Heisingin yliopisto
Matematiikan ja tilastotieteen laitos

Otantamenetelmät
Syksy 2008

Uudelleenpainotus ja imputointi
Perusteita

Prof. Risto Lehtonen, Helsingin yliopisto

2.12.2008

Uudelleenpainotus – Otostasoisen tiedon käyttö[1]

Tyypilliset otannasta riippumattomat virheet (nonsampling errors)

Vastauskato (nonresponse)
Peitto- ja kehikkovirheet (coverage and frame errors)
Mittausvirheet (measurement errors)
Processing errors

Tavoite: Vastauskadon vaikutusten arviointi ja adjustointi

Vastauskato viittaa kahteen tilanteeseen:

Yksikkökato (Unit nonresponse)

- Mitään tietoja ei ole saatu kerättyä joiltakin otosyksiköiltä
- Kaikki tutkimusmuuttujat saavat puuttuvan tiedon arvon näille yksiköille

Eräkato (Item nonresponse)

- Joitakin tietoja on jäänyt keräämättä joiltakin otosyksiköiltä

[1] Source: Lehtonen R. and Pahkinen E. (2003) Practical Methods for Design and Analysis of Complex Surveys. Second Edition. Chichester: John Wiley & Sons, Ltd (Chapter 4).

- Jotkin tutkimusmuuttujat saavat puuttuvan tiedon arvon näille yksiköille

HUOM: Molemmat puuttuvan tiedon tyypit voivat aiheuttaa harhaa estimointiin

ESIMERKKI

Yksikkökato tyypillisissä survey-tutkimuksissa

Table 4.1 Vastausprosentti eräissä otostutkimuksissa

Name of the survey	Sampling unit	Sample size	Response rate (%)
(1) Mini-Finland Health Survey	Person	8000	96 %
(2) Occupational Health Care Survey	Establishment	1542	88 %
(3) Health Security Survey	Household	6998	84 %
(4) PISA 2000 Survey	School	6638	85 %
(5) Passenger Transport Survey	Person	18250	65 %
(6) Wages Survey	Business firm	1572	80 %

PISA 2000: Median of country-level response rate is presented due to heavy country-level variation

YKSIKKÖKATO (UNIT NONRESPONSE

Estimoitava parametri

Totaali $T = \sum_{k=1}^{N} Y_k$

HT estimaattori $\hat{t}_{ht} = \sum_{k=1}^{n} y_k / \pi_k$

Otanta-asetelma: SRSWOR
Otoskoko: *n* alkiota

HT-estimaattorin varianssi (SRSWOR)

$$V_{srs}(\hat{t}_{ht}) = N^2(1 - n/N)S^2/n$$

Jakajana alkuperäinen otoskoko *n*
Vastauskadon vallitessa saadun aineiston koko pienenee

Saadun aineiston koko: $n_{(r)} < n$

Siis varianssi kasvaa!

YKSIKKÖKADON AIHEUTTAMA HARHA

Alkion *k* vastaustodennäköisyys

$$\theta_k, k = 1,...,N$$

Harmillinen (non-ignorable) vastauskato
Little and Rubin (1987):

> Vastaustodennäköisyys θ_k riippuu tulosmuuttujan y arvosta Y_k

Harmiton (ignorable) vastauskato

> Vastaustodennäköisyys θ_k ei riipu tulosmuuttujan y arvosta Y_k

Esimerkiksi: ”Ignorable” tilanne

Vastaustodennäköisyys θ_k on vakio kaikille alkioille $k = 1,\ldots,N$

ESIMERKKI

Harmillinen (non-ignorable) vastauskato

Oletetaan, että haastattelututkimuksessa yksi osajoukko jättäytyy kokonaisuudessaan tutkimuksen ulkopuolelle

Perusjoukko voidaan tällöin jakaa kahteen osaperusjoukkoon

A. Osallistuva osajoukko, N_1 alkiota
B. Ei-osallistuva osajoukko (kato) N_2 alkiota

Totaalin T estimaattori $\hat{t}_{ht(r)} = N \times \bar{y}_{(r)}$
missä $\bar{y}_{(r)}$ on osajoukosta A saadun aineiston keskiarvo

Tällöin $E(\bar{y}_{(r)}) = \bar{Y}_1$ (osajoukon A keskiarvo)

Jos $\overline{Y}_1 \neq \overline{Y}_2$ niin estimaattori $\hat{t}_{ht(r)}$ on harhainen

BIAS $(\hat{t}_{ht(r)})$ =

$$E(\hat{t}_{ht(r)}) - T = N\overline{Y}_1 - (N_1\overline{Y}_1 + N_2\overline{Y}_2) = N_2(\overline{Y}_1 - \overline{Y}_2)$$

Käytännössä harhan suuruutta on vaikea arvioida

Varianssin sijasta variaation mittana tulisi käyttää keskineliövirhettä

$$\mathrm{MSE}(\hat{t}_{ht(r)}) = V_{p(s)}(\hat{t}_{ht(r)}) + \mathrm{BIAS}^2(\hat{t}_{ht(r)}),$$

Jos harhaa ei tiedetä, niin MSE ei voida laskea

ESIMERKKI

Vastauskato ja harha datassa Province'91

Oletetaan, että seuraavat 5 kuntaa kuuluvat kato-osajoukkoon B: Kuhmoinen, Joutsa, Luhanka, Leivonmäki, Toivakka

Osajoukko A: N_1=27
Osajoukko B (kato): N_2=5

$T_1 = 14\,475 \quad N_1 = 27 \quad \overline{Y}_1 = 536.11$
$T_2 = 623 \quad N_2 = 5 \quad \overline{Y}_2 = 124.60$
$T = 15\,098 \quad N = 32 \quad \overline{Y} = 471.81$

SRSWOR-otos (n = 8 kuntaa)

Estimaattori $\hat{t}_{ht(r)}$, odotusarvo:

$$E(\hat{t}_{ht(r)}) = N \times \overline{Y}_1 = 32 \times 536.11 = 17\,156.$$

BIAS($\hat{t}_{ht(r)}$)

$$= E(\hat{t}_{ht(r)}) - T = N_2(\overline{Y}_1 - \overline{Y}_2)$$
$$= 5 \times (536.11 - 124.60) = 2058$$

eli varsin suuri

UUDELLEENPAINOTUS Reweighting

Yksikkökadon (Unit non-response) hallinta

Lisäinformaation käyttö

Koko otoksesta saatava lisäinfo

Perusjoukon tasoinen lisäinfo

Yksinkertainen esimerkki

Oletus: Kaikkien perusjoukon alkioiden osallistumistodennäköisyys on vakio, eli

$\theta_k = \theta$ kaikille $k \in U$

Aineistosta estimoitu $\hat{\theta} = n_{(r)} / n$

Uudelleenpainotettu HT-estimaattori

$$\hat{t}^*_{ht} = \sum_{k=1}^{n(r)} w_k y_k = \sum_{k=1}^{n(r)} y_k / (\hat{\theta} \times \pi_k)$$

tai

$$\hat{t}^*_{ht} = (1/\hat{\theta}) \times \sum_{k=1}^{n(r)} y_k / \pi_k = (1/\hat{\theta}) \times \hat{t}_{ht}$$

missä $\hat{t}_{ht} = \sum_{k=1}^{n(r)} y_k / \pi_k$

Vakio-osallistumistodennäköisyyden oletus on käytännössä epärealistinen

1) Diskreetti lisäinfo: RHG-menetelmä

Response Homogeneity Groups

Jaetaan perusjoukko tai koko otos vastaustodennäköisyyden suhteen sisäisesti homogeenisiin osajoukkoihin käyttäen hyväksi perusjoukosta tai koko otoksesta käytettävissä olevaa lisäinformaatiota, joka korreloi osallistumisalttiuden kanssa

Otostasoinen lisäinfo:

Osajoukot: $1,\ldots,c,\ldots,C$
Osajoukkojen otoskoot: $n_1,\ldots,n_c,\ldots,n_C$
Saadun aineiston koot: $n_{1(r)},\ldots,n_{c(r)},\ldots,n_{C(r)}$

Oletus: Vastaustodennäköisyys θ_c on vakio kunkin osajoukon sisällä, mutta voi vaihdella osajoukkojen välillä

Estimoitu osajoukon *c* osallistumistn

$$\hat{\theta}_c = n_{c(r)} / n_c,\ c = 1,\ldots,C$$

Uudelleenpainotettu HT-estimaattori

$$\hat{t}^{*}_{rhg} = \sum_{k=1}^{n(r)} w^{*}_{rhg,k} y_k = \sum_{c=1}^{C} \sum_{k=1}^{n_{c(r)}} (1/\hat{\theta}_c) \times w_{ck} y_{ck}$$

missä uusi paino on $w^{*}_{rhg,k} = (1/\hat{\theta}_c) \times w_{ck}$

ja $w_{ck} = 1/\pi_{ck}$ on asetelmapaino, $c = 1,...,C$ ja $k = 1,...,n_{c(r)}$

RHG-menetelmä on tehokas jos osajoukkojen konstruointi onnistuu niin, että sisäinen homogeenisuusehto täyttyy

Edellyttää lisäinformaation hyvää saatavuutta ja (voimakasta) korrelaatiota osallistumisalttiuden kanssa

RHG-menetelmä käyttää **diskreettiä lisäinformaatiota** (yhden tai useamman otostasoisen diskreetin muuttujan käyttö osajoukkojen muodostamisessa)

2) Jatkuvatyyppinen lisäinformaatio

Jatkuva lisäinfomuuttuja *z* tunnetaan kaikilta otosalkioilta $k = 1,\ldots,n$

Muuttuja korreloi voimakkaasti osallistumisalttiuden θ_k kanssa

Uudet painot (reweights)

$$w^*_{rat,k} = [(1/\hat{\theta}) \times (\bar{z} / \bar{z}_{(r)})] \times w_k$$

missä $\bar{z}$ on muuttujan *z* keskiarvo, joka on laskettu koko otoksesta

$\bar{z}_{(r)}$ on keskiarvo, joka on laskettu saadusta aineistosta, $\hat{\theta} = n_{(r)} / n$ ja $w_k = 1/\pi_k$

Uudelleenpainotettu HT-estimaattori

$$\hat{t}^*_{rat} = \sum_{k=1}^{n_{(r)}} w^*_{rat,k} y_k = \frac{\bar{z}}{\hat{\theta} \times \bar{z}_{(r)}} \sum_{k=1}^{n_{(r)}} w_k y_k$$

Suhdetehosteinen estimointi/ Ratio estimation

UUDELLEENPAINOTETUN HT-ESTIMAATTORIN VARIANSSIN ESTIMOINTI

Uudelleenpainotuksessa painot ovat muotoa

$$w_k^* = 1/(\pi_k \hat{\theta}_k)$$

missä sisältymistodennäköisyydet π_k ovat tunnettuja parametreja (ei satunnaismuuttujia)

Estimoidut vastaustodennäköisyydet $\hat{\theta}_k$ ovat satunnaismuuttujia

Uudelleenpainotetun HT-estimaattorin asetelmavarianssi on siten muotoa

$$V(\hat{t}_{ht}^*) = V_{sam}(\hat{t}_{ht}^*) + V_{rew}(\hat{t}_{ht}^*)$$

missä

V_{sam} Asetelmavarianssi (otantavirheen hallinta)

V_{rew} Lisävarianssi (uudelleenpainotuksen aiheuttama lisäepävarmuus)

ESIMERKKI (Example 4.2)
Province'91 Population

N = 32 kuntaa

SRSWOR otos, *n* = 8 kuntaa, $\pi_k = \pi = 0.25$

Kaksi katokuntaa: Kuhmoinen ja Toivakka

Saadun datan koko $n_{(r)} = 6$ kuntaa

Lisäinformaatiomuuttuja *z* (jatkuva)

HOU85 Asuntokuntien lkm 1985

Lisäinfo tiedossa kaikista otoskunnista

Estimoitu vastaustodennäköisyys

$$\hat{\theta}_k = \hat{\theta} = n_{(r)} / n = 6/8 = 0.75$$

RHG:

Kaupungit $c = 1$ $\hat{\theta}_1 = 3/3 = 1.00$

Muut kunnat $c = 2$ $\hat{\theta}_2 = 3/5 = .60$

Lisäinfo:

Koko otos ($n = 8$): $\bar{z} = 5154.75$
Saatu data ($n = 6$): $\bar{z}_{(r)} = 6490.17$

(1) Estimaattori $\hat{t}^*_{ht}$
RHG: Koko otos
Naiivi uudelleenpainotus:

$$w^*_{ht} = 1/(\pi_k \hat{\theta}_k) = 1/(0.25 \times 0.75) = 5.3333$$

(2) Estimaattori $\hat{t}^*_{rhg}$
RHG: Kaupungit / Muut kunnat
Uudelleenpainotus:
Kaupungit $w^*_{rhg,1} = (1/1) \times 4 = 4$
Muut kunnat $w^*_{rhg,2} = (1/0.60) \times 4 = 6.6667$

(3) Estimaattori $\hat{t}^*_{rat}$
RHG: Koko otos
Uudelleenpainotus:

$$w^*_{rat,k} = w_k \times [(1/\hat{\theta}) \times (\frac{\bar{z}}{\bar{z}_{(r)}})]$$

$$= 4 \times (1/0.75) \frac{5154.75}{6490.17} = 4.2359$$

Table 4.2 SRSWOR otos perusjoukosta *Province'91*.

Sample design identifiers			Element	Response data (Sample)			Reweight by nonresponse model		
STR	CLU	WGHT	LABEL	UE91	HOU85	RHG	REW_HT w*ht	RHG w*rhg	RATIO w*rat
1	18	4	Kuhmoinen	. .	1 463	2	0.0000	0.0000	0.0000
1	30	4	Toivakka	. .	834	2	0.0000	0.0000	0.0000
1	26	4	Pihtipudas	331	1 946	2	5.3333	6.6667	4.2359
1	31	4	Uurainen	219	932	2	5.3333	6.6667	4.2359
1	15	4	Konginkangas	142	556	2	5.3333	6.6667	4.2359
1	1	4	Jyväskylä	4 123	26 881	1	5.3333	4.0000	4.2359
1	4	4	Keuruu	760	4 896	1	5.3333	4.0000	4.2359
1	5	4	Saarijärvi	721	3 730	1	5.3333	4.0000	4.2359

A missing value is denoted as ". . "

Uudelleenpainotusestimaattorin varianssi

Totaaliestimaattorin asetelmavarianssi:

$$V_{sam}(\hat{t}) = N^2(1-\frac{n}{N}) \times S^2_{(r)} / n_{(r)}$$

missä $$S^2_{(r)} = \sum_{k=1}^{N(r)} \frac{(Y_k - \bar{Y}_{(r)})^2}{N_{(r)} - 1}$$

Asetelmavarianssin estimaatti:

$$\hat{v}_{sam}(\hat{t}) = N^2(1-\frac{n}{N}) \times \hat{s}^2_{(r)} / n_{(r)}$$

$$= 32^2(1-\frac{8}{32}) \times 1527.59^2 / 6 = 14\,967^2$$

missä $$\hat{s}^2_{(r)} = \sum_{k=1}^{n(r)} \frac{(y_k - \bar{y}_{(r)})^2}{n_{(r)} - 1}$$

$V_{sam}(\hat{t})$ on sama kaikille estimaattoreille (1)-(3)

(1) Estimaattori $\hat{t}^*_{ht}$

Uudelleenpainotuksesta johtuva varianssikomponentti:

$$V_{rew}(\hat{t}^*_{ht}) = N^2(1-\frac{n_{(r)}}{n})\times S^2_{(r)}/n_{(r)}$$

missä $$S^2_{(r)} = \sum_{k=1}^{N_{(r)}}\frac{(Y_k - \overline{Y}_{(r)})^2}{N_{(r)}-1}$$

Varianssikomponentin estimaatti:

$$\hat{v}_{rew}(\hat{t}^*_{ht}) = N^2(1-\frac{n_{(r)}}{n})\times \hat{s}^2_{(r)}/n_{(r)}$$

$$= 32^2(1-\frac{6}{8})\times 1527.59^2/6 = 9978.18^2$$

(2) Estimaattori $\hat{t}^*_{rhg}$

RHG:

Kaupungit Otoskoko $n_1 = 3$

$$\hat{N}_1 = (n_1 / n) \times N = (3/8) \times 32 = 12$$

Muut kunnat Otoskoko $n_2 = 5$

$$\hat{N}_2 = (n_2 / n) \times N = (5/8) \times 32 = 20$$

Uudelleenpainotuksesta johtuva varianssikomponentti:

$$V_{rew}(\hat{t}^*_{rhg})$$

$$= \hat{N}_1^2 (1 - \frac{n_{1(r)}}{n_1}) \times S^2_{1(r)} / n_{1(r)}$$

$$+ \hat{N}_2^2 (1 - \frac{n_{2(r)}}{n_2}) \times S^2_{2(r)} / n_{2(r)}$$

missä $S^2_{h(r)} = \sum_{k=1}^{N_{h(r)}} \frac{(Y_{hk} - \overline{Y}_{h(r)})^2}{N_{h(r)} - 1}$

Varianssikomponentin estimaatti:

$$\hat{v}_{rew}(\hat{t}^{*}_{rhg}) = 12^2(1-\frac{3}{3})\times 1952.99^2/3$$

$$+ 20^2(1-\frac{3}{5})\times 95.04^2/3$$

$$= 0 + 694.07^2$$

$$= 694.07^2$$

(3) Estimaattori $\hat{t}^*_{rat}$

Määritellään jäännökset

$$E_{k(r)} = Y_{k(r)} - \frac{\overline{Y}_{(r)}}{\overline{Z}_{(r)}} \times Z_{k(r)}$$

Uudelleenpainotuksesta johtuva varianssikomponentti:

$$V_{rew}(\hat{t}^*_{rat}) = N^2(1 - \frac{n_{(r)}}{n}) \times S^2_{E_{(r)}} / n_{(r)}$$

missä $S^2_{E_{(r)}} = \sum_{k=1}^{N_{(r)}} (E_{k(r)} - \overline{E})^2 / (N_{(r)} - 1)$ ja

$$\overline{E} = \sum_{k=1}^{N_{(r)}} E_{k(r)} / N_{(r)}.$$

Estimoidut jäännökset

$$\hat{e}_{k(r)} = y_{k(r)} - \frac{\overline{y}_{(r)}}{\overline{z}_{(r)}} \times z_{k(r)}$$

Varianssikomponentin estimaatti:

$$\hat{v}_{rew}(\hat{t}^{*}_{rat}) = N^2(1-\frac{n_{(r)}}{n})\hat{s}^2_{\hat{e}_{(r)}}/n_{(r)}$$

$$= 32^2(1-\frac{6}{8})\times 120.29^2/6 = 785.73^2$$

missä $\hat{s}^2_{\hat{e}_{(r)}} = \sum_{k=1}^{n_{(r)}}(\hat{e}_{k(r)} - \bar{\hat{e}}_{(r)})^2/(n_{(r)}-1)$

Poimintasuhteet:

Estimaattorit $\hat{t}^*_{ht}$ ja $\hat{t}^*_{rat}$

$$n_{(r)} / N = 6 / 32 = 0.1875$$

Estimaattori $\hat{t}^*_{rhg}$

Kaupungit $n_{1(r)} / \hat{N}_1 = 3/12 = 0.25$

Muut kunnat: $n_{2(r)} / \hat{N}_2 = 3/20 = 0.15$

Vertailuestimaattorit:

(0) Estimaattori $\hat{t}_{ht(r)} = N \times \bar{y}_{(r)}$

Poimintasuhde $n_{(r)} / N = 6 / 32 = 0.1875$

(4) Estimaattori $\hat{t}_{ht}$ "Full response"

Poimintasuhde $n/N = 8/32 = 0.25$

Table 4.3 Varianssikomponentit ja kokonaisvarianssi eri estimaattoreille (*Province'91* population).

Model and estimator	**Estimate for a Total**	$\hat{v}$	$\hat{v}_{sam}$	$\hat{v}_{rew}$
(0) Respondent data $(n_{(r)} = 6)$ $\hat{t}_{ht(r)}$	33 579	$17\,988^2$	$17\,988^2$	0
(1) Reweighted estimator $\hat{t}^*_{ht}$	33 579	$17\,988^2$	$14\,967^2$	9978^2
(2) Response homogeneity group $\hat{t}^*_{rhg}$	27 029	$14\,983^2$	$14\,967^2$	694^2
(3) Ratio estimator $\hat{t}^*_{rat}$	26 669	$14\,988^2$	$14\,967^2$	786^2
(4) "Full response" $(n = 8)$ $\hat{t}_{ht}$	26 440	$13\,282^2$	$13\,282^2$	0

IMPUTOINTI Imputation

Eräkadon (item non-response) hallinta

Tavoite: Täydellinen datamatriisi

Tulosmuuttuja y

Puuttuva mittaustulos y_k alkiolle k

Imputoitu arvo $\hat{y}_k$

IMPUTOINTIMENETELMIÄ

(1) Keskiarvoimputointi
Respondent mean method RM

Jatkuva tulosmuuttuja y

Imputoitu arvo $\hat{y}_k = \bar{y}_{(r)}$

eli vastanneiden keskiarvo

Keskiarvoimputointi ei ole yleisesti suositeltava menetelmä

Kehittyneemmät menetelmät:

Lisäinformaation käyttö otosaineistosta tai perusjoukosta

(2) Lähimmän naapurin menetelmä Nearest neighbor method NN

Jatkuva tulosmuuttuja y
Puuttuva tieto y_k alkiolle k

Jatkuva lisäinformaatiomuuttuja z
Tiedossa kaikilta otosalkioilta

Lasketaan pareittaiset etäisyydet

$$\left|z_l - z_k\right|,\ l \neq k$$

Valitaan substituutti $\hat{y}_k = y_l$ jolle etäisyys on pienin, missä y_l on havaittu arvo

Alkio l on luovuttaja (donor)

(3) Suhde-estimointi
Ratio estimation method RA

Jatkuva tulosmuuttuja *y*

Puuttuva tieto y_k alkiolle *k*

Jatkuva lisäinformaatiomuuttuja *z*
Tiedossa kaikilta otosalkioilta

Imputoitu arvo

$$\hat{y}_k = z_k \times (\bar{y}_{(r)} / \bar{z}_{(r)})$$

missä $\bar{y}_{(r)}$ on tulosmuuttujan *y* keskiarvo havaitussa aineistossa

$\bar{z}_{(r)}$ on apumuuttujan *z* keskiarvo havaitussa aineistossa

(4) Hot deck –menetelmä HD

Tulosmuuttuja y (jatkuva tai diskreetti)
Puuttuva tieto y_k alkiolle k

Donor l ja vastaava imputoitu arvo $\hat{y}_k = y_l$
valitaan satunnaisesti **havaittujen arvojen** joukosta

(5) Moni-imputointi - Multiple imputation MI

Single imputation: Menetelmät (1)-(4)

> Alkion k puuttuva tieto y_k korvataan **yhdellä** imputoidulla arvolla $\hat{y}_k$

Multiple imputation:

> Alkion k puuttuva tieto y_k korvataan **usealla** imputoidulla arvolla $\hat{y}_{k1}, \hat{y}_{k2}, \ldots, \hat{y}_{km}$
>
> Saadaan m täydellistä havaintomatriisia

Usein valitaan arvo $m = 5$

TOTAALIESTIMAATTORIN VARIANSSIN ESTIMOINTI IMPUTOINNIN YHTEYDESSÄ

Imputointi tuottaa estimaattorin varianssi-lausekkeeseen lisäkomponentin (vastaavasti kuin uudelleenpainotusmenetelmien yhteydessä)

HT-estimaattorin
$\hat{t}_{ht}^{*} = \sum_{k=1}^{n} y_k / \pi_k$ varianssilauseke

$$V(\hat{t}_{ht}^{*}) = V_{sam}(\hat{t}_{ht}^{*}) + V_{imp}(\hat{t}_{ht}^{*})$$

missä

$V_{sam}(\hat{t}_{ht}^{*})$ on asetelmavarianssi

$V_{imp}(\hat{t}_{ht}^{*})$ on imputoinnin aiheuttama lisävarianssi (imputointivarianssi)

Lisävarianssin $V_{imp}(\hat{t}_{ht}^{*})$ lauseke riippuu imputointimenetelmästä

Moni-imputointi Multiple imputation (MI)

Varianssiestimaattori

$$\hat{v}(\hat{t}_{mi}) = \hat{v}_{sam}(\hat{t}_{mi}) + \hat{v}_{imp}(\hat{t}_{mi})$$

Alkiolle *k* imputoidaan *m* arvoa

$$\hat{y}_1,...,\hat{y}_j,...,\hat{y}_m$$

jolloin saadaan *m* täydellistä datamatriisia

Määritellään jokaiselle *m* matriisille totaaliestimaattori

$$\hat{t}_j^* = \sum_{k=1}^{n} w_k y_k \,,\ j = 1,...,m$$

missä $w_k = 1/\pi_k$

HUOM: Osa arvoista y_k on imputoituja!

Lasketaan totaaliestimaattien keskiarvo

$$\hat{t}_{mi}^* = \frac{1}{m} \times \sum_{j=1}^{m} \hat{t}_j^*$$

Määritellään **varianssikomponentit:**

Imputointien **sisäinen** varianssiestimaattori

$$\hat{v}_{sam}(\hat{t}_{mi}) = \left[\frac{1}{m} \times \sum_{j=1}^{m} \hat{v}_{p(s)}(\hat{t}_j^*) \right]$$

Imputointien **välinen** varianssiestimaattori

$$\hat{v}_{imp}(\hat{t}_{mi}) = \left[(1 + \frac{1}{m}) \times \sum_{j=1}^{m} \frac{(\hat{t}_j^* - \bar{\hat{t}}_{mi}^*)^2}{m-1} \right]$$

jolloin **kokonaisvarianssin** estimaattori on:

$$\hat{v}(\hat{t}_{mi}) = \hat{v}_{sam}(\hat{t}_{mi}) + \hat{v}_{imp}(\hat{t}_{mi})$$

$$= \left[\frac{1}{m} \times \sum_{j=1}^{m} \hat{v}_{p(s)}(\hat{t}_j^*) \right] +$$

$$\left[(1 + \frac{1}{m}) \times \sum_{j=1}^{m} \frac{(\hat{t}_j^* - \bar{\hat{t}}_{mi}^*)^2}{m-1} \right]$$

ESIMERKKI (Example 4.3)

Province'91 Population

N = 32 kuntaa

SRSWOR otos, n = 8 kuntaa, $\pi_k = \pi = 0.25$

Tulosmuuttuja y = UE91 (työttömien lukumäärä kunnassa)

Lisätietomuuttuja z = HOU85 (asuntokuntien lkm vuonna 1985, väestölaskenta)
Tiedossa kaikista kunnista

Puuttuva tieto muuttujalta UE91 kunnista:
Kuhmoinen ja Toivakka

Imputointimenetelmät:

(1) Keskiarvoimputointi RM
(2) Lähimmän naapurin menetelmä NN
(3) Suhde-estimointi RA
(4) Moni-imputointi MI

Puuttuva tieto alkiolla k

(1) Keskiarvoimputointi RM

Tulosmuuttujan *y* keskiarvo saadussa datassa ($n_{(r)} = 6$)

$$\hat{y}_k = \bar{y}_{(r)} = 1049.33$$

Imputointi:

Kuhmoinen $k = 18$ $\hat{y}_{18} = 1049.33$

Toivakka $k = 30$ $\hat{y}_{30} = 1049.33$

(2) Lähimmän naapurin menetelmä NN

Tutkitaan, millä alkiolla $l \neq k$ etäisyys $|z_l - z_k|$ saavuttaa minimin. Imputointi:

Kuhmoinen $k = 18$

Minimi on $|1949 - 1463| = 486$

Donor: Pihtipudas $\hat{y}_{18} = y_{26} = 331$

Toivakka $k = 30$

Minimi on $|932 - 834| = 98$

Donor: Uurainen $\hat{y}_{30} = y_{31} = 219$

(3) Suhde-estimointi RA

Lasketaan saadusta aineistosta suhde-estimaatti

$$\hat{B} = \bar{y}_{(r)} / \bar{z}_{(r)} = 1049.33 / 6490.17 = 0.1617$$

Lasketaan sovitteet

$$\hat{y}_k = \hat{B} \times z_k$$

Imputointi:

Kuhmoinen $k = 18$ $z_{18} = 1463$

$$\hat{y}_{18} = 236.57 = 0.1617 \times 1463$$

Toivakka $k = 30$ $z_{30} = 834$

$$\hat{y}_{30} = 134.86 = 0.1617 \times 834$$

Table 4.4 Completed data sets obtained by single imputation methods (The *Province'91* population).

ID	Element	Response data (Sample)		Imputed data sets by model			
	LABEL	UE91	HOU85	(1) Respondent mean RM	(2) Nearest neighbour NN	(3) Ratio estimation RA	Full response
18	Kuhmoinen	. .	1463	1049.33*	331*	236.57*	187
30	Toivakka	. .	834	1049.33*	219*	134.86*	127
1	Jyväskylä	4123	26 881	4123	4123	4123	4123
4	Keuruu	760	4 896	760	760	760	760
5	Saarijärvi	721	3 730	721	721	721	721
15	Kongink.	142	556	142	142	142	142
26	Pihtipudas	331	1 946	331	331	331	331
31	Uurainen	219	932	219	219	219	219

Imputed values are flagged with " * " and missing values with " . . "

Sampling rate for respondent data is 6/32 = 0.1875

Sampling rate for "Full response"and completed data sets is 8/32 = 0.2500

Totaaliestimaattorin varianssin estimointi

Varianssiestimaattori

$$\hat{v}(\hat{t}^*_{ht}) = \hat{v}_{sam}(\hat{t}^*_{ht}) + \hat{v}_{imp}(\hat{t}^*_{ht})$$

Asetelmavarianssin estimointi:

$$\hat{v}_{sam}(\hat{t}^*_{ht}) = N^2(1-\frac{n}{N}) \times \hat{s}^2_{(r)} / n_{(r)}$$

$$= 32^2(1-\frac{8}{32}) \times 1527.59^2 / 6 = 14967^2$$

missä $\hat{s}^2_{n_{(r)}} = \sum_{k=1}^{n_{(r)}} (y_k - \bar{y}_{(r)})^2 / (n_{(r)} - 1)$

on laskettu saadusta aineistosta

$\hat{v}_{sam}(\hat{t})$ on sama kaikille estimaattoreille (1)-(3)

Imputointivarianssin estimointi

Varianssiestimaattori:

$$\hat{v}_{imp}(\hat{t}^{*}_{ht}) = N^2(1-\frac{n_{(r)}}{n}) \times \frac{\sum_{k=1}^{n_{(r)}}(\hat{e}_k - \bar{\hat{e}})^2}{n_{(r)}-1} / n_{(r)}$$

missä $\bar{\hat{e}} = \sum_{k=1}^{n_{(r)}} \hat{e}_k / n_{(r)}$

on jäännösten $\hat{e}_k = y_k - \hat{y}_k$ keskiarvo

Jäännökset:

(1) RM: $\hat{e}_k = y_k - \bar{y}_{(r)}$.

(2) NN: $\hat{e}_k = y_k - y_{k(l)}$

missä $y_{k(l)}$ on donorin *y*-arvo

(3) RA: $\hat{e}_k = y_k - (\bar{y}_{(r)} / \bar{z}_{(r)}) \times z_k$

Imputointivarianssin estimaatit:

(1) RM (respondent mean):

$$\hat{v}_{imp}(\hat{t}^{*}_{rm}) = 32^2(1-\frac{6}{8})\times 1527.59^2/6$$
$$= 9978.18^2$$

(2) NN (nearest neighbour):

$$\hat{v}_{imp}(\hat{t}^{*}_{nn}) = 32^2(1-\frac{6}{8})\times 1365.21^2/6 = 8917.51^2$$

(3) RA (ratio estimation):

$$\hat{v}_{imp}(\hat{t}^{*}_{ra}) = 32^2(1-\frac{6}{8})\times 120.29^2/6 \ = 785.73^2.$$

HUOM:

Pienin imputointivarianssin estimaatti on RA-menetelmälle

(4) Moni-imputointi MI

Käytetään HD-menetelmää (Hot Deck)

Muodostetaan m = 5 täydellistä dataa

Imputoidut datat: Table 4.5

Varianssin estimointi

$$\hat{v}(\hat{t}_{mi}) = \hat{v}_{sam}(\hat{t}_{mi}) + \hat{v}_{imp}(\hat{t}_{mi})$$

Lasketaan totaaliestimaattien keskiarvo

$$\hat{t}^*_{mi} = \sum_{j=1}^{m} \hat{t}^*_j / m = (1/5)(28792 + 31108 + 28944 + 44716 + 29100)$$

$$= 32532$$

Table 4.5 Imputed data sets obtained by multiple imputation (*m*=5). Hot deck imputation is used for each completed data set (The *Province'91* population).

ID	Element	Response data (sample) UE91	Repeated samples including imputed values and flagged as " * "					Full response
			1	2	3	4	5	
18	Kuhm.	. .	760*	760*	721*	4123*	760*	187
30	Toivakka	. .	142*	721*	219*	760*	219*	127
1	Jyväskylä	4123	4123	4123	4123	4123	4123	4123
4	Keuruu	760	760	760	760	760	760	760
5	Saarijärvi	721	721	721	721	721	721	721
15	Kongink.	142	142	142	142	142	142	142
26	Pihtipudas	331	331	331	331	331	331	331
31	Uurainen	219	219	219	219	219	219	219
	Mean	1049.33	899.75	972.12	904,50	1397.38	909.37	826.25
	STD (y)	1527.59	1330.71	1298.98	1325.42	1699.99	1324.72	1355.15

Imputointien sisäinen varianssikomponentti:

$$\hat{v}_{sam} = \frac{1}{m} \times \sum_{j=1}^{m} \hat{v}_{srswor}(\hat{t}_j^*)$$

$$= \frac{1}{5} \times (1 - \frac{8}{32}) \times 32^2 \times (1330.715^2 + 1298.982^2$$

$$+ 1325.416^2 + 1699.989^2 + 1324.716^2) / 6$$

$$= 13758.87^2$$

Imputointien välinen varianssikomponentti:

$$\hat{v}_{imp} = (1 + \frac{1}{m}) \times \sum_{j=1}^{m} \frac{(\hat{t}_j^* - \bar{\hat{t}}_{mi}^*)^2}{m-1}$$

$$= 1.2 \times 6876.444^2 = 7532.39^2$$

Estimaattorin $\hat{t}_{mi}^*$ varianssiestimaatti:

$$\hat{v}(\hat{t}_{mi}^*) = \hat{v}_{sam} + \hat{v}_{imp} = 13758.87^2 + 7532.39^2$$

$$= 15686.86^2$$

Table 4.6 Estimates of a total and its standard error under various imputation methods (the *Province'91* population).

Model type	Estimator	Estimate for a total	$\hat{v}$	$\hat{v}_{sam}$	$\hat{v}_{imp}$
(0) No adj. $n_{(r)} = 6$	$\hat{t}_{ht(r)}$	33 579	$17\,988^2$	$17\,988^2$	0
(1) RM	$\hat{t}^*_{ma}$	33 579	$17\,988^2$	$14\,967^2$	$9\,978^2$
(2) NN	$\hat{t}^*_{nn}$	27 384	$17\,422^2$	$14\,967^2$	$8\,918^2$
(3) RA	$\hat{t}^*_{ra}$	26 669	$14\,988^2$	$14\,967^2$	786^2
(4) MI $m = 5$	$\hat{t}^*_{mi}$	32 532	15686^2	13759^2	7532^2
(5) Full $n = 8$	$\hat{t}_{ht}$	26 440	$13\,282^2$	$13\,282^2$	0