

**TEKNINEN YHTEENVETO I****Otanta-asetelmat ja estimointiasetelmat****Perusjoukko ja muuttujat**

Äärellinen perusjoukko  $U = \{1, \dots, k, \dots, N\}$

*Tulosmuuttujan y tuntemattomat* arvot  $Y_1, \dots, Y_k, \dots, Y_N$

*Apumuuttujan z tunnetut* arvot  $Z_1, \dots, Z_k, \dots, Z_N$

**Perusjoukon parametrit**

Äärellisen perusjoukon  $U$  parametrit

*Kokonaismäärä*  $T = \sum_{k=1}^N Y_k = Y_1 + Y_2 + \dots + Y_N$

*Keskiarvo*  $\bar{Y} = T/N$

*Suhteellinen osuus*  $R = T_1 / T_2$

**Otanta-asetelma ja otos**

*Otos s* on perusjoukon  $U$  osajoukko

Perusjoukon  $U$  kaikkien mahdollisten  $n$  ( $n < N$ ) kokoisten otosten joukko  $S$

Toteutunut otos  $s = \{1, \dots, k, \dots, n\}$ , missä  $s$  on yksi mahdollisista otoksista joukossa  $S$

Otosyksiköt poimitaan soveltuvaa arpomismenettelyä eli *otantamenetelmää* (SRS, SYS, PPS) käyttäen

Otoksen  $s$  *poimintatodennäköisyys*  $p(s)$

Perusjoukon alkion  $k$  *sisältymistodennäköisyys*  $\pi_k$  ( $0 < \pi_k \leq 1$ )

***Otanta-asetelmaksi (sampling design),  $p(\cdot)$ , sanotaan niiden sääntöjen ja menetelmien kokonaisuutta, joilla otos poimitaan määritellystä perusjoukosta.***

**Perusjoukon parametrin  $\theta$  estimaattori  $\hat{\theta}$ :**

Laskentakaava tai laskenta-algoritmi

**Estimaattorin odotusarvo**  $E(\hat{\theta}) = \sum_{s \in S} p(s) \hat{\theta}_s$ *Harhaton* (unbiased) estimaattori:  $E(\hat{\theta}) - \theta = 0$ *Harha* (Bias):  $Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$ *Tarkentuva* (consistent) estimaattori:  $E(\hat{\theta})$  lähestyy parametria  $\theta$  kun  $n$  kasvaa, ja yhtyy parametriin, kun  $n = N$ .**Estimaatti:** Otoksesta laskettu estimaattorin numeerinen arvo**Estimaattorin asetelmavarianssi  $V(\hat{\theta})$ :**

$$V(\hat{\theta}) = \sum_{s \in S} p(s) (\hat{\theta}_s - E(\hat{\theta}))^2 = E(\hat{\theta} - E(\hat{\theta}))^2$$

missä otoksen  $s$  poimintatodennäköisyys on  $p(s) > 0$ **Estimaattorin keskineliövirhe** (Mean squared error MSE)

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = V(\hat{\theta}) + Bias^2(\hat{\theta})$$

**Varianssiestimaattori  $\hat{v}_{p(s)}$ :** Otanta-asetelmaspesifi analyttinen lauseke tai approksimatiivinen varianssiestimaattoriEstimoitu **keskivirhe:**  $s.e(\hat{\theta}) = \sqrt{\hat{v}(\hat{\theta})}$  (*standard error*)Estimaattorin estimoitu **suhteellinen keskivirhe** (*relative standard error*) eli **variaatiokerroin** (*coefficient of variation*):

$$c.v(\hat{\theta}) = \sqrt{\hat{v}(\hat{\theta})} / \hat{\theta} = s.e(\hat{\theta}) / \hat{\theta}$$

Estimoitu **asetelmakerroin** (design effect)  $deff(\hat{\theta}) = \frac{\hat{v}_{p(s)}(\hat{\theta})}{\hat{v}_{SRS}(\hat{\theta})}$ missä  $p(s)$  viittaa käytettyyn otanta-asetelmaan

SRS on yksinkertainen satunnaisotanta (WR tai WOR)

 $deff = 1$  Otanta-asetelma on **yhtä tehokas** kuin SRS $deff < 1$  Otanta-asetelma on **tehokkaampi** kuin SRS $deff > 1$  Otanta-asetelma on **tehottomampi** kuin SRS

## Yksinkertainen satunnaisotanta SRS

Sisältymistodennäköisyys  $\pi_k = n/N$  on vakio

**Kokonaismäärän  $T$  estimaattori** (harhaton)

$$\hat{t} = N\bar{y} = N \sum_{k=1}^n y_k / n,$$

missä  $\bar{y}$  on otoskeskiarvo ja  $N$  perusjoukon koko

$$\hat{t} = \sum_{k=1}^n w_k y_k = \sum_{k=1}^n \frac{1}{\pi_k} y_k = \frac{N}{n} \sum_{k=1}^n y_k,$$

missä  $w_k = N/n$  on otospaino (alkion  $k$  sisältymistodennäköisyyden  $\pi_k = n/N$  käänteisluku)

**Asetelmavarianssi (parametri) SRSWOR-poiminnalle**

$$V_{SRS}(\hat{t}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) \sum_{k=1}^N (Y_k - \bar{Y})^2 / (N-1) = N^2 \left(1 - \frac{n}{N}\right) \left(\frac{1}{n}\right) S^2$$

missä  $\bar{Y} = \sum_{k=1}^N Y_k / N$  on perusjoukon keskiarvo

$S^2 = \sum_{k=1}^N (Y_k - \bar{Y})^2 / (N-1)$  on perusjoukon varianssi

$\left(1 - \frac{n}{N}\right)$  on äärellisyyskorjaus (fpc, *finite population correction*)

**Varianssiestimaattori** (harhaton)

$$\hat{v}_{SRS}(\hat{t}) = N^2 \left(1 - \frac{n}{N}\right) \left(\frac{1}{n}\right) \sum_{k=1}^n (y_k - \bar{y})^2 / (n-1) = N^2 \left(1 - \frac{n}{N}\right) \left(\frac{1}{n}\right) \hat{s}^2,$$

missä  $\bar{y} = \sum_{k=1}^n y_k / n$  on otoskeskiarvo

$\hat{s}^2 = \sum_{k=1}^n (y_k - \bar{y})^2 / (n-1)$  on otosvarianssi

HUOM: SRSWR-otannassa fpc =  $\left(1 - \frac{1}{N}\right)$

HUOM: Erikoistapauksena **Bernoulli-poiminta** (ks. Survey sampling reference manual, s. 15 ja Appendix 1, katsotaan lähemmin demoissa).

## Systemaattinen otanta SYS

Sisällymistodennäköisyys  $\pi_k = n/N$  on vakio

**Kokonaismäärän  $T$  estimaattori** (harhaton)

$$\hat{t} = N \sum_{k=1}^n y_k / n$$

**Asetelmavarianssi**

$$V_{\text{sys}}(\hat{t}) = \sum_{j=1}^q (\hat{t}_j - T)^2 / q = V_{\text{SRS}}(\hat{t})(1 + (n-1)\rho_{\text{int}}) = N \times \text{SSB},$$

missä  $\hat{t}_j$  on  $j$ :nnen systemaattisen otoksen kokonaismäärän estimaattori  
 $q=N/n$  on poimintaväli

$\rho_{\text{int}} = 1 - \frac{n}{n-1} \times \frac{\text{SSW}}{\text{SST}}$  on sisäkorrelaatiokerroin, missä käytetään

ANOVA-neliösummahajoitelmaa  $\text{SST} = \text{SSW} + \text{SSB}$ .

**Asetelmakerroin** (parametri)

$$\text{DEFF}_{\text{sys}}(\hat{t}) = \frac{V_{\text{sys}}(\hat{t})}{V_{\text{SRS}}(\hat{t})} = 1 + (n-1)\rho_{\text{int}}$$

Systemaattinen otanta on yksinkertaiseen satunnaisotantaan verrattuna:

- tehokkaampi, jos  $-1/(n-1) < \rho_{\text{int}} < 0$ ,
- yhtä tehokas, jos  $\rho_{\text{int}} = 0$ ,
- tehottomampi, jos  $0 < \rho_{\text{int}} < 1$

**Varianssiestimaattori**

Kuten SRS, jos oletetaan, että kyseessä on **satunnaisjärjestyksessä** oleva perusjoukko (jolloin sisäkorrelaation = 0)

Kuten STR (ositettu otanta, suhteellinen kiintiöinti), jos oletetaan **implisiittinen** ositus (perusjoukon alkioiden lajittelu ennen SYS-poimintaa)

**Ositettu otanta STR**

Ositteiden koot, ositteet  $1, \dots, h, \dots, H$ :

$$N_1 + N_2 + \dots + N_h + \dots + N_H = N,$$

missä  $N_h$  on ositteen  $h$  alkioiden lukumäärä

$H$  on ositteiden lukumäärä

$N$  on perusjoukon alkioiden lukumäärä

STR-otos poimitaan kustakin ositteesta itsenäisesti

Otoskoot:

$$n_1 + n_2 + \dots + n_h + \dots + n_H = n$$

Estimaattorit ovat ositekohtaisten estimaattoreiden painotettuja summia, painoina ositepainot  $W_h = N_h / N$ .

**Kokonaismäärän  $T$  estimaattori**  $\hat{t}_{str}$  on painotettu summa ositekeskiarvoista  $\bar{y}_h = \sum_{k=1}^{n_h} y_k / n_h$

$$\hat{t}_{str} = N \sum_{h=1}^H W_h \bar{y}_h = \sum_{h=1}^H \hat{t}_h = \hat{t}_1 + \dots + \hat{t}_h + \dots + \hat{t}_H,$$

missä  $\hat{t}_h = N_h \bar{y}_h$  on kokonaismäärän estimaattori ositteessa  $h$

**Asetelmavarianssi** (SRS ositteissa)

$$V_{str}(\hat{t}_{str}) = \sum_{h=1}^H V_{srs}(\hat{t}_h)$$

**Varianssiestimaattori** (harhaton)

$$\hat{v}_{str}(\hat{t}_{str}) = \sum_{h=1}^H \hat{v}_{srs}(\hat{t}_h)$$

**Kiintiöinti** (*allocation*)Suhteellinen kiintiöinti (*proportional allocation*)Tasakiintiöinti (*equal allocation*)Optimaalinen (*optimal allocation*) eli Neyman kiintiöintiBankier kiintiöinti (*Bankier or power allocation*)*Suhteellinen kiintiöinti:*Lisätieto: ositteen koko  $N_h$ Otoskoko  $n_h$  ositteessa  $h$ 

$$n_{h,pro} = n \times \frac{N_h}{N} = n \times W_h$$

Sisältymistodennäköisyys on vakio  $\pi_k = \pi = n / N$ Kokonaismäärän estimaattori  $\hat{t}_{str} = \hat{t} = N \sum_{h=1}^H \sum_{k=1}^{n_h} y_{hk} / n$ Menetelmää kutsutaan itsepainottuvaksi (*self-weighting*), koska ositekohtaisia keskiarvoja ei lasketa

HUOM: Muissa kiintiöintimenetelmissä sisältymistodennäköisyydet vaihtelevat ositteiden välillä (mutta ovat vakioita ositteiden sisällä)

*Tasakiintiöinti:*  $n_h = n / H$  kussakin ositteessa  $h$ . Jos ositteiden koot  $N_h$  vaihtelevat, niin sisältymistodennäköisyydet vaihtelevat:

$$\pi_{hk} = n_h / N_h = n / (H \times N_h) \text{ alkionle } k \text{ ositteessa } h$$

Asetelmapainot ovat  $w_{hk} = H \times N_h / n$ *Optimaalinen eli Neyman-kiintiöinti:*Ositteiden otoskoot määräytyvät yhtälöstä  $n_h = \frac{N_h S_h}{\sum_{h=1}^H N_h S_h}$ .missä  $S_h$  (lisätieto) on muuttujan  $y$  (tunnettu) keskihajonta ositteessa  $h$

**PPS-otanta** (*Probability Proportional to Size*)

Oletetaan, että perusjoukon alkion kokoa mittaavan muuttujan arvo  $Z_k$  on tunnettu jokaiselle perusjoukon alkiolle  $k$

Alkion  $k$  suhteellinen koko  $p_k = Z_k / T_z$ ,  $k=1, \dots, N$  missä  $T_z = \sum_{k=1}^N Z_k$

**Kriteerit estimoinnin tehostumiselle**

Kokoa mittaavan muuttujan  $z$  oma vaihtelu muistuttaa tutkittavan muuttujan  $y$  vaihtelua (voimakas korrelaatio)

Apumuuttujan  $z$  ja tutkittavan muuttujan  $y$  suhde on mahdollisimman lähellä vakiota

Jos suhde on lähes vakio kaikilla perusjoukon yksiköillä, niin estimaattorin asetelmavarianssi saa pienen arvon

**PPS-otoksen poiminta**, eri tapoja:

PPS_SYS	Systemaattinen PPS
PPS_WOR	Kumulatiivisen summan menetelmä (WOR)
PPS_WR	Kumulatiivisen summan menetelmä (WR)
PPS_RHC	Rao-Hartley-Cochran-poiminta
PPS_Poisson	Poisson-poiminta

Sisällymistodennäköisyydet  $\pi_k$  ovat suhteessa yksiköiden suhteellisiin kokoihin  $p_k = Z_k / T_z$ .

Esim PPS\_WR ja PPS\_SYS:

$$\pi_k = n \times p_k$$

HUOM: SRS\_WR-poiminnassa  $p_k = 1/N$  jokaiselle  $k$ . Lukua  $1/N$  kutsutaan alkion  $k$  yksittäisen poiminnan poimintatodennäköisyydeksi (*single-draw selection probability*) Sisällymistodennäköisyys  $n$  kokois-  
sen otoksen alkiolle  $k$  on siten  $\pi_k = n \times p_k = n/N$

**PPS\_WR: Kumulatiivisen summan PPS-poiminta**

Työvaiheet:

1) Laske kullekin alkioille  $k$  apumuuttujan  $z$  kumulatiivinen summa:

$$G_k = \sum_{j=1}^k Z_j, k=1, \dots, N, G_N = T_z.$$

2) Perusjoukon ensimmäiseen alkioon ( $a_1$ ) liitetään välin  $[1, G_1]$  kokonaisluvut

Toiseen alkioon ( $a_2$ ) liitetään välin  $[G_1 + 1, G_2]$  kokonaisluvut

Yleisesti alkioille  $k$  ( $a_k$ ) liitetään välin  $[G_{k-1} + 1, G_k]$  kokonaisluvut

3) Poimi satunnaisluku väliltä  $[1, G_N]$ . Se alkio tulee otokseen, jonka poimintaväliin satunnaisluku kuuluu

4) Toista vaihe 3) kunnes  $n$  alkion otos on poimittu.

Perusjoukon alkion  $k$  suhteellinen koko  $p_k$ :

$$p_k = \frac{Z_k}{\sum_{k=1}^N Z_k} = \frac{Z_k}{T_z}.$$

ja sisältymistodennäköisyys  $\pi_k$ :

$$\pi_k = n \times p_k = n \times \frac{Z_k}{T_z}$$



**PPS\_SYS: Systemaattinen PPS-poiminta**

Työvaiheet:

1) Laske poimintaväli  $q = T_z / n$

2) Generoi satunnaismuuttuja suljetulta väliltä  $[1, q]$ . Olkoot se  $q_0$ .

Poimintanumerot  $n$  alkion otosta varten ovat:

$$q_0, q_0 + q, q_0 + 2q, \dots, q_0 + (n-1)q$$

3) Kussakin poiminnassa otokseen otetaan ensimmäinen alkio kehikkolistalta, jossa kumulatiivinen koko  $G_k$  on suurempi tai yhtäsuuri kuin poimintanumero.

Sisältymistodennäköisyys on  $\pi_k = n \times p_k$

Alkiotason painokerroin  $w_k = 1 / \pi_k = 1 / (n \times p_k) = T_z / (Z_k \times n)$

HUOM: Sisältymistodennäköisyyden tulee täyttää ehto  $\pi_k \leq 1$ .

Jos  $Z_k$  on hyvin suuri, voi sisältymistodennäköisyys olla  $> 1$ .

Tällaiset alkiot otetaan otokseen ns. varmoina alkioina eli niille alkiuille sisältymistodennäköisyys  $\pi_k = 1$  joilla  $nZ_k > \sum_{k=1}^N Z_k$ . Varmat alkiot laitetaan kukin omaan ositteeseensa (ositettu PPS).

Jäljelle jäävien yksiköiden sisältymistodennäköisyys  $\pi_k$  määritellään uudelleen kokoa mittaavan muuttujan suhteessa.

Esim: Asetelma PPS\_SYS\_STR Keski-Suomen kunta-aineistossa.

## Kokonaismäärän estimaattorit

PPS\_WOR: **Horvitz-Thompson-estimaattori**

$$\hat{t}_{HT} = \sum_{k=1}^n \frac{y_k}{\pi_k} = \sum_{k=1}^n w_k y_k \quad \text{missä } \pi_k \text{ on alkion } k \text{ sisältymistodennäköisyys}$$

PPS\_WR: **Hansen-Hurwitz-estimaattori**

$$\hat{t}_{hh} = \frac{1}{n} \sum_{k=1}^n \frac{y_k}{p_k} = \frac{1}{n} (\hat{t}_1 + \dots + \hat{t}_k + \dots + \hat{t}_n),$$

missä kukin  $\hat{t}_k = y_k / p_k$  on kokonaismäärän  $T$  estimaatti

## Asetelmavarianssi

$$V_{ppswr}(\hat{t}_{hh}) = \frac{N^2}{n} \sum_{k=1}^N p_k \left( \frac{Y_k}{Np_k} - \bar{Y} \right)^2 = \frac{1}{n} \sum_{k=1}^N p_k (T_k - T)^2,$$

missä  $T_k = Y_k / p_k$  ja  $\bar{Y}$  on perusjoukon keskiarvo.

HUOM: Jos jokaiselle perusjoukon alkion  $k$  on voimassa  $Y_k / Z_k = C$  eli suhde on vakio, niin asetelmavarianssi = 0

## Varianssiestimaattori (harhaton)

$$\hat{v}_{ppswr}(\hat{t}_{hh}) = \frac{N^2}{n(n-1)} \sum_{k=1}^n \left( \frac{y_k}{Np_k} - \bar{y} \right)^2 = \frac{1}{n(n-1)} \sum_{k=1}^n (\hat{t}_k - \hat{t}_{hh})^2,$$

missä  $\bar{y}$  on otoskeskiarvo

HUOM: WR-varianssiestimaattoria käytetään approksimaationa PPS\_SYS- ja PPS\_WOR-otannassa