HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Otanta-aineistojen analyysi
(78136 , 78405)
Kevät 2010

TEEMA 2: Estimaattoreiden varianssin estimointi

Risto Lehtonen
risto.lehtonen@helsinki.fi

Estimaattoreiden asetelmaperusteinen varianssien estimointi 1

- **Linearisointimenetelmä**
  - Johdantoa Diat 2a
  - Lehtonen&Pahkinen (2004) Diat 2b
  - SAS 9.1.3 SURVEYMEANS
  - Laskentaesimerkki

Estimaattoreiden asetelmaperusteinen varianssien estimointi 2

- **Pseudotoisto-otantaan perustuvat menetelmät**
  - Johdantoa
- **Jackknife-menetelmä**
  - SAS 9.2. Jackknife-proseduurit
- **Bootstrap-menetelmä**
- Diat 3b

Estimaattoreiden asetelmaperusteinen varianssien estimointi 3

- **Laskentatyökaluja**

VLISS-Virtual laboratory in survey sampling

http://mathstat.helsinki.fi/VLISS/

- Chapter 5
- SAS-koodit ja makrot
  - Linearization method
  - JRR technique **%macro JRR**
  - Bootstrap method **%macro BOOT**

The SURVEYMEANS Procedure

Statistical Computations

The SURVEYMEANS procedure uses the Taylor expansion method to estimate sampling errors of estimators based on complex sample designs. This method obtains a linear approximation for the estimator and then uses the variance estimate for this approximation to estimate the variance of the estimate itself (Woodruff 1971, Fuller 1975). When there are clusters, or PSUs, in the sample design, the procedure estimates variance from the variation among PSUs. When the design is stratified, the procedure pools stratum variance estimates to compute the overall variance estimate. For *t* tests of the estimates, the degrees of freedom equals the number of clusters minus the number of strata in the sample design.

For a multistage sample design, the variance estimation method depends only on the first stage of the sample design. So, the required input includes only first-stage cluster (PSU) and first-stage stratum identification. You do not need to input design information about any additional stages of sampling. This variance estimation method assumes that the first-stage sampling fraction is small, or the first-stage sample is drawn with replacement, as it often is in practice.

Quite often in complex surveys, respondents have unequal weights, which reflect unequal selection probabilities and adjustments for nonresponse. In such surveys, the appropriate sampling weights must be used to obtain valid estimates for the study population.

For more information on the analysis of sample survey data, refer to Lee, Forthoffer, and Lorimor (1989), Cochran (1977), Kish (1965), and Hansen, Hurwitz, and Madow (1953).

Definition and Notation

For a stratified clustered sample design, together with the sampling weights, the sample can be represented by an $n \times (P+1)$ matrix

$$\begin{aligned}(\mathbf{w}, \mathbf{Y}) &= (w_{hij}, \mathbf{y}_{hij}) \\ &= \left(w_{hij}, y_{hij}^{(1)}, y_{hij}^{(2)}, \ldots, y_{hij}^{(P)}\right)\end{aligned}$$

where

- $h = 1, 2, \ldots, H$ is the stratum number, with a total of H strata
- $i = 1, 2, \ldots, n_h$ is the cluster number within stratum h, with a total of n_h clusters
- $j = 1, 2, \ldots, m_{hi}$ is the unit number within cluster i of stratum h, with a total of m_{hi} units
- $p = 1, 2, \ldots, P$ is the analysis variable number, with a total of P variables
- $n = \sum_{h=1}^{H} \sum_{i=1}^{n_h} m_{hi}$ is the total number of observations in the sample
- w_{hij} denotes the sampling weight for observation j in cluster i of stratum h
- $\mathbf{y}_{hij} = (y_{hij}^{(1)}, y_{hij}^{(2)}, \ldots, y_{hij}^{(P)})$ are the observed values of the analysis variables for observation j in cluster i of stratum h, including both the values of numerical variables and the values of indicator variables for levels of categorical variables.

For a categorical variable C, let I denote the number of levels of C, and denote the level values as $c_1, c_2, \ldots, c_I$. Then there are I indicator variables associated with these levels. That is, for level $C=c_k$ ($k = 1, 2, \ldots, I$), a $y^{(q)}\left(q \in \{1, 2, \ldots, P\}\right)$ contains the values of the indicator variable for the category $C=c_k$, with the value of observation j in cluster i of stratum h:

$$y_{hij}^{(q)} = I_{\{C=c_k\}}(h,i,j) = \begin{cases} 1 & \text{if } C_{hij} = c_k \\ 0 & \text{otherwise} \end{cases}$$

Therefore, the total number of analysis variables, P, is the total number of numerical variables plus the total number of levels of all categorical variables.

Also, f_h denotes the sampling rate for stratum h. You can use the TOTAL= option or the RATE= option to input population totals or sampling rates. See the section "Specification of Population Totals and Sampling Rates" for details. If you input stratum totals, PROC SURVEYMEANS computes f_h as the ratio of the stratum sample size to the stratum total. If you input stratum sampling rates, PROC SURVEYMEANS uses these values directly for f_h. If you do not specify the TOTAL= option or the RATE= option, then the procedure assumes that the stratum sampling rates f_h are negligible, and a finite population correction is not used when computing variances.

This notation is also applicable to other sample designs. For example, for a sample design without stratification, you can let H=1; for a sample design without clusters, you can let m_{hi}=1 for every h and i.

Mean

When you specify the keyword MEAN, the procedure computes the estimate of the mean (mean per element) from the survey data. Also, the procedure computes the mean by default if you do not specify any statistic-keywords in the PROC SURVEYMEANS statement.

PROC SURVEYMEANS computes the estimate of the mean as

$$\hat{\bar{Y}} = \left(\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\, y_{hij}\right) / w_{...}$$

where

$$w_{...} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}$$

is the sum of the weights over all observations in the sample.

Variance and Standard Error of the Mean

When you specify the keyword STDERR, the procedure computes the standard error of the mean. Also, the procedure computes the standard error by default if you specify the keyword MEAN, or if you do not specify any statistic-keywords in the PROC SURVEYMEANS statement. The keyword VAR requests the variance of the mean.

PROC SURVEYMEANS uses the Taylor series expansion theory to estimate the variance of the mean $\hat{\bar{Y}}$. The procedure computes the estimated variance as

$$\hat{V}(\hat{\bar{Y}}) = \sum_{h=1}^{H} \hat{V}_h(\hat{\bar{Y}})$$

where if n_h>1,

$$\widehat{V_h}(\widehat{\bar{Y}}) = \frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}(e_{hi\cdot}-\bar{e}_{h\cdot\cdot})^2$$

$$e_{hi\cdot} = \left(\sum_{j=1}^{m_{hi}} w_{hij}\,(y_{hij}-\widehat{\bar{Y}})\right)/w_{\cdots}$$

$$\bar{e}_{h\cdot\cdot} = \left(\sum_{i=1}^{n_h} e_{hi\cdot}\right)/n_h$$

and if n_h=1,

$$\widehat{V_h}(\widehat{\bar{Y}}) = \begin{cases} \text{missing} & \text{if } n_{h'}=1 \text{ for } h'=1,2,\ldots,H \\ 0 & \text{if } n_{h'}>1 \text{ for some } 1<h'<H \end{cases}$$

The standard error of the mean is the square root of the estimated variance.

$$\text{StdErr}(\widehat{\bar{Y}}) = \sqrt{\widehat{V}(\widehat{\bar{Y}})}$$

Ratio

When you use a RATIO statement, the procedure produces statistics requested by the statistics-keywords in the PROC SURVEYMEANS statement.

Suppose that you want to calculate the ratio of variable *Y* over variable *X*. Let x_{hij} be the value of variable *X* for the *j*th member in cluster *i* in the *h*th stratum.

The ratio of *Y* over *X* is

$$\widehat{R} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\,y_{hij}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\,x_{hij}}$$

PROC SURVEYMEANS uses the Taylor series expansion method to estimate the variance of the ratio $\widehat{R}$ as

$$\widehat{V}(\widehat{R}) = \sum_{h=1}^{H}\widehat{V_h}(\widehat{R})$$

where if n_h>1,

$$\widehat{V_h}(\widehat{R}) = \frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}(g_{hi\cdot}-\bar{g}_{h\cdot\cdot})^2$$

$$g_{hi\cdot} = \frac{\sum_{j=1}^{m_{hi}} w_{hij}\,(y_{hij}-x_{hij}\widehat{R})}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\,x_{hij}}$$

$$\bar{g}_{h\cdot\cdot} = \left(\sum_{i=1}^{n_h} g_{hi\cdot}\right)/n_h$$

and if n_h=1,

$$\widehat{V_h}(\widehat{R}) = \begin{cases} \text{missing} & \text{if } n_{h'}=1 \text{ for } h'=1,2,\ldots,H \\ 0 & \text{if } n_{h'}>1 \text{ for some } 1<h'<H \end{cases}$$

The standard error of the ratio is the square root of the estimated variance.

$$\text{StdErr}(\widehat{R}) = \sqrt{\widehat{V}(\widehat{R})}$$

t Test for the Mean

If you specify the keyword T, PROC SURVEYMEANS computes the *t*-value for testing that the population mean equals zero, $H_0 : \bar{Y}=0$. The test statistic equals

$$t(\widehat{\bar{Y}}) = \widehat{\bar{Y}} \,/\, \text{StdErr}(\widehat{\bar{Y}})$$

The two-sided *p*-value for this test is

$$\text{Prob}(\,|T| > |t(\widehat{\bar{Y}})|\,)$$

where *T* is a random variable with the *t* distribution with *df* degrees of freedom.

PROC SURVEYMEANS calculates the degrees of freedom for the *t* test as the number of clusters minus the number of strata. If there are no clusters, then *df* equals the number of observations minus the number of strata. If the design is not stratified, then *df* equals the number of clusters minus one. The procedure displays *df* for the *t* test if you specify the keyword DF in the PROC SURVEYMEANS statement.

If missing values or missing weights are present in your data, the number of strata, the number of observations, and the number of clusters are counted based on the observations in non-empty strata. See the section "Missing Values" for details. For degrees of freedom in domain analysis, see the section "Domain Statistics."

Confidence Limits for the Mean

If you specify the keyword CLM, the procedure computes two-sided confidence limits for the mean. Also, the procedure includes the confidence limits by default if you do not specify any statistic-keywords in the PROC SURVEYMEANS statement.

The confidence coefficient is determined by the value of the ALPHA= option, which by default equals 0.05 and produces 95% confidence limits. The confidence limits are computed as

$$\widehat{\bar{Y}} \pm \mathrm{StdErr}(\widehat{\bar{Y}}) \cdot t_{df,\,\alpha/2}$$

where $\widehat{\bar{Y}}$ is the estimate of the mean, $\mathrm{StdErr}(\widehat{\bar{Y}})$ is the standard error of the mean, and $t_{df,\,\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the *t* distribution with *df* calculated as described in the section "t Test for the Mean."

If you specify the keyword UCLM, the procedure computes the one-sided upper $100(1-\alpha)$ confidence limit for the mean:

$$\widehat{\bar{Y}} + \mathrm{StdErr}(\widehat{\bar{Y}}) \cdot t_{df,\,\alpha}$$

If you specify the keyword LCLM, the procedure computes the one-sided lower $100(1-\alpha)$ confidence limit for the mean:

$$\widehat{\bar{Y}} - \mathrm{StdErr}(\widehat{\bar{Y}}) \cdot t_{df,\,\alpha}$$

Coefficient of Variation

If you specify the keyword CV, PROC SURVEYMEANS computes the coefficient of variation, which is the ratio of the standard error of the mean to the estimated mean.

$$cv(\bar{Y}) = \mathrm{StdErr}(\widehat{\bar{Y}}) \,/\, \widehat{\bar{Y}}$$

If you specify the keyword CVSUM, PROC SURVEYMEANS computes the coefficient of variation for the estimated total, which is the ratio of the standard deviation of the sum to the estimated total.

$$cv(Y) = \mathrm{Std}(\widehat{Y}) \,/\, \widehat{Y}$$

Proportions

If you specify the keyword MEAN for a categorical variable, PROC SURVEYMEANS estimates the proportion, or relative frequency, for each level of the categorical variable. If you do not specify any statistic-keywords in the PROC SURVEYMEANS statement, the procedure estimates the proportions for levels of the categorical variables, together with their standard errors and confidence limits.

The procedure estimates the proportion in level c_k for variable *C* as

$$\hat{p} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\, y_{hij}^{(q)}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}}$$

where $y_{hij}^{(q)}$ is the value of the indicator function for level $C=c_k$, defined in the section "Definition and Notation," and $y_{hij}^{(q)}$ equals 1 if the observed value of variable *C* equals c_k, and $y_{hij}^{(q)}$ equals 0 otherwise. Since the proportion estimator is actually an estimator of the mean for an indicator variable, the procedure computes its variance and standard error according to the method outlined in the section "Variance and Standard Error of the Mean." Similarly, the procedure computes confidence limits for proportions as described in the section "Confidence Limits for the Mean."

Total

If you specify the keyword SUM, the procedure computes the estimate of the population total from the survey data. The estimate of the total is the weighted sum over the sample.

$$\widehat{Y} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\, y_{hij}$$

For a categorical variable level, $\widehat{Y}$ estimates its total frequency in the population.

Variance and Standard Deviation of the Total

When you specify the keyword STD or the keyword SUM, the procedure estimates the standard deviation of the total. The keyword VARSUM requests the variance of the total.

PROC SURVEYMEANS estimates the variance of the total as

$$\widehat{V}(\widehat{Y}) = \sum_{h=1}^{H} \widehat{V_h}(\widehat{Y})$$

where if $n_h>1$,

$$\widehat{V_h}(\widehat{Y}) = \frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}(y_{hi\cdot}-\bar{y}_{h\cdot\cdot})^2$$

$$y_{hi\cdot} = \sum_{j=1}^{m_{hi}} w_{hij}\, y_{hij}$$

$$\bar{y}_{h\cdot\cdot} = \left(\sum_{i=1}^{n_h} y_{hi\cdot}\right) / n_h$$

and if $n_h=1$,

$$\widehat{V_h}(\widehat{Y}) = \begin{cases} \text{missing} & \text{if } n_{h'}=1 \text{ for } h'=1,2,\ldots,H \\ 0 & \text{if } n_{h'}>1 \text{ for some } 1<h'<H \end{cases}$$

The standard deviation of the total equals

$$\mathrm{Std}(\widehat{Y}) = \sqrt{\widehat{V}(\widehat{Y})}$$

Confidence Limits of a Total

If you specify the keyword CLSUM, the procedure computes confidence limits for the total. The confidence coefficient is determined by the value of the ALPHA= option, which by default equals 0.05 and produces 95% confidence limits. The confidence limits are computed as

$$\widehat{Y} \pm \mathrm{Std}(\widehat{Y}) \cdot t_{df,\,\alpha/2}$$

where $\widehat{Y}$ is the estimate of the total, $\mathrm{Std}(\widehat{Y})$ is the estimated standard deviation, and $t_{df,\,\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the *t* distribution with *df* calculated as described in the section "t Test for the Mean."

If you specify the keyword UCLSUM, the procedure computes the one-sided upper $100(1-\alpha)$ confidence limit for the sum:

$$\widehat{Y} \;+\; \mathrm{Std}(\widehat{Y}) \cdot t_{df,\,\alpha}$$

If you specify the keyword LCLSUM, the procedure computes the one-sided lower $100(1-\alpha)$ confidence limit for the sum:

$$\widehat{Y} \;-\; \mathrm{Std}(\widehat{Y}) \cdot t_{df,\,\alpha}$$

Domain Statistics

When you use a DOMAIN statement to request a domain analysis, the procedure computes the requested statistics for each domain.

For a domain *D*, let I_D be the corresponding indicator variable:

$$I_D(h,i,j) \;=\; \begin{cases} 1 & \text{if observation } (h,i,j) \text{ belongs to } D \\ 0 & \text{otherwise} \end{cases}$$

Let

$$z_{hij} \;=\; y_{hij} I_D(h,i,j) \;=\; \begin{cases} y_{hij} & \text{if observation } (h,i,j) \text{ belongs to } D \\ 0 & \text{otherwise} \end{cases}$$

The requested statistics for variable *y* in domain *D* are computed based on the values of *z*.

Domain Mean The estimated mean of *y* in the domain *D* is

$$\widehat{\bar{Y}}_D \;=\; \left(\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} v_{hij}\, z_{hij} \right) / \, v_{...}$$

where

$$v_{hij} \;=\; w_{hij} I_D(h,i,j) \;=\; \begin{cases} w_{hij} & \text{if observation } (h,i,j) \text{ belongs to } D \\ 0 & \text{otherwise} \end{cases}$$

$$v_{...} \;=\; \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} v_{hij}$$

The variance of $\widehat{\bar{Y}}_D$ is estimated by

$$\widehat{V}(\widehat{\bar{Y}}_D) \;=\; \sum_{h=1}^{H} \widehat{V}_h(\widehat{\bar{Y}}_D)$$

where if n_h>1,

$$\widehat{V}_h(\widehat{\bar{Y}}_D) \;=\; \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h} (r_{hi\cdot} - \bar{r}_{h\cdot\cdot})^2$$

$$r_{hi\cdot} \;=\; \left(\sum_{j=1}^{m_{hi}} v_{hij}\, (z_{hij} - \widehat{\bar{Y}}_D) \right) / \, v_{...}$$

$$\bar{r}_{h\cdot\cdot} \;=\; \left(\sum_{i=1}^{n_h} r_{hi\cdot} \right) / \, n_h$$

and if n_h=1,

$$\widehat{V}_h(\widehat{\bar{Y}}_D) \;=\; \begin{cases} \text{missing} & \text{if } n_{h'} = 1 \text{ for } h' = 1,2,\ldots,H \\ 0 & \text{if } n_{h'} > 1 \text{ for some } 1 < h' < H \end{cases}$$

Domain Total The estimated total in domain *D* is

$$\widehat{Y}_D \;=\; \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} v_{hij}\, z_{hij}$$

and its estimated variance is

$$\widehat{V}(\widehat{Y}_D) = \sum_{h=1}^{H} \widehat{V}_h(\widehat{Y}_D)$$

where if n_h>1,

$$\widehat{V}_h(\widehat{Y}_D) \;=\; \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h} (z_{hi\cdot} - \bar{z}_{h\cdot\cdot})^2$$

$$z_{hi\cdot} \;=\; \sum_{j=1}^{m_{hi}} v_{hij}\, z_{hij}$$

$$\bar{z}_{h\cdot\cdot} \;=\; \left(\sum_{i=1}^{n_h} z_{hi\cdot} \right) / \, n_h$$

and if n_h=1,

$$\widehat{V}_h(\widehat{Y}_D) \;=\; \begin{cases} \text{missing} & \text{if } n_{h'} = 1 \text{ for } h' = 1,2,\ldots,H \\ 0 & \text{if } n_{h'} > 1 \text{ for some } 1 < h' < H \end{cases}$$

Degrees of Freedom For domain analysis, PROC SURVEYMEANS computes the degrees of freedom for *t* tests as the number of clusters in the non-empty strata minus the number of non-empty strata. When the sample design has no clusters, the degrees of freedom equals the number of observations in non-empty strata minus the number of non-empty strata. As discussed in the section "Missing Values," missing values and missing weights can result in empty strata. In domain analysis, an empty stratum can also occur when the stratum contains

no observations in the specified domain. If no observations in a whole stratum belong to a domain, then this stratum is called an empty stratum for that domain.

For example,

```
data new;
   input str clu y w d;
   datalines;
1 1 . 40 9
1 2 2  . 9
1 3 . 25 9
2 4 5 20 9
2 5 8 15 9
3 6 5 30 7
3 7 9 89 7
3 8 6 23 7
;
proc surveymeans df nobs nclu nmiss;
   strata str;
   cluster clu;
   var y;
   weight w;
   domain d;
run;
```

Table 70.2: Calculations of *df* for Y

	Domain D=7	Domain D=9
Non Empty Strata	STR=3	STR=2
Clusters Used in the Analysis	CLU=6, CLU=7, and CLU=8	CLU=4 and CLU=5
df	3-1=2	2-1=1

Although there are three strata in the data set, STR=1 is an empty stratum for variable Y because of missing values and missing weights. In addition, no observations in stratum STR=3 belong to domain D=9. Therefore, STR=3 becomes an empty stratum as well for variable Y in domain D=9. As a result, the total number of non-empty strata for domain D=9 is one. The non-empty stratum for domain D=9 and variable Y is stratum STR=2. The total number of clusters for domain D=9 is two, which belong to stratum STR=2. Thus, for variable Y in domain D=9, the degrees of freedom for the *t* tests of the domain mean is *df*=2-1=1. Similarly, for domain D=7, strata STR=1 and STR=2 are both empty strata, so the total number of strata is one (STR=3), and the total number of clusters is three (CLU=6, CLU=7, and CLU=8). Table 70.2 illustrates how domains affect the total number of clusters and total number of strata in the *df* calculation. Figure 70.8 shows the *df* computed by the procedure.

The SURVEYMEANS Procedure

Domain Analysis: d

d	Variable	N	N Miss	Clusters	DF
7	y	3	0	3	6
9	y	2	2	2	4

* 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

* 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

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

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

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