Helsingin yliopisto
Sosiaalitieteiden laitos

Pienalue-estimointi (78189)

Kevät 2013

Risto Lehtonen

OSA 2

Horvitz-Thompson-estimaattorin ja GREG-estimaattorin varianssien estimointi

Esimerkkejä

Tämän osan kirjallisuus:

Lehtonen-Pahkinen (2004) *Practical Methods for Design and Analysis of Complex Surveys.* 2nd Edition. Chichester: Wiley. Chapter 6. (jaetaan luennolla)

Lehtonen R. and Veijanen A. (2009) Design-based methods of estimation for domains and small areas. Chapter 31 in Rao C.R. and Pfeffermann D. (Eds.). *Handbook of Statistics. Vol. 29B. Sample Surveys: Inference and Analysis.* New York: Elsevier. Pages 219-249. (jaetaan luennolla)

VARIANSSIEN ESTIMOINTI

(1) Suunniteltu domainrakenne
(*Planned domains*)

Oletetaan ositettu SRSWOR = STR-SRSWOR

Kustakin osajoukosta U_d poimitaan SRSWOR-otos
Osajoukkojen otoskoot n_d on kiinnitetty ositetun otanta-asetelman mukaisesti

Kiintiöintimenetelmiä:
- Optimaalinen (Neyman) -kiintiöinti
- Bankierin kiintiöinti
- Tasakiintiöinti
- Suhteellinen kiintiöinti

Otos s_d kokoa n_d alkiota poimitaan **ositteesta** U_d, jossa on N_d alkiota, $d = 1,...,D$

Asetelmapainot ovat $w_k = N_d / n_d$ kaikille $k \in U_d$

HUOM: Oletetaan siis että N_d **on tunnettu** kaikille d

HUOM: Yleinen tilanne (*unequal probability sampling*): Lehtonen and Veijanen (2009)

Domain-totaalit:

$$T_d = \sum_{k \in U_d} Y_k \ ,$$

$d = 1, \ldots, D$

HT-estimaattori (STR-SRSWOR):

$$\hat{t}_{dHT} = \sum_{k \in s_d} w_k y_k = \frac{N_d}{n_d} \sum_{k \in s_d} y_k = N_d \bar{y}_d \qquad (21)$$

HT-estimaattorin (21) varianssiestimaattori:

$$\hat{v}_{str-srs}(\hat{t}_{dHT}) = N_d^2 (1 - \frac{n_d}{N_d})(\frac{1}{n_d}) \sum_{k \in s_d} \frac{(y_k - \bar{y}_d)^2}{n_d - 1} \qquad (22)$$

HUOM: $\hat{s}_d^2 = \sum_{k \in s_d} \frac{(y_k - \bar{y}_d)^2}{n_d - 1}$ on **tulosmuuttujan** y

otosvarianssi domainissa d

Vaihtoehto estimaattorille (22):
Lehtonen and Veijanen (2009)
Kaava (4)

Domain-totaalin $T_d = \sum_{k \in U_d} Y_k$ **GREG-estimaattori (STR-SRSWOR):**

$$\begin{aligned} \hat{t}_{dGREG} &= \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} w_k (y_k - \hat{y}_k) \\ &= \sum_{k \in U_d} \hat{y}_k + \frac{N_d}{n_d} \sum_{k \in s_d} (y_k - \hat{y}_k) \end{aligned} \qquad (23)$$

GREG-estimaattorin (23) varianssiestimaattori:

$$\hat{v}_{srs}(\hat{t}_{dGREG}) = N_d^2 (1 - \frac{n_d}{N_d})(\frac{1}{n_d}) \sum_{k \in s_d} \frac{(\hat{e}_k - \bar{\hat{e}}_d)^2}{n_d - 1} \qquad (24)$$

missä

$\hat{e}_k = y_k - \hat{y}_k$, $k \in s_d$ ovat jäännöksiä (*residuals*)

$\bar{\hat{e}}_d = \sum_{k \in s_d} \hat{e}_k / n_d$ on jäännösten keskiarvo domainissa d, $d = 1, ..., D$

HUOM:

$$\hat{s}_{\hat{e}}^2 = \sum_{k \in s_d} \frac{(\hat{e}_k - \bar{\hat{e}}_d)^2}{n_d - 1}$$

on **jäännösten** $\hat{e}_k$ otosvarianssi domainissa d

Vertaa HT-estimaattorin varianssiestimaattoriin!

(2) Ei-suunniteltu domainrakenne
(Unplanned domains)

Poimitaan SRSWOR-otos *s*. Osajoukon U_d otoskoko n_{s_d} on satunnaismuuttuja, $E(n_{s_d}) = n_d = nN_d / N$

HUOM: Satunnaisuus tuottaa lisävariaatiota, joka on otettava huomioon varianssiestimaattoreissa

HUOM: Kaksi eri tilannetta:

N_d **on tunnettu: Oletus tässä luvussa**

N_d ei ole tunnettu, estimaattori $\hat{N}_d = \sum_{k \in s_d} w_k$

Olkoon SRSWOR, poimitaan otos kokoa *n* alkiota *N* alkion perusjoukosta

Otantasuhde (sisältymistn) on $\pi_k = n / N$

Asetelmapainot: $w_k = N / n$ kaikille $k \in U$

Määritellään:

Uudet muuttujat $y_{dk} = \delta_{dk} y_k$

Jäännökset $\hat{e}_{dk} = y_{dk} - \hat{y}_k$, $d = 1,...,D$

missä domain-indikaattorit ovat

$\delta_{dk} = 1$ kun $k \in U_d$, nolla muulloin

Muuttujat y_{dk} ovat **domain-kohtaisia tulos-muuttujia** (*extended domain variables of interest*), Lehtonen ja Veijanen (2009)

Domain-totaalin $T_d = \sum_{k \in U_d} Y_k$ **HT-estimaattori (SRSWOR):**

$$\hat{t}_{dHT} = \sum_{k \in s_d} w_k y_k = \frac{N}{n} \sum_{k \in s} y_{dk} = \frac{N}{n} n_d \bar{y}_d \qquad (25)$$

HT-estimaattorin (25) varianssiestimaattori:

$$\hat{v}_{srs}(\hat{t}_{dHT}) = N^2 (1 - \frac{n}{N})(\frac{1}{n}) p_d \sum_{k \in s_d} \frac{(y_k - \bar{y}_d)^2}{n_d - 1} (1 + \frac{q_d}{\text{c.v}_{dy}^2}) \qquad (26)$$

$d = 1, ..., D$, missä

$p_d = n_d / n$ ja $q_d = 1 - p_d$

$\text{c.v}_{dy} = \hat{s}_{dy} / \bar{y}_d$ tulosmuuttujan y otosvariaatiokerroin domainissa d

$\hat{s}_{dy}$ tulosmuuttujan y otoskeskihajonta domainissa d

Varianssiestimaattori (26) vastaa **Bernoulli-otannassa** yleisesti käytettyä varianssiestimaattoria

Vaihtoehto estimaattorille (26):
ks. Lehtonen and Veijanen (2009)
Kaavat (5) ja (6)

Domain-totaalin $T_d = \sum_{k \in U_d} Y_k$ **GREG-estimaattori (SRSWOR):**

$$\begin{aligned} \hat{t}_{dGREG} &= \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} w_k (y_k - \hat{y}_k) \\ &= \sum_{k \in U_d} \hat{y}_k + \frac{N}{n} \sum_{k \in s_d} (y_k - \hat{y}_k) \end{aligned} \qquad (27)$$

GREG-estimaattorin (27) varianssiestimaattori:

$$\hat{v}_{srs}(\hat{t}_{dGREG}) = N^2 (1 - \frac{n}{N})(\frac{1}{n}) \sum_{k \in s} \frac{(\hat{e}_{dk} - \bar{\hat{e}}_d)^2}{n-1} \qquad (28)$$

HUOM: Varianssiestimaattorissa (28) ovat mukana myös alkiot domainin d ulkopuolelta, koska $\hat{e}_{dk} = -\hat{y}_k$ alkioille $k \notin U_d$ ja $k \in s$

Vaihtoehtoinen varianssiestimaattori:

$$\hat{v}_{srs}(\hat{t}_{dGREG}) = N^2 (1 - \frac{n}{N})(\frac{1}{n}) p_d \sum_{k \in s_d} \frac{(\hat{e}_k - \bar{\hat{e}}_d)^2}{n_d - 1} (1 + \frac{q_d}{\text{c.v}^2_{d\hat{e}}}) \qquad (29)$$

$d = 1, ..., D$, missä

$p_d = n_d / n$ ja $q_d = 1 - p_d$

$\text{c.v}_{d\hat{e}} = \hat{s}_{d\hat{e}} / \bar{\hat{e}}_d$ jäännösten $\hat{e}_k$ otosvariaatiokerroin domainissa d

$\hat{s}_{d\hat{e}}$ jäännösten $\hat{e}_k$ otoskeskihajonta domainissa d

ESIMERKKI

Osajoukkojen totaalien estimointi asetelmaperusteisilla menetelmillä SRSWOR-otannan tilanteessa

Perusjoukko: Occupational Health Care Survey (OHC) -aineisto, $N = 7841$ henkilöä

Otanta-asetelma: SRSWOR-otanta, otoskoko $n = 1960$ henkilöä

Tavoite: Estimoidaan pitkäaikaissairaiden lukumäärä $D = 30$ osajoukossa

Binäärinen tulosmuuttuja:
CHRON (0: Ei ole, 1: On)

Alkiotason apumuuttuja *z*:
AGE (vuosina)

Muuttujien CHRON ja AGE **korrelaatio** domaineissa vaihtelee välillä 0.08 - 0.55
Koko aineistossa korrelaatio on 0.28

Ei-suunniteltu (*unplanned*) domainrakenne
Osajoukkojen otoskokoja n_{s_d} ei ole kiinnitetty otanta-asetelmassa vaan ne ovat satunnaismuuttujia

Mallin valinta: Malli (1b) on muotoa

$$y_k = \beta \times z_k + \varepsilon_k$$

Malli tuottaa yhteisen (*common*) suhdeparametrin

$$R = T / T_z = 7.778 \times 10^{-3}$$

kaikissa osajoukoissa (P-tyypin malli)

Mallin parametrin R estimaatti

$$\hat{r} = \hat{t}_{HT} / \hat{t}_{zHT} = 7.651 \times 10^{-3}$$

missä

$\hat{t}_{HT}$ on tulosmuuttujan y totaalin T HT-estimaatti

$\hat{t}_{zHT}$ on apumuuttujan z totaalin T_z HT-estimaatti

Mallilla saadaan y-muuttujalle sovitteet

$$\hat{y}_k = \hat{r} \times z_k, \qquad k = 1,...,7841$$

Domain-totaalien estimaattorit

HT-estimaattori:

$$\hat{t}_{dHT} = \sum_{k \in s_d} w_k y_k = N/n \sum_{k \in s_d} y_k$$

GREG-estimaattori:

$$\hat{t}_{dGREG-P} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} w_k (y_k - \hat{y}_k)$$
$$= \hat{t}_{dHT} + \hat{r}(T_{dz} - \hat{t}_{dzHT})$$

missä $w_k = N/n = 7841/1960 = 4.001$

T_{dz} on z-muuttujan **tunnettu** perusjoukon kokonaismäärä domainissa *d*

$\hat{t}_{dzHT} = \sum_{k \in s_d} w_k z_k$ vastaava HT-estimaattori

HUOM: Vastaava synteettinen estimaattori:

$$\hat{t}_{dSYN-P} = \sum_{k \in U_d} \hat{y}_k = T_{dz} \times \hat{r}$$

joka perustuu samaan yksinkertaiseen malliin kuin GREG-estimaattori

HUOM: P-malliin perustuvat GREG ja SYN ovat **epäsuoria** (*indirect*)

Tehokkustarkastelu: Estimaattorin $\hat{t}_d$

Estimoidut keskivirheet s.e$(\hat{t}_d)$

Prosenttiset variaatiokertoimet

c.v$(\hat{t}_d)\% = 100 \times$ s.e$(\hat{t}_d) / \hat{t}_d$

Varianssiestimaattorit (26) ja (29):

$$\hat{v}_{srs}(\hat{t}_{dHT}) = N^2(1-\frac{n}{N})(\frac{1}{n})p_d\hat{s}_{dy}^2(1+\frac{q_d}{\text{c.v}_{dy}^2})$$

$$\hat{v}_{srs}(\hat{t}_{dGREG-P}) = N^2(1-\frac{n}{N})(\frac{1}{n})p_d\hat{s}_{d\hat{e}}^2(1+\frac{q_d}{\text{c.v}_{d\hat{e}}^2}),$$

missä $p_d = n_d / n,\ q_d = 1 - p_d$

Varianssiestimaattorit ovat:

$\hat{s}_{dy}^2 = \sum_{k \in s_d}(y_k - \bar{y}_d)^2 / (n_d - 1)$

$\hat{s}_{d\hat{e}}^2 = \sum_{k \in s_d}(\hat{e}_k - \bar{\hat{e}}_d)^2 / (n_d - 1)$

Estimoidut variaatiokertoimet ovat:

c.v$_{dy} = \hat{s}_{dy} / \bar{y}_d$

c.v$_{d\hat{e}} = \hat{s}_{d\hat{e}} / \bar{\hat{e}}_d$, missä

$\bar{y}_d = \sum_{k \in s_d} y_k / n_d$ ja $\bar{\hat{e}}_d = \sum_{k \in s_d} \hat{e}_k / n_d$

Jäännökset ovat: $\hat{e}_k = y_k - \hat{r} \times z_k$

Taulukko 7. HT-estimaattorin ja GREG-estimaattorin keskimääräinen absoluuttinen suhteellinen erotus (*Mean absolute relative difference*, MARD) ja keskimääräinen variaatiokerroin (*Mean coefficient of variation*, MCV) osajoukon kokoluokan mukaan

	MARD (%)		***MCV (%)***	
Koko-luokka	**HT**	**GREG**	**HT**	**GREG**
-39	10.6	10.2	30.8	24.7
40-79	2.0	3.4	23.5	19.8
80-	3.2	3.7	16.0	13.6
Kaikki	1.8	1.7	23.0	19.0

$$\text{MARD} = |\bar{\hat{t}} - \bar{T}| / \bar{T}$$

laskettuna erikseen kussakin kokoluokassa, missä

$\bar{\hat{t}}$ kokoluokassa estimoitujen domain-totaalien $\hat{t}_d$ keskiarvo

$\bar{T}$ kokoluokan perusjoukon domain-totaalien T_d keskiarvo

Taulukko 8. Pitkäaikaissairaiden kokonaismäärien estimaatit osajoukoissa, SRSWOR-otos jossa otoskoko n = 1960), OHC-aineisto

	Otos-koko	**Pj:n koko**	**Para-metri**	**Totaali-estimaatti**		**Keskivirhe**		**Variaatio-kerroin**	
d	n_d	N_d	T_d	$\hat{t}_{dHT}$	$\hat{t}_{dGREG}$	s.e($\hat{t}_{dHT}$)	s.e($\hat{t}_{dGREG}$)	c.v($\hat{t}_{dHT}$)	c.v($\hat{t}_{dGREG}$)
Osajoukon otoskoko $n_d < 40$									
20	24	101	31	32.0	31.6	9.77	7.13	30.5	22.5
10	26	81	27	32.0	25.6	10.83	8.05	33.8	31.5
18	26	129	36	20.0	27.2	7.60	6.95	38.0	25.5
23	31	156	57	44.0	53.2	10.82	9.10	24.6	17.1
8	35	141	29	24.0	24.5	8.57	7.88	35.7	32.2
30	36	146	34	32.0	33.8	9.86	8.56	30.8	25.3
3	37	133	29	36.0	32.6	10.77	8.73	29.9	26.8
16	37	165	45	52.0	54.8	12.14	9.15	23.3	16.7
Osajoukon otoskoko $40 \le n_d < 80$									
1	41	181	33	40.0	43.0	10.80	9.15	27.0	21.3
21	43	153	48	64.0	55.3	14.55	10.93	22.7	19.8
6	45	188	52	24.0	26.6	8.51	7.67	35.5	28.9
28	51	194	74	88.0	85.4	16.61	11.65	18.9	13.6
24	53	200	55	56.0	55.7	13.21	11.06	23.6	19.9
22	57	242	96	112.0	115.0	17.79	13.08	15.9	11.4
15	58	252	61	60.0	66.4	13.20	11.90	22.0	17.9
11	59	187	47	52.0	39.5	13.30	10.89	25.6	27.6
13	69	305	89	80.0	88.5	15.10	12.86	18.9	14.5
12	73	311	95	56.0	65.9	12.85	11.40	22.9	17.3
4	76	295	65	68.0	68.1	14.39	12.17	21.2	17.9
7	78	292	52	40.0	36.3	11.09	10.17	27.7	28.0
Osajoukon otoskoko $n_d \ge 80$									
2	84	352	86	76.0	78.6	14.95	13.49	19.7	17.2
5	86	323	66	76.0	70.5	15.31	13.62	20.1	19.3
26	89	364	124	124.0	126.0	19.07	15.72	15.4	12.5
29	90	365	128	124.0	124.5	19.12	15.10	15.4	12.1
25	91	339	114	112.0	101.6	18.68	14.81	16.7	14.6
17	99	426	139	176.0	183.3	22.11	16.72	12.6	9.1
9	103	366	89	88.0	79.3	16.66	13.82	18.9	17.4
19	115	490	165	152.0	160.0	20.81	17.13	13.7	10.7
14	116	447	130	136.0	128.4	20.31	16.28	14.9	12.7
27	132	517	197	176.0	173.8	22.94	17.51	13.0	10.1
All	1960	7841	2293	2252.3	2254.8	69.42	66.88	3.1	3.0

Asetelmaperusteisen malliavusteisen epäsuoran GREG-estimaattorin ja malliperusteisen epäsuoran SYN-estimaattorin vertailu

SYN-estimaattori:

$$\hat{t}_{dSYN-P} = \sum_{k \in U_d} \hat{y}_k = T_{dz} \times \hat{r}$$

Domain 1:

$$\hat{t}_{1GREG-P} = \sum_{k \in U_1} \hat{y}_k + \sum_{k \in s_1} w_k (y_k - \hat{y}_k)$$

$$= 45.43 + 4.001 \times (-0.5974) = 43.04$$

$$\hat{t}_{1SYN-P} = \sum_{k \in U_1} \hat{y}_k = T_{1z} \times \hat{r} = 5937 \times 0.0076515 = 45.43$$

Perusjoukon parametri: $T_1 = 33$

Domain 19:

GREG: $\hat{t}_{19GREG-P} = 160.00$

SYN: $\hat{t}_{19SYN-P} = 138.09$

Perusjoukon parametri: $T_{19} = 165$