_0 uskalletaan hylätä.

- Usein päätetään etukäteen, kuinka suuri hylkäämisvirheen riski voidaan enimmillään sietää, jos H_0 päätetään hylätä.

Tällaisina kynnysarvoina käytetään yleisesti

rajoja (α =) 0.05, 0.01 ja 0.001.

Jos hylkäämisvirheen riskin p ylärajaksi on valittu ***α = 0.05*** ja H_0 hylätään ($p<0.05$), niin tulosta sanotaan ***melkein merkitseväksi*** ja sanotaan,

että H_0 hylätään 5 %:n merkitsevyystasolla.

Jos H_0 hylätään ***α =1 %:n*** merkitsevyystasolla ($p<0.01$), tulos on ***tilastollisesti merkitsevä*** ja

jos H_0 hylätään ***α = 0.1 %:n*** merkitsevyystasolla ($p<0.001$), tulos on ***erittäin merkitsevä***.

Seuraavassa on joitain tärkeitä erikoistapauksia hypoteesien testaamisesta.

Suhteellisen osuuden testi yhden perusjoukon tapauksessa

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

Paikallislehti tekee otantatutkimuksen, jossa selvitettiin (mm.), onko puolueen Ö kannatus pienentynyt.

Puolueen epäillään sotkeutuneen törkyiseen lahjusskandaaliin ja ”poliittisen ilmaston” perusteella voidaan testata 1-suuntaisesti

H_0: $\pi = 0.35$ ($=\pi_0$) eli Ö:n kannatus ei ole muuttunut.

H_1: $\pi < 0.35$ eli Ö:n kannatus on pienentynyt.

Etukäteen päätetään, että

testaamisessa voidaan sietää korkeintaan α = 1 %:n suuruinen hylkäämisvirheen tekemisen riski.

Kunnan N = 40000 äänestysikäisestä poimittiin n = 1500 suuruinen otos palauttamatta.

Otoksessa Ö:tä kannatti $\hat{p}$ = 32 % vastaajista.

Jos nollahypoteesi H_0: π = 0.35 pitäisi paikkansa,

niin sattuma olisi generoinut otoksen otantajakauman

H_0:ssa oletettu π_0 N n

↙ ↓ ↘ ↓ ↓

$\hat{P} \sim N(0.35, \frac{0.35(1-0.35)}{1500} \cdot \frac{40000-1500}{40000-1})$ = $N(0.35, 0.0121^2)$ mukaan.

↑ ↑

n N

Silloin on todennäköisyys, että otoksessa ”vain sattumalta” korkeintaan 32 % kannattaa Ö:tä

eli **hylkäämisvirheen tekemisen riski** on

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

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

$= 0.0066$

$p = 0.66\ \% < 1\ \% = \alpha$, joten H_0 uskalletaan hylätä.

Päätellään, että Ö:n kannatus on pienentynyt.

Esim. (jatkoa)

Otantatutkimuksessa kysyttiin myös, aikovatko haastateltavat ensi vuonna tilata tutkimuksen teettäneen lehden.

- Kuluvana vuonna kunnan äänestysikäisistä 60 % on tilannut lehden ja

- otoksen 1500:sta haastatellusta 62.0 % arveli, että lehti tilataan.

Voidaanko tämän perusteella päätellä, että tilaajien määrä on **muuttunut**?

Lehdessä ei ole ennen tutkimusta selvää käsitystä, onko lehden suosio muuttunut ja, jos on, niin mihin suuntaan.

Silloin on testattava 2-suuntaisesti ja hypoteesit ovat

H_0: π = 0.60 (=π_0) eli tilaajien osuus on edelleen sama kuin ennen.

H_1: $\pi \neq 0.60$ eli tilaajien osuus on muuttunut (kasvanut tai pienentynyt.)

Jos H_0 tulisi hylätyksi, vaikka mitään muutosta ei olisikaan tapahtunut, seuraukset eivät ole mitenkään vakavat. Päätetään testata **5 %:n merkitsevyystasolla**.

Jos H_0 on tosi eli tilaavia olisi edelleen 60 %, niin sattuma on tuottanut otoksen sisällön otantajakauman

$\hat{P} \sim N(\mathbf{0.60}, \frac{\mathbf{0.60}(1-\mathbf{0.60})}{1500} \cdot \frac{40000-1500}{40000-1}) = N(0.60, 0.0124^2)$ mukaan.

Jos otoksessa havaittu 62 % on vain ”sattuman leikkiä”, olisi yhtä hyvin voinut käydä niin, että tilaajia olisi ollut 2 % -yksikköä 60 %:ia vähemmän.

Hylkäämisvirheen tekemisen riski on silloin

Otoksessa sattui käymään näin.

↓

p = P($\hat{P} \leq 0.58$ tai $\mathbf{\hat{P} \geq 0.62}$ | $\pi = 0.60$)

↑

Yhtä hyvin olisi voinut käydä näin.

$$= P\left(\frac{\hat{P} - 0.60}{0.0124} \leq \frac{0.58 - 0.60}{0.0124}\right) + \mathbf{P}\left(\frac{\mathbf{\hat{P} - 0.60}}{\mathbf{0.0124}} \geq \frac{\mathbf{0.62 - 0.60}}{\mathbf{0.0124}}\right)$$

= P(Z ≤ - 1.61) + **P(Z ≥ 1.61)** = ϕ(- 1.61) + **1 - P(Z < 1.61)**

= 1 - ϕ(1.61) + **1 - ϕ(1.61)** = 1 – 0.9463 + **1 – 0.9463**

= 0.0537 + **0.0537** = 2·**0.0537** = 0.1074.

↑ ↑

2 – suuntaisessa testissä hylkäämisvirheen riski p on **aina** 2 – kertainen 1 – suuntaiseen testiin verrattuna

p = 0.1074 > 0.05 = α ja H_0:aa ”ei uskalleta” hylätä.

Huom. Tulos ei suinkaan tarkoita, että olisi saatu lopullinen totuus asiasta.

- Otoksen informaation ”todistusvoima H_0:aa vastaan” vain ei ole riittävä

- ja edelleen on pidettävä mahdollisena, että tilanne ei ole muuttunut.

Kuitenkin on voitu tehdä **hyväksymisvirhe**.

- Todellinen suhteellinen osuus π on voinut muuttua ”vähän”, mutta

- **testin voimakkuus** ei riitä paljastamaan muutosta

- näin **pienen otoksen** informaation avulla.

Yleensäkin testauksessa H_0:n voimaan jääminen ei ole välttämättä lopullinen totuus. Otoksen informaatio ei vain riitä ”kaatamaan” H_0:aa.

Jos epäillään, että H_0 ei ole sittenkään tosi, se paljastuu kyllä (lisää resursseja käyttämällä) suuremmasta otoksesta.

Suhteellisen osuuden testaamisen vaiheet

ovat myös yleisesti samat kuin esimerkeissä:

- **Perusjoukossa** ominaisuuden A suhteellinen osuus tilastoyksiköissä on $\boldsymbol{\pi}$

- ja n:n suuruisesta **otoksesta** on saatu π:n estimaatiksi **$\hat{p}$**.

1. Asetetaan hypoteesit:

Nollahypoteesi $\mathbf{H_0: \pi = \pi_0}$,

jossa π_0 on jokin oletettu, väitetty, aikaisempi, tms. arvo.

Vastahypoteesi $\mathbf{H_1: \pi \neq \pi_0}$,

jos joudutaan käyttämään 2-suuntaista testiä,

ja $\mathbf{H_1: \pi < \pi_0}$ tai $\mathbf{H_1: \pi > \pi_0}$,

jos otoksen ulkopuolelta saadun tiedon perusteella ”tiedetään” mahdollisen poikkeaman suunta H_0:n mukaisesta tilanteesta.

2. Pohditaan, kuinka suuri on suurin siedettävissä oleva **hylkäämisvirheen riski α korkeintaan.**

3. Oletetaan, että **H_0 olisi tosi**, ja selvitetään, minkä **otantajakauman**

mukaan sattuma silloin olisi generoinut otoksen.

Jakauma olisi

<u>kaikissa muissa tapauksissa</u>: paitsi

$$\hat{P} \sim N\left(\pi_0, \frac{\pi_0(1-\pi_0)}{n}\right)$$

↑

<u>kun otos poimitaan palauttamatta N:n suuruisesta perusjoukosta</u>:

$$\hat{P} \sim N\left(\pi_0, \frac{\pi_0(1-\pi_0)}{n} \cdot \frac{N-n}{N-1}\right)$$

↑

”Spekuloidaan” sillä, että π_0 olisi oikea suhteellinen osuus.

Otantajakauman avulla lasketaan

otoksessa havaitun ja vielä sitäkin poikkeavamman tilanteen (”hännän”) todennäköisyys

$P(\hat{P} \leq \hat{p} \mid \pi = \pi_0)$ (, jos $\hat{p} < \pi_0$,) tai $P(\hat{P} \geq \hat{p} \mid \pi = \pi_0)$ (, jos $\hat{p} > \pi_0$,)

4. Hylkäämisvirheen tekemisen riski p on

- edellä laskettu todennäköisyys, jos voidaan testata 1–suuntaisesti, ja

- ja p = **2**· $P(\hat{P} \leq \hat{p} \mid \pi = \pi_0)$ (tai **2**·$P(\hat{P} \geq \hat{p} \mid \pi = \pi_0)$),

jos joudutaan testaamaan 2-suuntaisesti.

5. Jos laskettu näin laskettu

hylkäämisvirheen riski **p** $<$ **α** = riskille asetettu yläraja,

niin **H_0 uskalletaan hylätä**.

Muuten H_0 jää voimaan.

Otantajakaumat perustuvat normaaliapproksimaatioon. Tässä vaaditaan vastaavasti kuin edellä,

että **n**· $\boldsymbol{\pi_0} > 5$ ja **n(1-** $\boldsymbol{\pi_0}$**)** > 5.

Ekonomisti E:n juomien tunnistuskykyä tutkittaessa testausasetelma oli aivan vastaava.

Siinä approksimointiehto ei täyty, kun $13 \cdot (1/3) \approx 4.3$.

Otantajakaumana käytettiin H_0:n mukaista binomijakaumaa.

Minkä tahansa parametrin arvon testaamisessa voidaan myös käyttää luottamusväliä apuna:

2-suuntaisen testin ja luottamusvälin yhteys

Esim. (jatkoa) Lehden otantatutkimuksessa kysyttiin, aikovatko haastateltavat ensi vuonna tilata tutkimuksen teettäneen lehden.

- Kuluvana vuonna kunnan lehden on tilannut 60 % äänestysikäisistä ja

- otoksen 1500:sta haastatellusta 62.0 % arveli, että lehti tilataan.

Voidaanko tämän perusteella päätellä, että tilaajien määrä on **muuttunut**?

Hypoteesit ovat samat kuin edelläkin. Siis tilaajien todellinen suhteellinen osuuden arvo π on

H_0: $\pi = 0.60$ kuten aikaisemminkin tai

H_1: $\pi \neq 0.60$ muuttunut.

Hypoteesit jätetään hetkeksi sivuun ja tutkitaan,

mitä otoksen perusteella voidaan sanoa π:n suuruudesta:

Kun tässä aiotaan testata α = 5 %:n merkitsevyystasolla,

- hyväksytään 5 %:n **riski** hylkäämisvirheen tekemiselle.

- Silloin toisaalta päättely on oikea 95 %:n **varmuudella**.

<u>95 %:n luottamusväli oikealle π:n arvolle:</u>

Otoskoko n = 1500 ja sitä kautta vapausasteluku f = 1500 - 1 = 1499 ovat tässä niin suuria, että luottamuskertoimena voidaan käyttää normaalijakaumasta saatua

$z_{0.95}$ = 1.96 (Vrt. $t_{0.95}(1499)$ = 1.961548 Excelistä.)

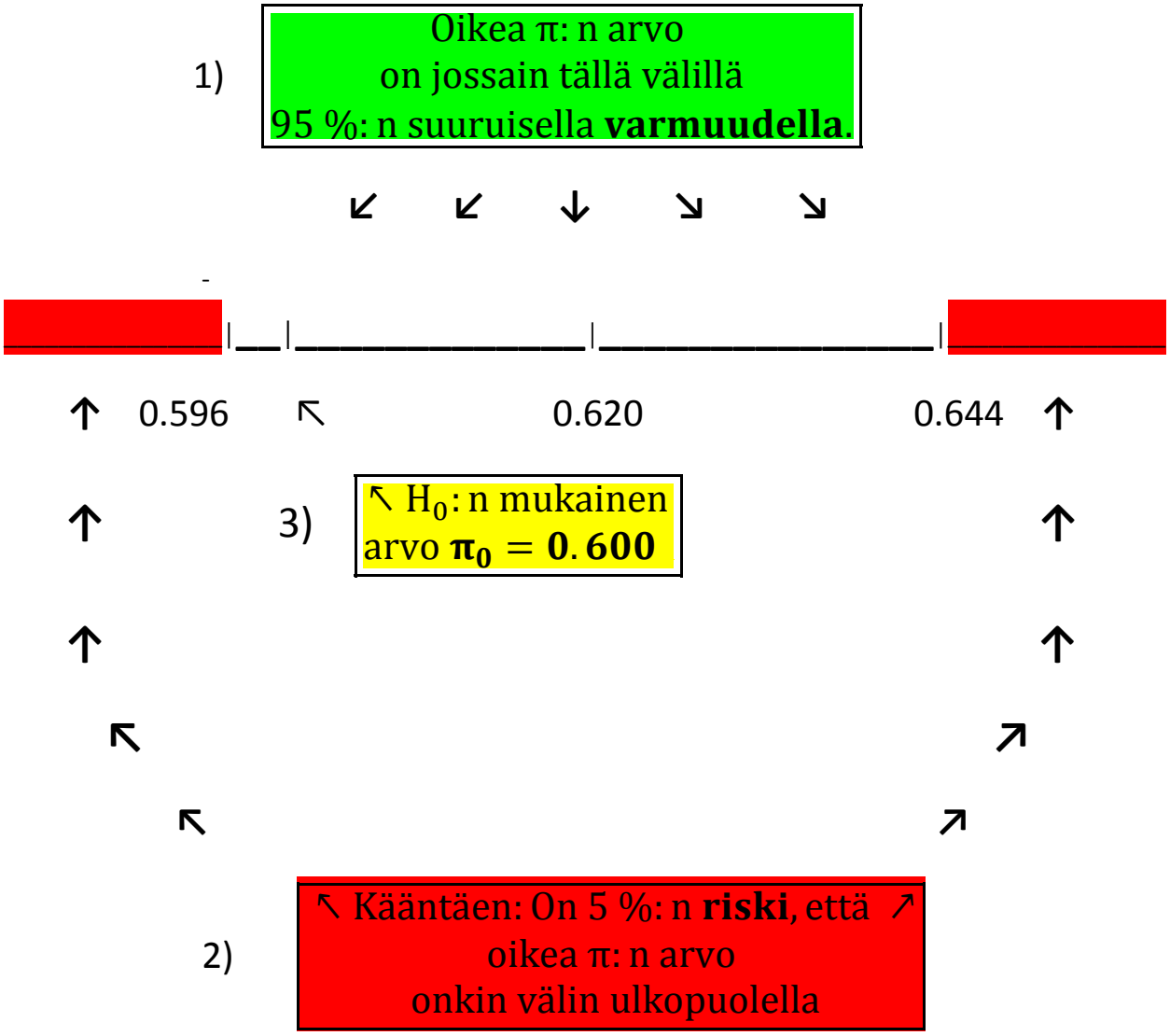
Päätepisteet ovat

$$0.620 \pm 1.96 \cdot \sqrt{\frac{0.620 \cdot (1-0.620)}{1500} \cdot \frac{40000-1500}{40000-1}} = 0.620 \pm 0.024$$

ja 95 %:n luottamusväli tilaajien todelliselle suhteelliselle osuudelle kunnassa on

(0.620 – 0.024, 0.620 + 0.024) = (0.596, 0.644).

Silloin päätellään:



4) H_0:n mukainen arvo π_0 = 0.60 on 95 %:n luottamusvälillä eli

”95 %:n varmuuden puitteisiin mahtuu vielä mahdollisuus, että π = 0.600.

5) Silloin on hylkäämisvirheen riski $p > 0.05 = \alpha$, jos H_0 hylättäisiin,

ja H_0 jää voimaan 5 %:n merkitsevyystasolla.

Menetelmä toimii myös yleisesti, kun testataan minkä tahansa parametrin arvon suuruutta 2-suuntaisesti:

- Jos suurin siedettävissä oleva hylkäämisvirheen tekemisen **riski on α**,

- tehdään parametrille luottamusväli **luottamustasolla c = 1 - α**.

- Jos H_0:n mukainen parametrin arvo on luottamusvälillä, H_0 jää voimaan.

Muuten H_0 hylätään.

Keskimääräisen suuruuden μ testi yhden perusjoukon tapauksessa

Esim. (jatkoa) Suklaatehdas väittää, että

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