Helsingin yliopisto

Otanta-aineistojen analyysi

Kevät 2014 Periodi III

Risto Lehtonen

Teema 2
Estimaattoreiden varianssien estimointi

Estimaattoreiden varianssin estimointi linearisointimenetelmällä ja pseudotoisto-otannan menetelmillä

Linearisointimenetelmä
Linearization method
Taylor series expansion

Pseudotoisto-otantaan perustuvat menetelmät
Pseudoreplication, Sample re-use

Jackknife-menetelmä (JACKKNIFE)
Balanced Repeated Replications (BRR)
Bootstrap-menetelmä (BOOT)

LINEARISOINTIMENETELMÄ

Menetelmä on yleisimmin käytetty survey-analyysin ohjelmistoissa:

SAS-proseduurit: SURVEYMEANS, SURVEYREG, SURVEYFREQ, SURVEYLOGISTIC, SURVEYPHREG
SPSS:n Complex Surveys-moduli
Stata: SVY-ohjelmat

Epälineaariset estimaattorit

Osajoukon koko satunnaismuuttuja
- Osajoukkojen osuusestimaattorit
- Osajoukkojen keskiarvoestimaattorit

Regressiokertoimien estimaattorit
Logitmallin kerroinestimaattorit

HUOM:
Ohjelmasovelluksissa (SAS, SPSS, Stata) estimaattoreiden varianssien estimointi perustuu otosrypäiden välisen varianssin estimointiin osittaittain

Poikkeus: SUDAAN-ohjelmisto

Point and variance estimation 1

- **Basic principles**

In software products (SAS, SPSS, Stata), design-based variances are estimated based on the variation at the cluster (PSU) level only

- Motivation:

$$\hat{\mathrm{v}}(\hat{\theta}) \cong \frac{\hat{v}_1}{m} + \frac{\hat{v}_2}{m \times n}$$

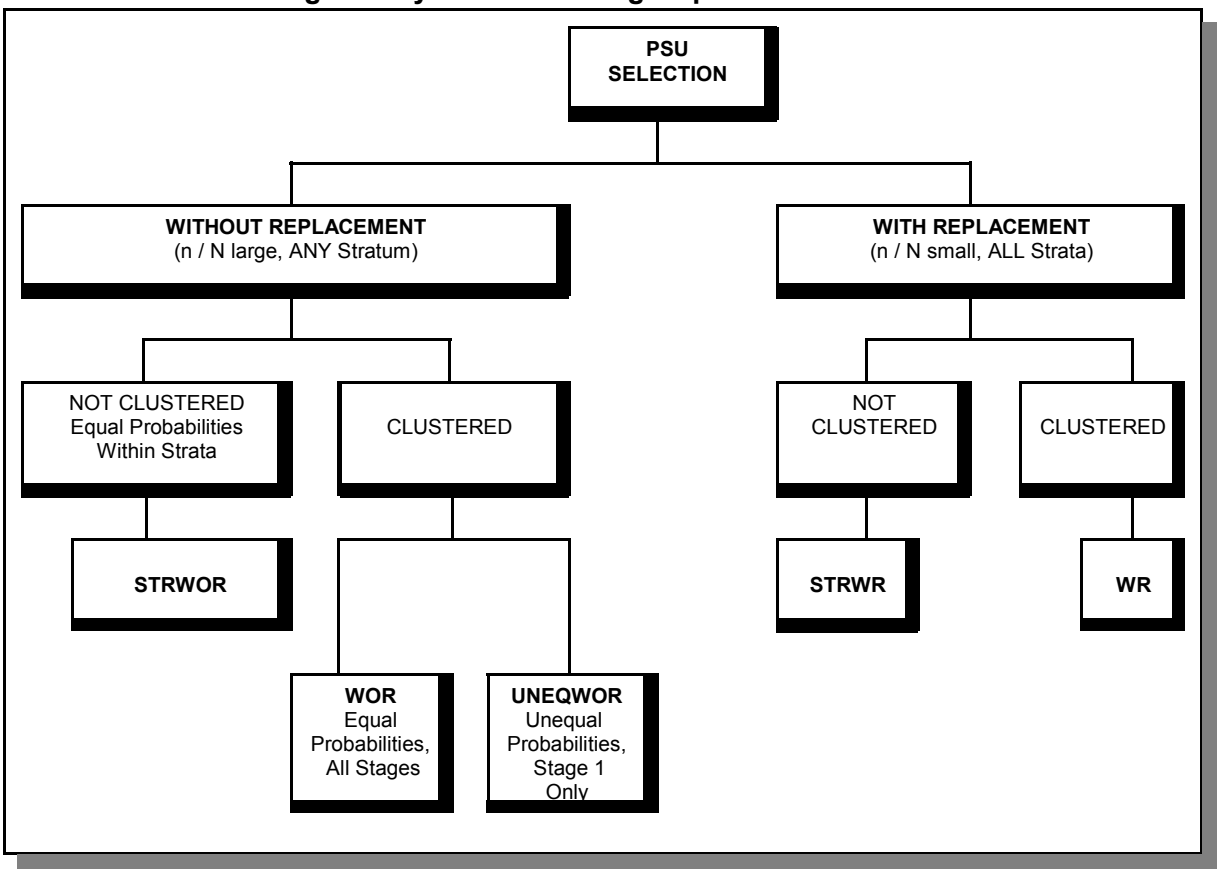
where $\hat{v}_1$ is PSU level variance term

$\hat{v}_2$ is element-level variance term

m is number of sample PSUs

n is number of sample elements

Exhibit 3-1. Choosing the Taylor Series Design Option



* PERUSJOUKON OSAJOUKKOJA KOSKEVIEN OSUUKSIEN JA KESKIARVOJEN ESTIMOINTI

Perusjoukko jaettu D **osajoukkoon** $U_1,...,U_D$

Binäärinen (0/1) indikaattorimuuttuja δ

$\delta_{jhik}=1$ jos ositteen h rypään i alkio $k \in U_j$
$= 0$ muulloin

Binäärinen (0/1) tulosmuuttuja y

$y_{hik}=1$ jos ositteen h rypään i alkiolla k on tutkittava ominaisuus
$= 0$ muulloin

Estimoitavana osuusparametri

$$p_j = \frac{\sum_{h=1}^{H} \sum_{i=1}^{M_h} \sum_{k=1}^{N_{hi}} \delta_{jhik} y_{hik}}{N_j} = \frac{T_j}{N_j} \qquad (j=1,...,D)$$

missä

H ositteiden lkm
M_h perusjoukon rypäiden lkm ositteessa h
N_{hi} perusjoukon alkioiden lkm ositteen h rypäässä i
T_j tulosmuuttujan y totaali osajoukossa j
N_j osajoukon alkioiden lkm

* SUHTEEN JA OSUUDEN ESTIMAATTORI
Combined ratio estimator

Osuusestimaattori $\hat{p}_j$, j=1,...,D (D osajoukkoa)

$$\hat{p}_j = \frac{y_j}{x_j} = \frac{\hat{t}_j}{\hat{N}_j} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{m_h} y_{jhi}}{\sum_{h=1}^{H}\sum_{i=1}^{m_h} x_{jhi}} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{m_h}\sum_{k=1}^{x_{hi}} \delta_{jhik} w_{hik} y_{hik}}{\sum_{h=1}^{H}\sum_{i=1}^{m_h}\sum_{k=1}^{x_{hi}} w_{hik}}$$

missä $\hat{t}_j$ Tulosmuuttujan y totaaliestimaattori osajoukossa j

$\hat{N}_j$ Osajoukon koon estimaattori

$y_j=(n/\hat{N})\hat{t}_j$ ja $x_j=(n/\hat{N})\hat{N}_j$ vastaavat skaalatut luvut

w_{hik} Painomuuttuja

HUOM: Analyyttisissa tutkimuksissa painot w skaalataan usein niin, että niiden keskiarvo koko aineistossa = 1

Suhteen estimaattori on yksinkertaisin esimerkki epälineaarisesta estimaattorista

HUOM: Merkintätapa x_j (eikä n_j) korostaa sitä, että myös nimittäjä on satunnaismuuttuja

* LINEARISOINTIMENETELMÄ

Suhteen estimaattorissa sekä osoittaja y_j että nimittäjä x_j ovat satunnaismuuttujia

Tästä syystä asetelmaperusteisen varianssiestimaattorin tulee käsittää:

- osoittajan varianssi *v(y)*
- nimittäjän varianssi *v(x)*
- osoittajan ja nimittäjän kovarianssi *cov(y,x)*

Osajoukon osuustunnusluvun $\hat{p}_j$ linearisointimenetelmään perustuva varianssiestimaattori on:

$$\hat{v}_{des}(\hat{p}_j)=\hat{p}_j^2(y_j^{-2}\hat{v}(y_j)+x_j^{-2}\hat{v}(x_j)-2(y_jx_j)^{-1}\hat{cov}(y_j,x_j))$$

HUOM: Vastaava malliperusteinen (binominen, SRS-perusteinen) varianssiestimaattori:

$$\hat{v}_{bin}(\hat{p}_j)=\hat{p}_j(1-\hat{p}_j)/\hat{n}_j$$

LINEARISOINTIMENETELMÄ

Lehtonen R. and Pahkinen E. (2004). *Practical Methods for Design and Analysis of Complex Surveys. Second Edition.* Wiley.

Linearisointimenetelmään perustuva asetelmaperusteinen **varianssiapproksimaatio** suhteen estimaattorille (osuusestimaattorille) (*combined ratio estimator*)

Linearisointimenetelmään perustuva osuusestimaattorivektorin asetelmaperusteinen **kovarianssimatriisiestimaattori**

ESIMERKKI: OHC Survey

Ositettu kaksiasteinen ryväsotanta

Point and variance estimation 2

- **Noninear statistics for complex samples involving clustering**
- **Domain means** $\bar{y}_d = \hat{t}_d / \hat{N}_d$ where the denominator is not fixed by the design but is estimated and thus is random variate
- **Regression coefficients**
- Design-based variance estimation: Analytical formulas not available
- Variance estimation by **approximate methods -** see Wolter (1985), Lehtonen and Pahkinen (2004)

Technical annex: Design-based covariance matrix of estimator vector

Vector of domain proportion estimators, two domains

$$\hat{\mathbf{p}} = \begin{bmatrix} \hat{p}_1 \\ \hat{p}_2 \end{bmatrix} \text{ where } \hat{p}_d = \frac{y_d}{x_d} = \frac{\sum_{h=1}^{H} \sum_{k=1}^{m_h} y_{dhk}}{\sum_{h=1}^{H} \sum_{k=1}^{m_h} x_{dhk}} \text{ , } d = 1,2$$

Binomial covariance matrix (diagonal by definition)

$$\hat{\mathbf{V}}_{bin} = \begin{bmatrix} \hat{p}_1(1-\hat{p}_1)/n_1 & 0 \\ 0 & \hat{p}_2(1-\hat{p}_2)/n_2 \end{bmatrix}$$

Design-based covariance matrix (nondiagonal by definition)

$$\hat{\mathbf{V}}_{des} = \begin{bmatrix} \hat{v}_{11} & \hat{v}_{21} \\ \hat{v}_{12} & \hat{v}_{22} \end{bmatrix}$$ where $\hat{v}_{dj}$ are design-based covariance estimators

Technical annex (contd.): Design-based covariance matrix of estimator vector

Why $\hat{\mathbf{V}}_{des}$ is allowed nondiagonal?

Because domain proportion estimators $\hat{p}_d$ can be correlated if the domain structure cuts across the cluster structure

Variance estimators $\hat{v}$ based on Taylor linearization:

$$\hat{v}_{11} = \hat{p}_1^2(\mathrm{y}_1^{-2}\hat{v}(y_1) + \mathrm{x}_1^{-2}\hat{v}(x_1) - 2(\mathrm{y}_1\mathrm{x}_1)^{-1}\hat{v}(y_1x_1)), \quad d = 2 \text{ similarly}$$

Covariance estimators derived similarly:

$$\hat{v}_{12} = \hat{p}_1\hat{p}_2((\mathrm{y}_1y_2)^{-1}\hat{v}(y_1,y_2) + (x_1x_2)^{-1}\hat{v}(x_1,x_2) - (\mathrm{y}_1x_2)^{-1}\hat{v}(y_1,x_2))$$
$$-(\mathrm{y}_2x_1)^{-1}\hat{v}(y_2,x_1)) = \hat{v}_{21}$$

ESIMERKKI. Osuusestimaattorin varianssin approksimointi linearisointimenetelmällä
Lehtonen&Pahkinen 2004, Example 5.5

OHC Survey demodata

Ositettu ryväsotanta-asetelma

H= 5 ositetta
m= 250 toimipaikkaa (otosryvästä)
n = 7841 henkilöä

Binäärinen tulosmuuttuja

PHYS Työn fysikaaliset terveyshaitat
0 = Ei ole
1 = On

Estimointi:

Työn fysikaalisista haitoista kärsivien miesten osuus

Osuuden estimaatti:

$$\hat{p}_1 = \frac{y_1}{x_1} = \frac{2061}{4485} = 0.4595$$

Varianssiapproksimaatio:

SAS / SURVEYMEANS

Osuusestimaattorin asetelmaperusteinen varianssiestimaatti linearisointimenetelmän avulla:

$$\hat{v}_{des}(\hat{p}_1)=\hat{p}_1^2(y_1^{-2}\hat{v}(y_1)+x_1^{-2}\hat{v}(x_1)-2(y_1x_1)^{-1}\hat{cov}(y_1,x_1))=0.2775\times10^{-3}$$

SRS-perusteinen (binomimalliin perustuva) varianssiestimaatti:

$$\hat{v}_{bin}(\hat{p}_1)=\hat{p}_1(1-\hat{p}_1)/\hat{n}_1=0.4595(1-0.4595)/4485=0.554\times10^{-4}$$

Estimoitu asetelmakerroin:

$$deff(\hat{p}_1) = 0.0002775/0.0000554 = 5.01$$

Suuri deff-estimaatti viittaa tulosmuuttujan PHYS voimakkaaseen positiiviseen sisäkorrelaatioon rypäissä

Binominen varianssiestimaatti aliestimoi selvästi todellista varianssia

* TOISTO-OTANTAAN PERUSTUVA ESTIMAATTORIN VARIANSSIN APPROKSIMOINTI
Replication / Pseudoreplication methods

“Aito” toisto-otanta (*replication*)

a) Perusjoukosta poimitaan useita toisistaan riippumattomia samankokoisia otoksia samalla otanta-asetelmalla niin, että kokonaisotoskoko on *n*

b) Estimaattoreiden varianssit estimoidaan toisto-otoksista havaitun variaation perusteella
Käytännössä verraten harvinainen menetelmä

“Pseudotoisto”-menetelmät (*pseudoreplication*)

a) Perusjoukosta poimitaan yksi kokoa *n* oleva otos annetulla otanta-asetelmalla

b) Poimitusta *n* alkion otoksesta poimitaan useita pseudotoisto-otoksia annetulla otanta-asetelmalla

c) Estimaattoreiden varianssit estimoidaan pseudotoisto-otoksista havaitun variaation perusteella
Käytännössä verraten yleinen menetelmä

Helsingin yliopisto
Otanta-aineistojen analyysi
Risto Lehtonen

Teema 2 Estimaattoreiden varianssien estimointi

Toisto-otanta

(a) Perusjoukosta poimitaan *K* toisistaan riippumatonta samankokoista otosta $s_1,...,s_K$ samalla otanta-asetelmalla niin, että kokonaisotoskoko on *n*

(b) Estimaattoreiden varianssit estimoidaan toisto-otoksista havaitun variaation perusteella

Estimoitava parametri $\theta = f(T_1,...,T_s)$ (totaalien funktio)

Esim: $\theta = T_1 / T_2$

Estimaattori $\hat{\theta} = f(\hat{t}_1,...,\hat{t}_s)$

Esim: $\hat{\theta} = \hat{t}_1 / \hat{t}_2$

Otoskoko *n*

Otokset $s_1,...,s_K$

Otoskoot $n_1,...,n_K, \quad \sum_{k=1}^{K} n_k = n$

Estimaatit $\hat{\theta}_1,...,\hat{\theta}_K$

Varianssiestimaatti $v(\hat{\theta}) = \sum_{k=1}^{K} (\hat{\theta}_k - \bar{\hat{\theta}})^2 / (K-1)$

$\bar{\hat{\theta}} = \sum_{k=1}^{K} \hat{\theta}_k / K$

* PSEUDOTOISTOMENETELMÄT

Otanta-asetelmat

Perusasetelma: ns. **“Paired clusters design”**

Paljon ositteita

Kustakin ositteesta on poimittu kaksi ryvästä otokseen

Voidaan yleistää mutkikkaampiin asetelmiin, joissa on vaihteleva määrä otosrypäitä per osite

Estimaattorityypit

Epälineaariset estimaattorit, jotka voidaan lausua totaaliestimaattoreiden funktioina

Varianssin approksimoinnin perusmenetelmä

Joustava

Soveltuu yleisesti epälineaarisille estimaattoreille

Laskentaintensiivinen linearisointimenetelmään verrattuna

* PSEUDOTOISTOMENETELMÄT

Varianssiestimaattorin perusmuoto:

$$\hat{v}(\hat{\theta}) = c\sum_{k=1}^{K} (\hat{\theta}_k - \hat{\theta})^2$$

missä $\hat{\theta}_k$ on pseudo-otoksesta *k* laskettu parametrin θ estimaatti

$\hat{\theta}$ on alkuperäisestä otoksesta laskettu parametrin θ estimaatti

c on vakio, joka riippuu valitusta pseudotoistomenetelmästä

K on kullekin pseudotoistomenetelmälle spesifi toistojen lukumäärä

HUOM: Lineaaristen estimaattoreiden tapauksessa kaikki pätevät pseudotoistoperusteiset varianssiestimaatit yhtyvät ja tuottavat vastaavan analyyttisen estimaattorin mukaisen estimaatin

HUOM: Linearisointimenetelmässä osittaisderivaattojen lausekkeet tarvitaan erikseen kullekin estimaattorityypille

* JACKKNIFE-TEKNIIKKA

“Jackknife repeated replications” JRR

McCarthy (1966), Frankel (1971), Wolter (1985)

Pseudo-otosten konstruointi

“Paired clusters design”

H Ositteiden lkm
$m_h = 2$ Otosrypäitä/osite
n Alkiotason otoskoko

Proseduuri:

1. pseudo-otos:
a) Poista ensimmäisen ositteen 1. ryväs
b) Painota toinen ryväs painolla 2
c) Jätä muut *H*-1 ositetta ennalleen

Toista proseduuri kullekin *H* ositteelle
Saadaan kaikkiaan *H* pseudo-otosta (tässä *K*=*H*)

Komplementtiotokset

Muuta rypäiden poistojärjestys kussakin ositteessa
Saadaan *H* komplementtiotosta

* JACKKNIFE-TEKNIIKKA

JRR-varianssiestimaattori
“Paired clusters design”

Estimaattorityypit:

- Osajoukon osuusestimaattorit
- Osajoukon keskiarvoestimaattorit
- Regressiokertoimen estimaattorit
- Logitmallin kerroinestimaattorit

JRR-varianssiestimaattorin perusmuoto:

$$\hat{v}_{JRR}(\hat{\theta}) = \sum_{k=1}^{H} (\hat{\theta}_k - \hat{\theta})^2$$

HUOM: Vakio *c* = 1 JRR-varianssiestimaattorin perusmuodolle

Menettelyllä voidaan konstruoida useita vaihtoehtoisia muotoja:

- Pseudo-otosten avulla
- Komplementtiotosten avulla
- Yhdistelmäestimaattoreina

Ks: Lehtonen&Pahkinen (2004) pp. 156-158

The JRR Technique

The particular jackknife method based on *jackknife repeated replications* has many features of the BRR technique, since only the method of forming the pseudosamples is different. Application of the JRR technique to a design where more than two sample clusters are drawn from a stratum is more straightforward than for BRR. We, however, consider the JRR technique in the simplest case where the number of sample clusters per stratum is exactly two, and the clusters are assumed to be drawn with replacement, i.e. with a design similar to that required for BRR. JRR variance estimators are derived for a ratio estimator $\hat{r}$, which is a subpopulation proportion or mean estimator.

We construct the pseudosamples following the method suggested by Frankel (1971). For the first pseudosample, we exclude the first cluster $h1$ from the first stratum and weight the second cluster $h2$ by the value 2, leaving the remaining $H-1$ strata unchanged. By repeating this procedure for all strata, we get a total of H pseudosamples. For a similar set of H complement pseudosamples, we change the order of the clusters that are excluded. The JRR variance estimators are derived using these two sets of pseudosamples.

Like the BRR technique, several alternative JRR variance estimators can be constructed for the parent ratio estimator $\hat{r}$. For these, we first derive the pseudosample estimators for each stratum. Let $\hat{r}_h$ denote a pseudosample estimator based on excluding cluster $h1$ and duplicating cluster $h2$ in stratum h:

$$\hat{r}_h = \frac{2y_{h2} + \sum_{h' \neq h}^{H} \sum_{i=1}^{2} y_{h'i}}{2x_{h2} + \sum_{h' \neq h}^{H} \sum_{i=1}^{2} x_{h'i}}, \quad h = 1, \ldots, H. \tag{5.19}$$

These estimators are constructed for each pseudosample. From the complement pseudosamples, we obtain corresponding estimators $\hat{r}_h^c$ by excluding cluster $h2$ and duplicating cluster $h1$. Using the pseudosample estimators and the complement pseudosample estimators, we can derive the first set of JRR variance estimators for the parent estimator $\hat{r}$. Hence we have

$$\hat{v}_{1.jrr}(\hat{r}) = \sum_{h=1}^{H} (\hat{r}_h - \hat{r})^2, \tag{5.20}$$

and from the complement pseudosamples

$$\hat{v}_{2.jrr}(\hat{r}) = \sum_{h=1}^{H} (\hat{r}_h^c - \hat{r})^2. \tag{5.21}$$

A combined variance estimator is

$$\hat{v}_{3.jrr}(\hat{r}) = (\hat{v}_{1.jrr}(\hat{r}) + \hat{v}_{2.jrr}(\hat{r}))/2. \tag{5.22}$$

Another set of variance estimators can be obtained using the so-called *pseudovalues* introduced by Quenouille (1956) to reduce the bias of an estimator. In the case considered above, pseudovalues are of the form

$$\hat{r}_h^p = 2\hat{r} - \hat{r}_h, \quad h = 1, \ldots, H, \tag{5.23}$$

and for the complement pseudosamples they are denoted by $\hat{r}_h^{pc}$. By using the first set of H pseudovalues $\hat{r}_h^p$, we obtain a bias-corrected estimator given by

$$\bar{\hat{r}}^p = \sum_{h=1}^{H} \hat{r}_h^p / H, \tag{5.24}$$

and using the pseudovalues $\hat{r}_h^{pc}$ from the complement pseudosamples we obtain

$$\bar{\hat{r}}^{pc} = \sum_{h=1}^{H} \hat{r}_h^{pc} / H. \tag{5.25}$$

Counterparts to the variance estimators (5.20)–(5.22) can be derived from the pseudovalues and the bias-corrected estimators, giving

$$\hat{v}_{4.jrr}(\hat{r}) = \sum_{h=1}^{H} (\hat{r}_h^p - \bar{\hat{r}}^p)^2, \tag{5.26}$$

and from the complement pseudosamples

$$\hat{v}_{5.jrr}(\hat{r}) = \sum_{h=1}^{H} (\hat{r}_h^{pc} - \bar{\hat{r}}^{pc})^2. \tag{5.27}$$

A combined variance estimator can also be derived:

$$\hat{v}_{6.jrr}(\hat{r}) = (\hat{v}_{4.jrr}(\hat{r}) + \hat{v}_{5.jrr}(\hat{r}))/2. \tag{5.28}$$

Finally, from all the $2H$ pseudosamples we obtain:

$$\hat{v}_{7.jrr}(\hat{r}) = \sum_{h=1}^{H} (\hat{r}_h - \hat{r}_h^c)^2 / 4. \tag{5.29}$$

A similar way of constructing the JRR variance estimators was used to that given for the BRR technique. For a linear estimator, the bias-corrected JRR estimators reproduce the parent estimator, and all the JRR variance estimators coincide. This is not the case for nonlinear estimators, but in practice all JRR variance estimators should give closely related results. Like BRR, the variance estimator $\hat{v}_{7.jrr}$ could be taken as the most natural estimator of the variance of the parent estimator $\hat{\theta}$.

The JRR technique can be extended to a more general case in which more than two clusters are drawn from each stratum, for without-replacement sampling of clusters. Pseudosamples and their complements are constructed by consecutively excluding a cluster and weighting the remaining clusters appropriately in a stratum (see Section 4.6 in Wolter 1985).

Like BRR, we use the JRR technique for variance estimation of a ratio estimator $\hat{r}$ for the MFH Survey design.

Example 5.3

The JRR technique in the MFH Survey. We continue to consider the estimation of variance of a ratio-type subpopulation proportion estimator $\hat{p}$ of CHRON (chronic morbidity) and a subpopulation mean estimator $\overline{y}$ of SYSBP (systolic blood pressure) for 30–64-year-old males. Using the cluster-level data set available, we calculate all the seven JRR variance estimates for $\hat{p}$ and $\overline{y}$.

Because $H = 24$, we construct 24 JRR pseudosamples with their complements by the Frankel method. The parent ratio and mean estimates $\hat{p}$ and $\overline{y}$, and the corresponding bias-corrected estimators given by (5.24) and (5.25) based on the pseudovalues $\hat{p}_h^p$, $\hat{p}_h^{pc}$, $\overline{y}_h^p$ and $\overline{y}_h^{pc}$ calculated from the pseudosamples and their complements, are first obtained. These are

$$\hat{p} = 0.3976,\ \overline{\hat{p}}^p = \sum_{k=1}^{24} \hat{p}_k^p/24 = 0.3972 \quad \text{and} \quad \overline{\hat{p}}^{pc} = \sum_{k=1}^{24} \hat{p}_k^{pc}/24 = 0.3980,$$

$$\overline{y} = 141.785,\ \overline{\hat{y}}^p = \sum_{k=1}^{24} \overline{y}_k^p/24 = 141.793 \quad \text{and} \quad \overline{\hat{y}}^{pc} = \sum_{k=1}^{24} \overline{y}_k^{pc}/24 = 141.777.$$

All three CHRON proportion estimates and SYSBP mean estimates are close. Next we calculate the JRR variance estimates. For a CHRON proportion estimator $\hat{p}$ the first variance estimate (5.20) is

$$\hat{v}_{1.jrr}(\hat{p}) = \sum_{h=1}^{24} (\hat{p}_h - 0.3976)^2 = 0.1099 \times 10^{-3},$$

ESIMERKKI. Varianssin approksimointi JRR-menetelmällä

OHC-demodata

a) Alkuperäinen otanta-asetelma

Ositettu ryväsotanta-asetelma
H=5 ositetta
m=250 otosryvästä
n=7841 henkilöä

b) Modifioitu asetelma

“Paired clusters design”-asetelma
H = 125 ositetta
m = 250 otosryvästä
2 otosryvästä per osite
n = 7841 henkilöä

Binäärinen tulosmuuttuja

PHYS Työn fysikaaliset terveyshaitat
0 = Ei ole
1 = On

Estimointi

Työn fysikaalisista haitoista kärsivien miesten osuus

Osuuden estimaatti

$$\hat{p}_1 = \frac{y_1}{x_1} = \frac{2061}{4485} = 0.4595$$

Osuuestimaattorin varianssiapproksimaatiot

	Varianssi-estimaatti	*deff*
a) Alkuperäinen asetelma		
JRR	0.0002788	5.03
Linearisointi	0.0002775	5.01
b) Modifioitu asetelma		
JRR	0.0002298	4.15
Linearisointi	0.0002298	4.15

* BOOTSTRAP-TEKNIIKKA

“Bootstrap repeated replications“

McCarthy and Snowden (1985)
Rao and Wu (1988)
Rao et al. (1992)

Pseudo-otosten konstruointi

Ositettu ryväsotanta-asetelma

H Ositteiden lkm
$m_h = a$ (≥ 2) Vakiomäärä otosrypäitä per osite
n Alkiotason otoskoko

Pseudo-otosten konstruointitapa poikkeaa huomattavasti JRR- ja BRR-tekniikoista

Bootstrap on laskentaintensiivisempi tapa

Ei toistaiseksi implementoitu survey-analyysin ohjelmistoihin

BOOT-proseduuri

Vaihe 1. Poimi kokoa *a* oleva SRS-WR-otos ositteen *h* otosrypäistä, *h*=1,...,*H*

HUOM: WR-tyyppinen poiminta
Poiminta suoritetaan toisistaan riippumattomasti jokaisessa *H* ositteessa

Saadaan kokoa *m* oleva bootstrap-otos

Vaihe 2. Toista vaihe 1 kaikkiaan *K* kertaa (Esim. *K*=1000)

Saadaan yhteensä *K* riippumatonta bootstrap-otosta

Bootstrap-varianssiestimaattori

Perusmuoto:

$$v_{BOOT}(\hat{\theta}) = \frac{a}{a-1} \sum_{k=1}^{K} (\hat{\theta}_k - \hat{\theta})^2 / K$$

ESIMERKKI: Tehdään harjoituksissa

$$\hat{v}_{4.jrr}(\overline{y}) = \sum_{h=1}^{24} (\overline{y}_h^p - 141.793)^2 = 0.2759,$$

$$\hat{v}_{5.jrr}(\overline{y}) = \sum_{h=1}^{24} (\overline{y}_h^{pc} - 141.777)^2 = 0.2789,$$

$$\hat{v}_{6.jrr}(\overline{y}) = (\hat{v}_{4.jrr}(\overline{y}) + \hat{v}_{5.jrr}(\overline{y}))/2 = 0.2774,$$

$$\hat{v}_{7.jrr}(\overline{y}) = \sum_{h=1}^{24} (\overline{y}_h - \overline{y}_h^c)^2/4 = 0.2788.$$

For SYSBP, the JRR variance estimates of $\overline{y}$ are also very close. All the JRR variance estimators of a proportion estimator and a mean estimator provided closely related numerical results. Therefore, either practical or computational considerations can guide the selection of an appropriate JRR variance estimator. The jackknife technique is available in some software products for the analysis of complex surveys.

The BOOT Technique

Similar to the other sample reuse methods, the bootstrap can be used for variance approximation of a nonlinear estimator under a complex sampling design. The method, however, differs from BRR and JRR in many respects, e.g. the generation of pseudosamples is quite different. We consider the bootstrap technique for variance estimation of a ratio estimator under a two-stage stratified epsem design where a constant number of clusters (which may be greater than two) is drawn with replacement from each stratum. We adopt a simple version of the bootstrap, introduced in Rao and Wu (1988) as a *naive bootstrap*, for this kind of design, and call it the *BOOT technique*.

Let us assume that $m_h = a\ (\geq 2)$ clusters are drawn with replacement from each of the H strata. The number of sample clusters is thus $m = a \times H$. We construct the bootstrap pseudosamples in the following way:

Step 1. From the a sample clusters in stratum h, draw a simple random sample of size a with replacement. This is performed independently in each stratum. The resulting H simple random samples together constitute a *bootstrap sample* of m clusters.

Step 2. Repeating Step 1 K times, a total of K independent bootstrap samples are obtained.

It is important in Step 1 that the simple random samples in each stratum are drawn with replacement, and the stratum-wise samples are drawn independently.

So, a particular sample cluster in a stratum may be included in a bootstrap sample many (even a) times, or not at all.

We consider the BOOT technique for the estimation of the variance of the ratio estimator $\hat{r}$. A ratio estimator for a bootstrap sample k is denoted by $\hat{r}_k$ $(k = 1, \ldots, K)$. The mean of the bootstrap sample estimates $\hat{r}_k$ provides a *bootstrap estimator*

$$\bar{\hat{r}} = \sum_{k=1}^{K} \hat{r}_k / K. \tag{5.30}$$

A Monte Carlo variance estimator based on $\hat{r}_k$ and the bootstrap estimator (5.30) is first derived for the parent estimator $\hat{r}$:

$$\hat{v}_{mc}(\hat{r}) = \sum_{k=1}^{K} (\hat{r}_k - \bar{\hat{r}})^2 / K. \tag{5.31}$$

Unfortunately, this intuitively attractive variance estimator is unacceptable because it is not consistent for the variance of $\hat{r}$ and, moreover, it is not unbiased even for the variance of a linear estimator, as Rao and Wu (1988) have shown. But in the case considered, where a constant number of clusters is drawn from each stratum, an appropriately rescaled Monte Carlo variance estimator provides a consistent variance estimator for the parent estimator $\hat{r}$. Hence the first BOOT variance estimator is

$$\hat{v}_{1.boot}(\hat{r}) = \frac{a}{a-1} \hat{v}_{mc}(\hat{r}) = \frac{a}{a-1} \sum_{k=1}^{K} (\hat{r}_k - \bar{\hat{r}})^2 / K. \tag{5.32}$$

By using the parent estimator $\hat{r}$ in place of the bootstrap estimator, another variance estimator is obtained:

$$\hat{v}_{2.boot}(\hat{r}) = \frac{a}{a-1} \sum_{k=1}^{K} (\hat{r}_k - \hat{r})^2 / K. \tag{5.33}$$

It should be noticed that for the naive bootstrap there is no obvious solution to the scaling problem in the case in which the number of sample clusters per stratum varies. Rao and Wu (1988) derive a *rescaling bootstrap* for these cases, based on drawing simple random samples of size m_h (≥ 1) clusters with replacement from a stratum. With appropriate selection of m_h, different versions of the bootstrap are provided. Sitter (1992) proposes a generalization of this method, based on resampling without replacement rather than with replacement, and repeating this many times with replacement. Rao *et al.* (1992) redefine the rescaling bootstrap to be also suitable for variance estimation of non-smooth functions such as the median.

In the BOOT technique, to obtain variance estimation results with sufficient precision the number K of bootstrap samples should be large, preferably 500 to 1000. The technique thus requires large processing capabilities and can consume a lot of computer resources. In this, the BOOT technique is more obviously computer-intensive than BRR and JRR.

Example 5.4

The BOOT technique in the MFH Survey. We apply the BOOT technique for variance approximation of subpopulation proportion and mean estimators $\hat{p}$ (for CHRON) and $\overline{y}$ (for SYSBP), both considered as ratio estimators. The MFH Survey subgroup consists of 2699 males aged 30–64 years. In the MFH Survey design there are $H = 24$ strata each with $a = 2$ sample clusters, so each bootstrap sample constitutes of $m = 2 \times 24 = 48$ clusters. In the generation of the bootstrap samples we use the cluster-level data set. We obtain a bootstrap sample by drawing a simple random sample of two clusters with replacement, independently from each stratum. Thus, a cluster in a stratum can appear in a bootstrap sample either 0, 1 or 2 times so that the sample size from a stratum is always 2 clusters. Note that the number of such samples can become large; e.g. if we have 1000 bootstrap samples, a total of 24 000 independent samples of size 2 must be drawn. In this example, $K = 1000$ bootstrap samples.

An estimate $\hat{r}_k$ mimicking the parent estimator $\hat{r}$ is calculated from each of the K bootstrap samples. A bootstrap estimate is then calculated as an average of the $\hat{r}_k$. By using the $\hat{r}_k$, the bootstrap estimate and the parent estimate, we finally obtain BOOT variance estimates $\hat{v}_{1.boot}(\hat{r})$ and $\hat{v}_{2.boot}(\hat{r})$.

With $K = 1000$ bootstrap samples, the distribution of the bootstrap sample estimates for CHRON and SYSBP are displayed in Figure 5.1. The parent estimates and the bootstrap estimates (5.30) for CHRON proportion and SYSBP mean are

$$\hat{p} = 0.3976, \text{ and the bootstrap estimate is } \overline{\hat{p}} = 0.3973,$$
$$\overline{y} = 141.785, \text{ and the bootstrap estimate is } \overline{\hat{y}} = 141.783.$$

The BOOT variance estimates (5.32) and (5.33) for CHRON proportion $\hat{p}$ are, respectively

$$\hat{v}_{1.boot}(\hat{p}) = 2 \times \sum_{k=1}^{1000} (\hat{p}_k - 0.3973)^2/1000 = 0.1039 \times 10^{-3}$$

and

$$\hat{v}_{2.boot}(\hat{p}) = 2 \times \sum_{k=1}^{1000} (\hat{p}_k - 0.3976)^2/1000 = 0.1040 \times 10^{-3}.$$

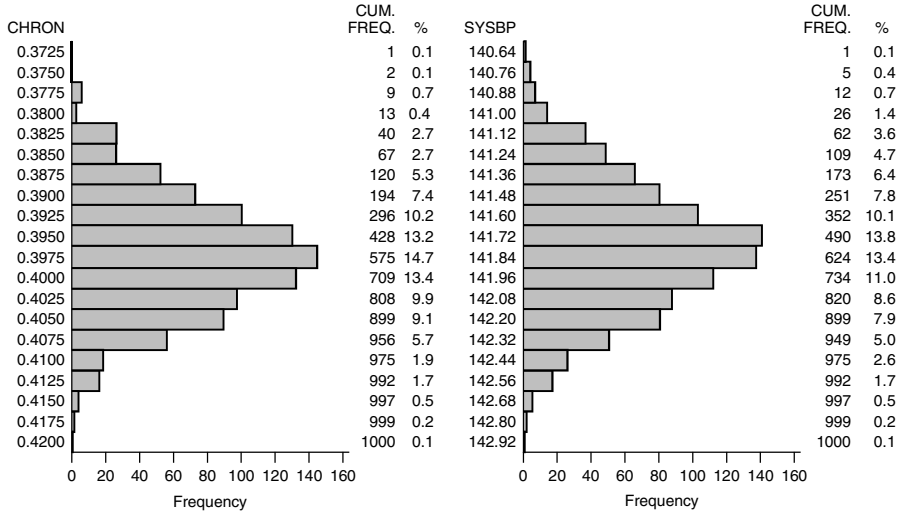


Figure 5.1 Bootstrap histograms for CHRON (a binary variable) and SYSBP (a continuous variable) from the bootstrap estimates $\hat{r}_k$ with $K = 1000$ bootstrap samples.

The BOOT variance estimates for SYSBP mean $\overline{y}$ are

$$\hat{v}_{1.boot}(\overline{y}) = 2 \times \sum_{k=1}^{1000} (\overline{y}_k - 141.783)^2/1000 = 0.2798$$

and

$$\hat{v}_{2.boot}(\overline{y}) = 2 \times \sum_{k=1}^{1000} (\overline{y}_k - 141.785)^2/1000 = 0.2798.$$

For a CHRON proportion estimator $\hat{p}$ and a SYSBP mean estimator $\overline{y}$, both BOOT variance estimates are approximately equal. As in the other reuse methods, any definite preference for the type of variance estimator has not been suggested. From a computational point of view, the estimator $\hat{v}_{2.boot}$ is simpler than $\hat{v}_{1.boot}$.

5.5 COMPARISON OF VARIANCE ESTIMATORS

The linearization method and sample reuse methods were used as basic approximation techniques for variance estimation of a nonlinear ratio estimator. It was assumed that the sample was from a two-stage epsem sampling design with at least two clusters drawn with replacement from each stratum. The linearization method was considered under a design with a varying number (≥ 2) of sample