TEKNINEN YHTEENVETO I
Otanta-asetelmat ja estimointiasetelmat

Perusjoukko ja muuttujat

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

Tulosmuuttujan y tuntemattomat arvot $Y_1,\ldots,Y_k,\ldots,Y_N$

Apumuuttujan z tunnetut arvot $Z_1,\ldots,Z_k,\ldots,Z_N$

Perusjoukon parametrit

Äärellisen perusjoukon *U parametrit*

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

Keskiarvo $\overline{Y} = T/N$

S*uhteellinen 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 $s=\{1,...,k,...,n\}$, 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 θ 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 θ 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 analyyttinen lauseke tai approksimatiivinen varianssiestimaattori

Estimoitu **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) $deff(\hat{\theta}) = \dfrac{\hat{v}_{p(s)}(\hat{\theta})}{\hat{v}_{SRS}(\hat{\theta})}$

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

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

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

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

Varianssiestimaattori (harhaton)

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

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 $\text{fpc} = (1-\frac{1}{N})$

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

Systemaattinen otanta SYS

Sisältymistodennäköisyys $\pi_k = n/N$ on vakio

Kokonaismäärän T estimaattori (harhaton)

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

Asetelmavarianssi

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

missä $\hat{t}_j$ on j:nnen systemaattisen otoksen kokonaismäärän estimaattori
$q=N/n$ on poimintaväli
$\rho_{\text{int}} = 1 - \frac{n}{n-1} \times \frac{SSW}{SST}$ on sisäkorrelaatiokerroin, missä käytetään ANOVA-neliösummahajoitelmaa $SST = SSW + SSB$.

Asetelmakerroin (parametri)

$$DEFF_{sys}(\hat{t}) = \frac{V_{sys}(\hat{t})}{V_{srs}(\hat{t})} = 1 + (n-1)\rho_{\text{int}}$$

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,…,*h*,…,*H*:

$$N_1 + N_2 + ...N_h + ... + N_H = N,$$

missä N_h on ositteen *h* alkioiden lukumäärä
H on ositteiden lukumäärä
N on perusjoukon alkioiden lukumäärä

STR-otos poimitaan kustakin ositteesta itsenäisesti

Otoskoot:

$$n_1 + n_2 + ... + n_h + ... + n_H = n$$

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

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

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

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 kiintiöinti
Bankier 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 osite-kohtaisia 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:

$\pi_{hk} = n_h / N_h = n / (H \times N_h)$ alkiolle k 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,…,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

Sisältymistodennäköisyydet π_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*) Sisältymistodennäköisyys n kokoisen otoksen alkiolle k on siten $\pi_k = n \times p_k = n / N$

PPS_WR: Kumulatiivisen summan PPS-poiminta

Työvaiheet:

1) Laske kullekin alkiolle *k* apumuuttujan *z* kumulatiivinen summa:

$$G_k = \sum_{j=1}^{k} Z_k \text{ , } k=1,\ldots,N,\ G_N = T_z.$$

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 alkiolle *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 π_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:
$q_0, q_0 + q, q_0 + 2q, ..., q_0 + (n-1)q$

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 alkioille 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 π_k määritellään uudelleen kokoa mittaavan muuttujan suhteessa.

Esim: Asetelma PPS_SYS_STR Keski-Suomen kunta-aineistossa.

Kokonaismäärän estimaattorit

PPS_WOR: **Horvitz-Thompson-estimaattori**

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

PPS_WR: **Hansen-Hurwitz-estimaattori**

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

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

Asetelmavarianssi

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

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

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

Varianssiestimaattori (harhaton)

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

missä $\bar{y}$ on otoskeskiarvo

HUOM: WR-varianssiestimaattoria käytetään approksimaationa PPS_SYS-ja PPS_WOR-otannassa

Of the approximate variance estimates, the value of $\hat{v}_{1.sys}$, being based on an assumption of SRSWOR, is the largest. The others fall more or less below it. This could indicate that, in this case, systematic sampling is more efficient than simple random sampling. The most efficient approximation method turns out to be autocorrelative modelling, which gave the value deff $= 0.35$. This model is based on the assumption of an autocorrelated superpopulation, of which the fixed population constitutes one realization. The design effect turns out to be DEFF $= 0.55$, confirming the result.

The results on variance estimation can be evaluated by studying the properties of the intra-class correlation coefficient ρ_{int}, which is the single design parameter under systematic sampling, and the efficiency of this sampling scheme. Moreover, it is illustrated how the sorting order in the frame register is related to the value of the intra-class correlation coefficient.

Intra-class Correlation

Systematic sampling is our first example of a design where a design parameter exists. This parameter, called the *intra-class correlation coefficient* ρ_{int}, will be included in the design variance V_{sys} of an estimator. The magnitude of the intra-class correlation, and consequently its effect on variance estimates, depends partly on the selected sampling interval and partly on whether there is a successive system of ordering the study variable's values in the population frame. Under systematic sampling, the design variance of $\hat{t}$ was given in (2.13) as $V_{sys}(\hat{t}) = N^2 \sum_{j=1}^{q} (\overline{Y}_j - \overline{Y})^2/q$. The design variance can also be written as

$$V_{sys}(\hat{t}) = \sum_{j=1}^{q} (N\overline{Y}_j - N\overline{Y})^2/(N/n) = N \times \sum_{j=1}^{q} n \times (\overline{Y}_j - \overline{Y})^2. \tag{2.19}$$

Let us analyse the design variance (2.19) in more detail. First we decompose population variance into the variation between the systematic samples and the variation within the systematic samples, as in standard one-way analysis of variance. In ANOVA terms, we have

$$SST = SSW + SSB, \tag{2.20}$$

where *SST* represents the total sum of squares, *SSW* the within sum of squares and *SSB* the between sum of squares. The decomposition (2.20) can be written as

$$\sum_{k=1}^{N} (Y_k - \overline{Y})^2 = \sum_{j=1}^{q} \sum_{k=1}^{n} (Y_{jk} - \overline{Y}_j)^2 + \sum_{j=1}^{q} n(\overline{Y}_j - \overline{Y})^2. \tag{2.21}$$

Thus, an alternative form for design variance is $V_{sys}(\hat{t}) = N \times SSB$.

By using the decomposition of the total sum of squares (2.20), the intra-class correlation is defined as

$$\rho_{int} = 1 - \frac{n}{n-1} \times \frac{SSW}{SST}. \tag{2.22}$$

If the variance between the means is zero, or $SSB = 0$, then the intra-class correlation reaches its minimum $-1/(n-1)$ and, correspondingly, where $SSW = 0$ it reaches its maximum, or $\rho_{int} = 1$.

Further, we can write the variance of the total estimator in the form

$$V_{sys}(\hat{t}) = N^2 \left(1 - \frac{n}{N}\right) \frac{S^2}{n}[1 + (n-1)\rho_{int}], \tag{2.23}$$

or alternatively as the product of the SRSWOR design variance times a correction factor including the intra-class correlation coefficient as a correction factor

$$V_{sys}(\hat{t}) = V_{srs}(\hat{t}) \times [1 + (n-1)\rho_{int}].$$

Hence, the design effect is

$$\text{DEFF}_{sys}(\hat{t}) = \frac{V_{sys}(\hat{t})}{V_{srs}(\hat{t})} \doteq 1 + (n-1)\rho_{int}. \tag{2.24}$$

Systematic sampling compared with simple random sampling with replacement is

1. more efficient, if $-1/(n-1) < \rho_{int} < 0$,
2. equally efficient, if $\rho_{int} = 0$, or
3. less efficient, if $0 < \rho_{int} < 1$.

This can be interpreted to mean that the more heterogeneous the sampling intervals (i.e. negative intra-class correlation), the more efficient systematic sampling will be. Therefore, in systematic sampling there is a connection between the design parameter ρ_{int} and the sorting order of the frame population, a fact that can be successfully utilized in practice.

Example 2.4

Intra-class correlation (ρ_{int}) in the *Province'91* population. We will now calculate the intra-class correlation under systematic sampling from the *Province'91* population, where the total of UE91 is to be estimated. The intra-class correlation is calculated for systematic sampling involving a single systematic sample of eight (8) elements. The decomposition of the total sum of squares (2.21) is given in Table 2.7.

Hence, the intra-class correlation coefficient is

$$\rho_{int} = 1 - \frac{n}{n-1}\frac{SSW}{SST} = 1 - \frac{8}{8-1} \times \frac{162.14 \times 10^5}{171.32 \times 10^5} = -0.082.$$