CS-BIGS 6(1):36-49

http://www.csbigs.fr

# Two-dimensional sampling in practice

**Hélène Juillard**
*Ined (Institut national d'études démographiques), France*

*This article explains the principles of the two-stage sampling design and presents the less known cross-classified sampling design. One purpose of the article is to allow the reader to differentiate between the two survey designs and put in practice the sampling and estimation steps. Respective variance estimators for these two designs are calculated in simple cases, and analogies with one-way and two-way ANOVA are proposed. The comparison is motivated by the ELFE french survey, and selections and estimations are illustrated using the softwares R, SAS and Stata.*

Keywords *: ANOVA, survey procedures in R / SAS / Stata, population observed bi-dimensionally, two-stage sampling, variance estimation.*

## 1. Introduction

Our population of interest is observed bi-dimensionally and can be represented by a rectangular array. In Figure 1, we illustrate the cross product of a population of rows and of a population of columns.

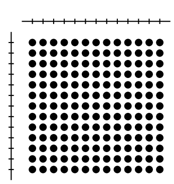

**Figure 1:** Population observed bi-dimensionally

Sampling in a population observed bi-dimensionally is discussed in the literature in different contexts: spatial sampling with the longitude and the latitude as the dimensions, as well as plane sampling or sampling in space and time in Vos (1964). The use of rows and columns in lattice sampling is presented in Bellhouse (1981) or Ohlsson (1996). Sampling of outlets and items for the consumer price index is presented in Dalén and Ohlsson (1995). A sampling of maternities and days is also used for the ELFE (Etude Longitudinale Française depuis l'Enfance) french cohort of infants.

Various sampling designs are possible in a population observed bi-dimensionally. The sample can be drawn directly with one phase of selection only (as shown in Figure 2), or with several steps of selections. For example, a standard two-stage sampling design can be used. This consists in drawing a sample of primary units, and then a second stage sample inside each primary unit independently. Figure 3 illustrates a case where rows are used as primary units: 4 rows are selected, and 3 columns are then drawn inside each selected row.

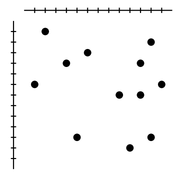

**Figure 2:** Direct sampling in a population observed bi-dimensionally

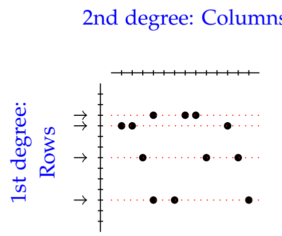


**Figure 3:** Two-stage sampling in a population observed bi-dimensionally with rows as primary units

A cross-classified sampling design (CCS) can also be used, which proceeds as follows: two samples are drawn independently, and then crossed. In Figure 4, a sample of 4 rows and a sample of 3 columns are selected, which results in a final sample of 12 units row × column.

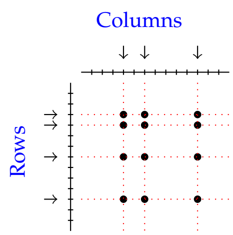


**Figure 4:** Cross-classified sampling

For the two-stage and the cross-classified designs, we distinguish two steps of sampling: one on rows and one on columns. Nevertheless, the CCS design can not be regarded as a classic two-stage design. A classic two-stage design requires two assumptions: independence between the drawings made at each stage, also called the invariance property (Särndal et al., 1992); independence between the various drawings at the second stage, conditionally on the first stage sample. For a CCS design, the invariance property is verified (independence between the sample of rows and the sample of columns), but the independence property is not (a same sample of columns is used for each row).

If the two-stage sampling design is well known, the CCS design presents a limited literature, recently completed by Skinner (2015) and Juillard et al. (2016). In practice, it is specifically used in the Consumer Price Index designs in different countries like the United States (Wilkerson, 1957) and Sweden (Dalén and Ohlsson, 1995). One purpose of this article is to allow the reader to differentiate between these two sampling designs, and to put in practice the sampling and estimation steps. In practice, softwares like R, SAS or Stata propose sampling and estimation procedures for two-stage sampling, but to the best of our knowledge there is no such offer for the CCS design. This case study aims at illustrating the error committed by users, when treating the CCS design as a two-stage sampling design for variance computation and variance estimation. A R program which enables to perform variance estimation for a CCS design is available as supplementary material.

The comparison between two-stage sampling and CCS is motivated by the ELFE survey presented in Section 2 with the data used for this case study. For these two designs, the total and ratio parameters are studied and corresponding variances as well as variance estimators are computed in a simple case. Analogies with one-way and two-way ANOVA are proposed, which enables to interpret the variance formulas in terms of column effect or row effect. In Section 3, we focus on the two-stage sampling design and in Section 4, we focus on the CCS design. We will compare the softwares advantages (R, SAS, Stata) in terms of selection procedures and variance estimation when estimating totals and ratios. The various estimators will be progressively illustrated in this article. A comparison between the different methods of estimation for the two designs through simulations is proposed in Section 5.

## 2. ELFE survey, data and softwares

The ELFE[1] french cohort consists of more than 18,000 children whose parents consented to their inclusion. In each of the 320 selected maternity units, targeted babies born during 25 days (during four specific periods representing each of the four seasons) in 2011 were selected. In the ELFE survey, spatial (metropolitan France) and temporal (year 2011) variabilities was sought. In practice, logistical and administrative reasons oriented the sample design: a direct sampling (as illustrated in Figure 2) or a two-stage sampling design (as illustrated in Figure 3) could not be used. A CCS was implemented, crossing independently a sample of maternities and a sample of days. Stratified simple random sampling was used for the two populations, but in our study, we will consider a simple random sampling for the two designs. Owing to its two selection steps, the CCS design may be considered by data users as a two-stage sampling design, leading to erroneous variance estimation. This article aims at differentiating these two sampling designs, and at quantifying the bias in variance induced by such approximation of the survey design.

The dataset delivered with this article represents the ELFE population with $N_M = 544$ maternities in the population $U_M$ and $N_D = 365$ days in the population $U_D$ in 2011. Given the confidentiality issues, the interest variables in the dataset are count variables simulated taking into account different maternity and day effects. So as to mimic the variables in the ELFE survey, we consider the *Number of infants with a mother followed by a midwife* for the variable $Y_{ik}$ and the *Number of infants born by caesarean* for $Z_{ik}$ where $i$ denotes the index for the maternity and $k$ the index for the day. In this article, we will focus on the estimation of total and ratio parameters and the variable $X_{ik}$ in the dataset, that will be used as the denominator for the ratio, can be considered as the *Number of births*. The construction of this count variables is detailed in Appendix 6.1.

The code is provided in order to replicate all results obtained in this article. Three softwares are used and compared: R 3.2.2 (R Core Team, 2015), SAS 9.4 (SAS Institute Inc., 2015), Stata 13.1 (StataCorp., 2013). R is available from Comprehensive R Archive Network (CRAN) at http://CRAN.R-project.org/.

## 3. Two-stage sampling: selection and estimation

We begin by describing the basic principles of two-stage sampling. Assume that we are interested in some population $U_M = \{u_1, \ldots, u_i, \ldots, u_{N_M}\}$ of non-overlapping Primary Sampling Units (PSUs), where each PSU $u_i$ is itself a population of Secondary Sampling Units (SSUs) of size $N_i$. A sample $S_M$ of size $n_M$ is selected in $U_M$ by means of some sampling design $p_M(\cdot)$. Inside each $u_i \in S_M$, a second stage sample $S_i$ of size $n_i$ is then selected according to some sampling design $p_{iD}(\cdot|S_M)$. The final sample of SSUs is $S = \bigcup_{u_i \in S_M} S_i$.

A two-stage sampling design is usually required to match the following assumptions:

**H1** Invariance: the design $p_{iD}(\cdot|S_M)$ used in the second stage for a PSU $u_i$ does not depend on the first-stage sample $S_M$ selected, that is

$$\forall u_i \in U_M, \quad p_{iD}\left(.|S_M\right) = p_{iD}\left(.\right).$$

**H2** Independence: conditionally on $S_M$, the sub-sampling inside the selected PSUs is independent from one PSU to another. That is,

$$Pr\left(\bigcup_{u_i \in S_M} S_i|S_M\right) = \prod_{u_i \in S_M} Pr\left(S_i|S_M\right).$$

[1] http://www.elfe-france.fr/index.php/en/

### 3.1. *Selecting a two-stage sample*

In this part, the possibilities to draw two-stage samples using the softwares R, SAS and Stata are scanned. In our case study, a SI (simple random) sampling is drawn in $U_M$ and the SI sampling is also used in each $u_i \in U_M$ (which we denote {SI,SI}); in order to mimic the ELFE sample size, the same number $n_D = 25$ of SSUs is drawn inside each of the $n_M = 320$ selected PSUs.

**R implementation** The function *mstage* of the sampling package (Tillé and Matei, 2015) in R can be used to select a two-stage sample in a single step (see the frame Code 1). With the argument *stage*, four methods of selection can be used but it has to be the same for the two stages: simple random sampling without replacement or with replacement, Poisson sampling or systematic sampling. The option *pik* has to be applied in the case of unequal probabilities of selection. The argument *size* used indicates the sample size of PSUs, and the vector of sample sizes of SSUs.

```
library(sampling)
tableR=read.csv2(".../Data2stCCS.csv")
n_m=320; n_d=25; N_m=544; N_d=365; N=N_m*N_d

m=mstage(tableR,stage=list("cluster","cluster"),
     varnames=list("ID_i","ID_k"),size=list(n_m,c
     (rep(n_d,n_m))), method=c("srswor","srswor")
     )

ech=getdata(tableR,m)[[2]]
```

**Code 1:** An R code to select a two-stage sample in a population observed bi-dimensionally

**SAS implementation** The SAS software proposes to call two procedures *SURVEYSELECT* as proposed in the frame Code 2. In order to identify the PSUs, the first procedure uses the *cluster* statement and the second the *strata* statement. The *strata* statement can also be applied at both stages and a lot of different methods of selection are available (simple random sampling with or without replacement, Bernoulli sampling, sampling with probabilities proportional to size, with sequential or systematic selection, . . . ).

```
proc import datafile = ".../Data2stCCS.csv"
out=pop dbms=csv replace;DELIMITER=";" ; run;

proc SURVEYSELECT data=pop method=srs n=320 seed
     =1357 out=ech1;
   cluster ID_i;
run;

proc SURVEYSELECT data=ech1 method=srs n=25 seed
     =7548 out=ech;
   strata ID_i;
run;
```

**Code 2:** A SAS code to select a two-stage sample in a population observed bi-dimensionally

**Stata implementation** The software Stata proposes the command *sample* (*bsample*, respectively) to draw a random sample without replacement (with replacement, respectively). The command *sample* can be used with the option *by* followed by the name of the stratum. In this case the same number or the same percentage of units is drawn inside each stratum. In the frame Code 3, a table 'ech1' containing only one row by PSU is created and a first SI sample of size 320 is selected. The command *merge* enables to create the sampling base for the second step of selection. A SI sample of 25 units is drawn in each selected PSU using *by id_i: sample 25, count*.

```
. clear
. insheet using /.../Data2stCCS.csv,delimiter(;)
. save POP, replace
. contract id_i
. sample 320, count

. sort id_i
. keep id_i
. save /.../ech1.dta, replace
. clear
. use POP
. sort id_i
. merge m:1 id_i using /.../ech1.dta
. drop if _merge != 3

. sort id_i
. by id_i: sample 25, count
. count
```

**Code 3:** A Stata code to select a two-stage sample in a population observed bi-dimensionally

The same steps as in Stata could also be used with R and SAS. This would enable to make use at any one-stage sampling procedure available in each software.

### 3.2. Estimating a total

We consider a study variable $Y$ taking the value $Y_{ik}$ for the PSU $u_i$ and the SSU $k$. We are interested in estimating the total

$$t_Y = \sum_{u_i \in U_M} \sum_{k \in u_i} Y_{ik}.$$

In the particular case of SI sampling in $U_M$ and SI sampling inside each $u_i \in S_M$, the expansion estimator

$$\hat{t}_Y = \frac{N_M}{n_M} \sum_{u_i \in S_M} \frac{N_i}{n_i} \sum_{k \in S_i} Y_{ik}$$

is unbiased for $t_Y$ (Särndal et al., 1992).

### 3.3. Calculating the variance

Under the invariance and the independence assumptions, the variance of $\hat{t}_Y$ is obtained by conditioning on the first stage sample $S_M$. This leads to

$$\mathbf{V}_{2d}\left(\hat{t}_Y\right) = \mathbf{V}_{PSU}\left(\hat{t}_Y\right) + \mathbf{V}_{SSU}\left(\hat{t}_Y\right).$$

In case of {SI,SI}, we obtain

$$\mathbf{V}_{PSU}\left(\hat{t}_Y\right) = N_M^2 \left(\frac{1}{n_M} - \frac{1}{N_M}\right) S^2_{Y_{\circ\bullet}}, \quad (1)$$

$$\mathbf{V}_{SSU}\left(\hat{t}_Y\right) = \frac{N_M}{n_M} \sum_{u_i \in U_M} N_i^2 \left(\frac{1}{n_i} - \frac{1}{N_i}\right) S^2_{Y_{i\circ}}, \quad (2)$$

with

$$S^2_{Y_{\circ\bullet}} = \frac{1}{N_M - 1} \sum_{u_i \in U_M} \left(Y_{i\bullet} - \frac{1}{N_M} \sum_{u_j \in U_M} Y_{j\bullet}\right)^2,$$

$$S^2_{Y_{i\circ}} = \frac{1}{N_i - 1} \sum_{k \in u_i} \left(Y_{ik} - \frac{1}{N_i} \sum_{l \in u_i} Y_{il}\right)^2.$$

In the particular case where two-stage sampling is used inside a product population $U_M \times U_D$ (as illustrated in Figure 3), all the PSUs $u_i$ (with associated size $N_i$) in the above formulas can be replaced by a same notation $U_D$ (with associated size $N_D$) for all $i \in U_M$. In this case, if the same number of SSUs is drawn inside each selected PSU, we may note $n_i = n_D$ for any $i \in S_M$.

An analogy can be made between the two-stage variance decomposition and the analysis of variance (ANOVA) which uses the partitioning of sums of squared deviations. For one-way ANOVA, the total sum of squares $SS_T = \sum_{u_i \in U_M} \sum_{k \in u_i} (Y_{ik} - \bar{Y}_{\bullet\bullet})^2$ may be written as

$$SS_T = SS_M + SS_E$$

where $SS_M$ is the explained sum of squares (a.k.a. the sum of squares between classes) and $SS_E$ denotes the residual sum of squares (a.k.a. sum of squares within classes), see Appendix 6.3 for details. For example, in our case study, the variable *Number of infants born by caesarean* ($Z_{ik}$) presents a smaller $SS_M$ than the variable *Number of births* ($X_{ik}$).

We consider the {SI,SI} sampling case, and assume for simplicity that all the PSUs are of the same size $N_i = N_D$, and that the same sample size $n_i = n_D$ is used inside each selected PSU. In this case, we have

$$SS_M = \frac{N_M - 1}{N_D} S^2_{Y_{\circ\bullet}},$$

$$SS_E = (N_D - 1) \sum_{u_i \in U_M} S^2_{Y_{i\circ}}.$$

The variance in (1) due to the selection of PSUs may be rewritten as

$$\mathbf{V}_{PSU}\left(\hat{t}_Y\right) = N_M^2 \left(\frac{1}{n_M} - \frac{1}{N_M}\right) \frac{N_D}{N_M - 1} SS_M,$$

and depends on the explained sum of squares $SS_M$. The variable $X_{ik}$ will present a more important part of first-stage variance than the variable $Z_{ik}$. The variance in (2) due to the selection of SSUs may be rewritten as

$$\mathbf{V}_{SSU}\left(\hat{t}_Y\right) = \frac{N_M}{n_M} N_D^2 \left(\frac{1}{n_D} - \frac{1}{N_D}\right) \frac{1}{N_D - 1} SS_E$$

and depends on the residual sum of squares $SS_E$.

### 3.4. *Estimating the variance*

An unbiased variance estimator of $\hat{t}_Y$ can be written as

$$\hat{\mathbf{V}}_{2d}\left(\hat{t}_Y\right) = \hat{\mathbf{V}}_{2d,a}\left(\hat{t}_Y\right) + \hat{\mathbf{V}}_{2d,b}\left(\hat{t}_Y\right) \quad (3)$$

where

$$\hat{\mathbf{V}}_{2d,a}\left(\hat{t}_Y\right) = N_M^2\left(\frac{1}{n_M} - \frac{1}{N_M}\right) s^2_{\hat{Y}_{\circ\bullet}}, \quad (4)$$

$$\hat{\mathbf{V}}_{2d,b}\left(\hat{t}_Y\right) = \frac{N_M}{n_M}\sum_{u_i \in S_M} N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right) s^2_{Y_{i\circ}}, \quad (5)$$

with

$$s^2_{\hat{Y}_{\circ\bullet}} = \frac{1}{n_M - 1}\sum_{u_i \in S_M}\left(\hat{Y}_{i\bullet} - \frac{1}{n_M}\sum_{u_j \in S_M}\hat{Y}_{j\bullet}\right)^2,$$

$$s^2_{Y_{i\circ}} = \frac{1}{n_i - 1}\sum_{k \in S_i}\left(Y_{ik} - \frac{1}{n_i}\sum_{l \in S_i} Y_{il}\right)^2,$$

and where

$$\hat{Y}_{i\bullet} = \sum_{k \in S_i}\frac{N_i}{n_i} Y_{ik}$$

denotes the Horvitz-Thompson estimator of the sub-total $Y_i$. For an estimation term by term of the variance in formula (1), see the Appendix 6.2.

Using the same one-way ANOVA as in the previous section but calculated on the sample $s_M \times s_D$, the total sum of squares $ss_T$ may be written as

$$ss_T = ss_M + ss_E$$

where each term is defined in Appendix 6.3. The first part of the variance estimator in (4) can be rewritten as

$$\hat{\mathbf{V}}_{2d,a}\left(\hat{t}_Y\right) = N_M^2\left(\frac{1}{n_M} - \frac{1}{N_M}\right)\frac{n_D}{n_M - 1}\; ss_M,$$

and depends on the explained sum of squares $ss_M$. The second part in (5) can be rewritten as

$$\hat{\mathbf{V}}_{2d,b}\left(\hat{t}_Y\right) = \frac{N_M}{n_M} N_D^2\left(\frac{1}{n_D} - \frac{1}{N_D}\right)\frac{1}{n_D - 1}\; ss_E$$

and depends on the explained sum of squares $ss_E$. Note that the term $\hat{\mathbf{V}}_{2d,a}\left(\hat{t}_Y\right)$ is occasionally considered as a simplified variance estimator of $\hat{\mathbf{V}}_{2d}\left(\hat{t}_Y\right)$. The underestimation is seen as negligible when the first-stage inclusion probabilities $n_M/N_M$ are small (Särndal et al., 1992).

### 3.5. *Estimation in practice*

In this Section, we propose to study the estimation of a more complex parameter using different procedures from the R, SAS or Stata softwares. A ratio $R = t_Y/t_X$ can be easily estimated by $\hat{R} = \hat{t}_Y/\hat{t}_X$ using a plug-in principle. To estimate the variance, a linearization method can be used (Deville, 1999). The estimated linearized variable is then plugged into the formula (3).
From a particular selected sample (using variable *Dummy_2d* in the dataset, which takes the value 1 if the unit is selected in the {SI,SI} sample and 0 otherwise), the estimated ratios $\hat{t}_Y/\hat{t}_X$ and $\hat{t}_Z/\hat{t}_X$ and their estimated variance $\hat{\mathbf{V}}_{2d}$ can be calculated together with the approximation $\hat{\mathbf{V}}_{2d,a}$.

**R implementation** The functions *svydesign* and *twophase* of the R package survey (Lumley, 2014) can be used to describe the two-stage sample. Only the first one is illustrated in the frame Code 4. Note that other packages are available to estimate the sampling variance. In the argument *id*, the vector of PSU IDs has to be entered, followed by the vector of SSU IDs. The argument *fpc* can be specified as the PSU population size $N_M$ in the form of a vector, followed by the vector of the SSU populations size $N_D$. The appropriate set of weights can be set using the argument *weights*. Note that it is possible to take into account the stratified sampling by using the argument *strata*. The function *svyratio* estimates the ratio and its associated standard error. The command *SE(yxratio)^2* displays the estimated variance $\hat{\mathbf{V}}_{2d}\left(\hat{R}\right)$.

```
> tableR=read.csv2("... / Data2stCCS.csv")
> ech=tableR[tableR$Dummy_2d==1,]
> attach(ech)
> library(survey)
> n_m=320; n_d=25; N_m=544; N_d=365
```

```
> infoplan<−svydesign(id=~ID_i+ID_k,fpc=~N_M+N_D
    , weights=(N_m*N_d)/(n_m*n_d), data=ech)
> (yxratio <− svyratio(~Yik+Zik ,~Xik,infoplan))

Ratio estimator: svyratio.survey.design2(~Yik +
    Zik, ~Xik, infoplan)
Ratios=
            Xik
Yik 0.1507968
Zik 0.1510090

SEs=
                Xik
Yik 0.0008873011
Zik 0.0009162289

> SE(yxratio)^2

     Yik/Xik        Zik/Xik
7.873033e−07 8.394754e−07

> confint(yxratio)
#confint(yxratio, level=0.90)

               2.5 %     97.5 %
Yik/Xik 0.1490577 0.1525359
Zik/Xik 0.1492132 0.1528047
```

**Code 4:** R code and results when estimating the ratio and its variance $\hat{\mathbf{V}}_{2d}\left(\hat{R}\right)$

The command *vcov(yxratio)* permits also to display the estimated variance. The function *svytotal(* $\sim Xik + Yik + Zik$ *, infoplan)* can be used to estimate the totals $\hat{t}_X$, $\hat{t}_Y$ and $\hat{t}_Z$ while the function *svymean(* $\sim Xik + Yik + Zik$ *, infoplan)* can be used to estimate the respective means of $X_{ik}$, $Y_{ik}$ and $Z_{ik}$. By default the function *confint* produces a confidence interval of level 0.95 and it can be changed using the option *level*.
Note that $\hat{\mathbf{V}}_{2d,a}\left(\hat{R}\right)$ can also be calculated with R, with a simple modification of the previous procedures (see frame Code 5).

```
> infoplan<−svydesign(id=~ID_i,fpc=~N_M, weights
    =(N_m*N_d)/(n_m*n_d), data=ech)
> yxratio <− svyratio(~Yik+Zik ,~Xik, infoplan)
> SE(yxratio)^2

     Yik/Xik        Zik/Xik
3.377375e−07 3.732472e−07

> confint(yxratio)

               2.5 %     97.5 %
Yik/Xik 0.1496578 0.1519359
Zik/Xik 0.1498116 0.1522064
```

**Code 5:** R code and results when estimating the ratio and its part of variance $\hat{\mathbf{V}}_{2d,a}\left(\hat{R}\right)$

**SAS implementation** The procedure *SURVEYMEANS* is used in the frame Code 6 with the argument *cluster* to indicate the PSU IDs, and *weight* for the set of weights *wik*. The option *strata* is available. This procedure calculates $\hat{R}$ and only the first part $\hat{\mathbf{V}}_{2d,a}\left(\hat{R}\right)$ of the estimated variance $\hat{\mathbf{V}}_{2d}\left(\hat{R}_Y\right)$.

```
proc IMPORT datafile = ".../Data2stCCS.csv"
out  = ech (where= (Dummy_2d=1))
dbms = csv
replace;
DELIMITER=";" ;
run;

data ech; set ech; wik=(544*365)/(320*25); run;

proc SURVEYMEANS data=ech total=544 mean sum var
    varsum missing clm /* alpha=0.10 */;
CLUSTER ID_i ;
/* VAR Xik Yik Zik ; */
RATIO Yik Zik / Xik ;
WEIGHT wik ;
run ;

                    Ratio Analysis
Numerator Denominator Ratio Std Err Var 95% CL
    for Ratio
Yik Xik 0.150797 0.000581 0.000000338 0.149653
    0.151940
Zik Xik 0.151009 0.000611 0.000000373 0.149807
    0.152211
```

**Code 6:** SAS code and results when estimating the ratio and its part of variance $\hat{\mathbf{V}}_{2d,a}\left(\hat{R}\right)$

The default *alpha* option is 0.05. The line of code *VAR Xik Yik Zik ;* can be used to estimate the totals $\hat{t}_X$, $\hat{t}_Y$ and $\hat{t}_Z$ using results of options *sum* and *varsum*. The same command is used to estimate the means with options *mean* and *var*. Note that the second term of $\hat{\mathbf{V}}_{2d}\left(\hat{R}\right)$ can be calculated using a supplementary step (Aragon and Ruiz-Gazen, 2004).

**Stata implementation** The command *svyset* of Stata in the frame Code 7 is used to describe the two-stage sample. In the first place *id_i* stands for the PSU IDs, followed by the vector of weights *wik* which in this application equals $(N_m N_D)/(n_M n_D)$. The argument *fcp* takes into account the PSU population size. After the two vertical bars, the second stage is defined in the same way. The command *svy : ratio* calculates the estimated ratio $\hat{R}$ and its associated standard error which corresponds to the square root of $\hat{\mathbf{V}}_{2d}\left(\hat{R}\right)$.

```
. clear
. insheet using /.../Data2stCCS.csv,delimiter(;)
(14 vars, 198560 obs)
```

```
. save POP, replace
. keep if dummy_2d==1
. gen wik=(544*365)/(25*320)
. svyset id_i [pweight=wik], fpc(n_m) || id_k,
    fpc(n_d)

      pweight: wik
          VCE: linearized
  Single unit: missing
     Strata 1: <one>
         SU 1: id_i
        FPC 1: n_m
     Strata 2: <one>
         SU 2: id_k
        FPC 2: n_d

. svy : ratio (yik/xik) (zik/xik)
* svy : ratio (yik/xik) (zik/xik), level(90) ;
(running ratio on estimation sample)

Survey: Ratio estimation

Number of strata =   1  Number of obs   =   8000
Number of PSUs   = 320  Population size = 198560
                        Design df       =    319

     _ratio_1: yik/xik
     _ratio_2: zik/xik

--------------------------------------------------------
         |            Linearized
         |   Ratio    Std.Err.  [97.5% Conf.Interval]
---------+----------------------------------------------
_ratio_1 | .1507968  .0008873  .1490511  .1525425
_ratio_2 |  .151009  .0009162  .1492064  .1528116
--------------------------------------------------------
```

**Code 7:** Stata code and results when estimating the ratio and its variance $\hat{\mathbf{V}}_{2d}\left(\hat{R}\right)$

To estimate $\hat{t}_X$, $\hat{t}_Y$ and $\hat{t}_Z$, the command *svy : total xik yik zik* can be used. So as to estimate means, we may use the command *svy : mean xik yik zik*. The default *level* option for the confidence interval is 95 %.
Note that the variance estimator $\hat{\mathbf{V}}_{2d,a}\left(\hat{R}\right)$ can also be obtained with Stata in the frame Code 8.

```
. svyset id_i [pweight=wik], fpc(n_m)
. svy : ratio (yik/xik) (zik/xik)

--------------------------------------------------------
         |            Linearized
         |   Ratio    Std.Err.  [97.5% Conf.Interval]
---------+----------------------------------------------
_ratio_1 | .1507968  .0005812  .1496534  .1519402
_ratio_2 |  .151009  .0006109  .149807   .152211
--------------------------------------------------------
```

**Code 8:** Stata code and results when estimating the ratio and its part of variance $\hat{\mathbf{V}}_{2d,a}\left(\hat{R}\right)$

## 4. Cross-classified sampling: selection and estimation

We now consider the cross-classified sampling design. We consider a sampling design $p_M$ in $U_M$, leading to a sample $S_M$ of size $n_M$. We consider a sampling design $p_D$ in $U_D$, leading to a sample $S_D$ of size $n_D$. We assume that the two designs $p_M(\cdot)$ and $p_D(\cdot)$ are independent. This enables to define a sampling design $p(\cdot)$ on the product population $U = U_M \times U_D$ as

$$\begin{aligned} p(s) &= p_M(s_M) \times p_D(s_D) \\ \text{for any } s &= s_M \times s_D \subset U_M \times U_D. \end{aligned}$$

The assumption of independence for a cross-classified sampling design is equivalent to the standard assumption **H1** of invariance between two successive drawings in a two-stage sampling design.

### 4.1. *Selecting a cross-classified sample*

There is no standard procedure to perform CCS in one step, but all possible one-stage sampling procedures can be used to select $S_M$ and $S_D$ independently. The samples are then crossed to obtain the final sample $S_M \times S_D$. In our case study, we are interesting in the crossing of a SI sample of size $n_M = 320$ drawn in $U_M$, and of a SI sample of size $n_D$ drawn in $U_D$. Such design will be denoted as SI $\times$ SI.

**R implementation** A selection of a SI $\times$ SI sample with the software R is presented in the frame Code 9.

```
> tableR=read.csv2(".../Data2stCCS.csv")
> n_m=320; n_d=25; N_m=544; N_d=365
>
> s_m=sample(1:N_m,n_m) ; s_d=sample(1:N_d,n_d)

> Dummy_CCS2 <- rep(0,N); Dummy_CCS2[which(
    tableR$ID_i %in% s_m & tableR$ID_k %in% s_d)
    ] <-1
> echCCS=tableR[Dummy_CCS2==1, ]
```

**Code 9:** An R code to select the CCS sample

**SAS implementation** With SAS, the procedures *SURVEYSELECT* and *merge* can be used to select and cross the two samples (frame Code 10).

```
proc IMPORT datafile = ".../Data2stCCS.csv"
out=pop dbms=csv replace;DELIMITER=";" ; run;

proc freq data=pop;tables ID_i/out=popM;run;
proc SURVEYSELECT data=popM method=srs n=320
     seed=2289 stats out=echM ;
run;
proc freq data=pop;tables ID_k/out=popD;run;
proc SURVEYSELECT data=popD method=srs n=25 seed
     =2368 stats out=echD ;
run;

proc sort data=pop; by ID_i; run;
proc sort data=echM; by ID_i; run;
data echA; merge echM (in=A) pop; by ID_i; if A;
run;
proc sort data=echA; by ID_k; run;
proc sort data=echD; by ID_k; run;
data ech; merge echD (in=A) echA; by ID_k; if A;
/*Dummy_CCS2=1;*/
run;
```

**Code 10:** A SAS code to select the CCS sample

**Stata implementation** Following the same logic, the commands *sample* and *merge* may be used with the Stata software, as illustrated in the frame Code 11.

```
. clear
. insheet using /.../Data2stCCS.csv,delimiter(;)
. save POP, replace
. contract id_i
. sample 320, count
. sort id_i
. keep id_i
. save echM, replace

. clear
. use POP
. contract id_k
. sample 25, count
. sort id_k
. keep id_k
. save echD, replace

. clear
. use POP
. sort id_i
. merge m:1 id_i using echM.dta
. drop if _merge != 3
. drop _merge
. sort id_k
. merge m:1 id_k using echD.dta
. drop if _merge != 3

* gen Dummy_CCS2=1;
```

**Code 11:** A Stata code to select the CCS sample

### 4.2. *Estimating a total*

In the particular case of SI $\times$ SI, the total

$$t_Y = \sum_{i \in U_M} \sum_{k \in U_D} Y_{ik}$$

is unbiasedly estimated by the expansion estimator

$$\hat{t}_Y = \sum_{i \in S_M} \sum_{k \in S_D} \frac{N_M N_D}{n_M n_D} Y_{ik},$$

see Juillard et al. (2016) for details.

### 4.3. *Calculating the variance*

In this Section, the variance of $\hat{t}_Y$ is calculated in the SI $\times$ SI case. The analogy between the decomposition of the SI $\times$ SI variance and the decomposition of a two-way ANOVA was noted in Ohlsson (1996), and is described here. For a two-way ANOVA without replication, the total sum of squares may be written as

$$SS_T = SS_M + SS_D + SS_E \tag{6}$$

where the terms $SS_D$, $SS_M$ and $SS_E$ represent respectively the sum of squares explained by the factor D, the one explained by the factor M and the residual sum of squares. The details are given in Appendix 6.4. In our case study, the variable *Number of infants born by caesarean* presents a large $SS_D$, since caesarean sections are operations which are rarely scheduled during a week-end. On the other hand, the $SS_D$ is small for the variable *Number of infants with a mother followed by a midwife*. Using the different terms of this ANOVA, the variance of $\hat{t}_Y$ can be rewritten as

$$\mathrm{V}_{CCS}\left(\hat{t}_Y\right) = \mathrm{V}_1\left(\hat{t}_Y\right) + \mathrm{V}_2\left(\hat{t}_Y\right) + \mathrm{V}_3\left(\hat{t}_Y\right) \tag{7}$$

where

$$\begin{aligned}
\mathrm{V}_1\left(\hat{t}_Y\right) &= \left(\frac{1}{n_D} - \frac{1}{N_D}\right) \frac{N_D^2 N_M}{N_D - 1} \; SS_D \\
\mathrm{V}_2\left(\hat{t}_Y\right) &= \left(\frac{1}{n_M} - \frac{1}{N_M}\right) \frac{N_M^2 N_D}{N_M - 1} \; SS_M \\
\mathrm{V}_3\left(\hat{t}_Y\right) &= \left(\frac{1}{n_D} - \frac{1}{N_D}\right) \left(\frac{1}{n_M} - \frac{1}{N_M}\right) \\
&\quad \frac{N_D^2}{N_D - 1} \frac{N_M^2}{N_M - 1} \; SS_E.
\end{aligned}$$

We note that the CCS variance is divided into three terms associated respectively to a maternity effect, a day effect and a residual effect.

On the other hand, the two-stage variance was divided into two terms associated to a maternity effect and to a residual effect. The term $SS_M$ is the same in both decompositions, but the term $SS_E$ is obviously different.

### 4.4. *Estimating the variance*

A term by term unbiased estimator of the variance of $\hat{t}_Y$ in formula (7) is presented in Appendix 6.5. This variance estimator simplifies as

$$\hat{\mathbf{V}}_{CCS}\left(\hat{t}_Y\right) = \hat{\mathbf{V}}_D\left(\hat{t}_Y\right) + \hat{\mathbf{V}}_M\left(\hat{t}_Y\right) - \hat{\mathbf{V}}_E\left(\hat{t}_Y\right) \quad (8)$$

where

$$\begin{aligned} \hat{\mathbf{V}}_D\left(\hat{t}_Y\right) &= \left(\frac{1}{n_D} - \frac{1}{N_D}\right)\frac{N_D^2}{n_D - 1}\frac{N_M^2}{n_M}\; ss_D, \\ \hat{\mathbf{V}}_M\left(\hat{t}_Y\right) &= \left(\frac{1}{n_M} - \frac{1}{N_M}\right)\frac{N_M^2}{n_M - 1}\frac{N_D^2}{n_D}\; ss_M, \\ \hat{\mathbf{V}}_E\left(\hat{t}_Y\right) &= \left(\frac{1}{n_M} - \frac{1}{N_M}\right)\left(\frac{1}{n_D} - \frac{1}{N_D}\right) \\ &\quad \frac{N_M^2 N_D^2}{(n_M - 1)(n_D - 1)}\; ss_E, \end{aligned}$$

where the terms come from an ANOVA decomposition on the sample $s_M \times s_D$ as detailed in Appendix 6.4. The variance estimator is divided into three terms: $\hat{\mathbf{V}}_D\left(\hat{t}_Y\right)$ which represents an inter-day effect, $\hat{\mathbf{V}}_M\left(\hat{t}_Y\right)$ which represents an inter-maternity effect, and $\hat{\mathbf{V}}_E\left(\hat{t}_Y\right)$ which represents a residual effect.

### 4.5. *Estimation in practice*

To the best of our knowledge, there are no direct procedures in the softwares R, SAS and Stata to calculate CCS variance estimates. In this paper, we develop R functions to estimate a total and a ratio along with variance estimators. More precisely, from a selected sample (using variable *Dummy_CCS* in the dataset, which takes the value 1 if the unit is selected in the SI $\times$ SI sample and 0 otherwise), the estimated total $\hat{t}_X$ and its estimated variance can be calculated using the R functions *EstTccsSISI* and *EstVARTccsSISI* proposed in the supplementary material. In the frame Code 12, these functions require that you enter the cross-classified sample (matrix of size $n_D \times n_M$), the sample sizes $n_M$ and $n_D$ and the population sizes $N_M$ and $N_D$.

```
> echCCS=tableR[tableR$Dummy_CCS==1,];attach(
      echCCS)
> n_m=320; n_d=25; N_m=544; N_d=365
> echXCCS=matrix(Xik,nrow=n_d)
> EstTccsSISI(ECH=echXCCS,n_m,n_d,N_m,N_d)

[1] 3981426

> EstVARTccsSISI(ECH=echXCCS,n_m,n_d,N_m,N_d)

[1] 307219631
```

**Code 12:** R code and results when estimating the total and its variance $\hat{\mathbf{V}}_{CCS}\left(\hat{t}_Y\right)$

To estimate the ratio $t_Y/t_X$, the function *EstRccsSISI* can be used. The linearized variable for the ratio estimator is then calculated by *LinearizedR*, and is plugged in the function *EstVARTccsSISI* as illustrated in the frame Code 13.

```
> echYCCS=matrix(Yik,nrow=n_d)
> EstRccsSISI(ECHY=echYCCS,ECHX=echXCCS,n_m,n_d,
      N_m,N_d)

[1] 0.1495898

> LinR=LinearizedR(ECHY=echYCCS,ECHX=echXCCS,n_m
      ,n_d,N_m,N_d)
> EstVARTccsSISI(ECH=LinR,n_m,n_d,N_m,N_d)

[1] 1.006684e-06
```

**Code 13:** R code and results when estimating ratio and its variance $\hat{\mathbf{V}}_{CCS}\left(\hat{R}_Y\right)$

## 5. Illustration

A small simulation study is conducted to compare the performance of several variance estimators under a two-stage sampling design and under a CCS design. We also evaluate the performance of various variance estimators. For a two-stage sampling design where the primary units are the maternities and where the number of secondary units $n_D$ is the same inside all the primary units, we calculated the unbiased variance estimator $\hat{\mathbf{V}}_{2d}$ as well its first part $\hat{\mathbf{V}}_{2d,a}$. For the CCS design, the unbiased variance estimator $\hat{\mathbf{V}}_{CCS}$ is calculated as well as $\hat{\mathbf{V}}_{2d}$ and we also calculate the first part $\hat{\mathbf{V}}_{2d,a}$

of $\hat{\mathbf{V}}_{2d}$ in order to examine the error due to using the two-stage variance estimator instead of the cross-classified variance estimator. The two sampling designs and the various variance estimators are summarized in Table 1.

**Table 1:** Variance estimators of two-stage sampling and CCS

| SAMPLING DESIGN | |
|---|---|
| two-stage | cross-classified |
| UNBIASED VARIANCE ESTIMATOR | |
| $\hat{\mathbf{V}}_{2d}$ in (3) | $\hat{\mathbf{V}}_{CCS}$ in (8) |
| APPROXIMATION | |
| $\hat{\mathbf{V}}_{2d,a}$ in (4) | $\hat{\mathbf{V}}_{2d}$ in (3)<br>$\hat{\mathbf{V}}_{2d,a}$ in (4) |

For the two-stage sampling design, the {SI,SI} sampling is used: a sample $S_M$ of $n_M$ maternities is selected and in each selected maternity, a sample $s_D$ of size $n_D$ is selected. For the CCS design, the SI $\times$ SI sampling is used: a sample $S_D$ of $n_D$ days, and a sample $S_M$ of $n_M$ maternities are selected. We used various sample sizes are used, namely $n_M$ or $n_D$ equal to 5, 25 and 320 (the two last sizes corresponding to the true ELFE sample sizes). These two sample selections were respectively repeated $B = 10,000$ times. For CCS and for two-stage sampling, and in each of the $b = 1, \ldots, B$ samples, the estimator $\hat{R}^{(b)}$ of the ratio $R = t_Y/t_X$ is computed. Also, for each cross-classified sample, the unbiased variance estimator $\hat{V}^{(b)}_{CCS}$ and the simplified variance estimators $\hat{V}^{(b)}_{2d}$, $\hat{V}^{(b)}_{2d,a}$ are computed, and for each two-stage sample, the unbiased variance estimator $\hat{V}^{(b)}_{2d}$ and the simplified variance estimator $\hat{V}^{(b)}_{2d,a}$ are computed. For each variance estimator $\hat{V}$, the Monte Carlo Percent Relative Bias (RB), given by

$$\mathrm{RB}_{\mathbf{MC}}(\hat{V}) = 100 \times \frac{B^{-1}\sum_{b=1}^{B}\hat{V}^{(b)} - V}{V}$$

is computed, where the true variance $V$ was approximated through an independent set of $50,000$ simulations.

Results for two ratios are reported in Table 2. In the top part of the table (case 1), we consider the plug-in estimator $\hat{t}_Y/\hat{t}_X$ of the proportion of infants with a mother followed by a midwife. In the bottom part of the table (case 2), we consider the plug-in estimator $\hat{t}_Z/\hat{t}_X$ of the proportion of infants born by caesarean. As expected, the variance estimator $\hat{V}_{CCS}$ is unbiased for the CCS variance, and the variance estimator $\hat{V}_{2d}$ is unbiased for the two-stage sampling variance. For the two-stage sampling, the estimator $\hat{V}_{2d,a}$ gives a good approximation of $\hat{V}_{2d}$ when the sample size $n_M$ is small (5 or 25). But it presents an important underestimation when $n_M$ increases (320), especially when $n_D$ is small (25) : -57 % for both cases. For the CCS, in all cases, the relative biases of $\hat{V}_{2d}$ and $\hat{V}_{2d,a}$ increase when $n_M$ increases or when $n_D$ decreases. The relative bias of $\hat{V}_{2d,a}$ is always greater than the relative bias of $\hat{V}_{2d}$. In case 2, for all samples sizes, the relative biases are larger than in case 1. In this case, the variable *Number of infants born by caesarean* that we use presents an important day variability. The approximation of $\hat{V}_{CCS}$ by $\hat{V}_{2d}$ or $\hat{V}_{2d,a}$, which captures principally maternity effect, is therefore not appropriate. In case 1, the day effect (of $Y_{ik}$) is not as strong as for case 2, and the relative biases are therefore smaller.

**Table 2:** Comparison between variance estimators of the estimated ratio for CCS and two-stage sampling (2d)

| | | | | | | |
|---|---|---|---|---|---|---|
| | $n_M$ | 5 | 25 | 320 | 25 | 320 |
| | $n_D$ | 5 | 25 | 25 | 320 | 320 |
| Case 1: $\hat{t}_Y/\hat{t}_X$ | | | | | | |
| CCS | $\mathrm{RB}_{\mathbf{MC}}(\hat{V}_{CCS})$ | 0 | 0 | -1 | -1 | -1 |
| | $\mathrm{RB}_{\mathbf{MC}}(\hat{V}_{2d})$ | 1 | -1 | -16 | -1 | -5 |
| | $\mathrm{RB}_{\mathbf{MC}}(\hat{V}_{2d,a})$ | -0 | -5 | -64 | -1 | -18 |
| 2d | $\mathrm{RB}_{\mathbf{MC}}(\hat{V}_{2d})$ | -0 | 0 | -1 | -1 | 0 |
| | $\mathrm{RB}_{\mathbf{MC}}(\hat{V}_{2d,a})$ | -1 | -4 | -57 | -2 | -13 |
| Case 2: $\hat{t}_Z/\hat{t}_X$ | | | | | | |
| CCS | $\mathrm{RB}_{\mathbf{MC}}(\hat{V}_{CCS})$ | -2 | 1 | -0 | -1 | 1 |
| | $\mathrm{RB}_{\mathbf{MC}}(\hat{V}_{2d})$ | -28 | -63 | -96 | -16 | -83 |
| | $\mathrm{RB}_{\mathbf{MC}}(\hat{V}_{2d,a})$ | -29 | -65 | -98 | -17 | -85 |
| 2d | $\mathrm{RB}_{\mathbf{MC}}(\hat{V}_{2d})$ | -0 | -1 | -1 | 0 | -0 |
| | $\mathrm{RB}_{\mathbf{MC}}(\hat{V}_{2d,a})$ | -1 | -5 | -57 | -0 | -13 |

In the two-stage sampling design case, this simulation study recalls that the variance estimator $\hat{V}_{2d,a}$ is a fair approximation for $\hat{V}_{2d}$ only if the first stage sampling rate is small. The results also indicate that it seems hazardous to approximate a CCS variance estimator by a two-stage sampling variance estimator with a first stage on the maternity population. The behaviour of this simplified estimator depends on the importance of the day effect contained in the interest variable, and also depends on the sample sizes. In the ELFE case ($n_M = 320$ and $n_D = 25$), the underestimation is very high and its use is therefore not recommended. In Juillard et al. (2016), some alternative variance estimators are studied and proposed for a CCS design.

All the results of this paper are reproducible using the supplementary files which contain data and programming codes.

## Acknowledgements

I thank Guillaume Chauvet and Anne Ruiz-Gazen for their help.

## References

Aragon, Y. and Ruiz-Gazen, A. (2004). Utilisation des procédures sas dans l'enseignement des sondages. In Dunod, editor, *Echantillonnage et méthodes d'enquêtes*.

Bellhouse, D. (1981). Spatial sampling in the presence of a trend. *Journal of Statistical Planning and Inference*, 5:365–375.

Dalén, J. and Ohlsson, E. (1995). Variance estimation in the swedish consumer price indexy. *Journal of Business & Economic Statistics*, 13(3):347–356.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25(2):193–203.

Juillard, H., Chauvet, G., and Ruiz-Gazen, A. (2016). Estimation under cross-classified sampling with application to a chilhood survey. *to appear in Journal of the American Statistical Association*.

Lumley, T. (2014). survey: analysis of complex survey samples. R package version 3.30.

Ohlsson, E. (1996). Cross-classified sampling. *Journal of Official Statistics*, 12(3):241–251.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

SAS Institute Inc. (2015). *SAS/STAT® 14.1 User's Guide*. Cary, NC.

Skinner, C. (2015). Cross-classified sampling: some estimation theory. *Statistics and Probability Letters*, 104:163–168.

StataCorp. (2013). *Stata Statistical Software: Release 13*. StataCorp LP, College Station, TX.

Tillé, Y. and Matei, A. (2015). *sampling: Survey Sampling*. R package version 2.7.

Vos, J. (1964). Sampling in space and time. *Review of the International Statistical Institute*, 32(3):226–241.

Wilkerson, M. (1957). Sampling error in the consumer price index. *Journal of the American Statistical Association*, 62(319):899–914.

Correspondence: helene.juillard@ined.fr.

# 6. Appendix

### 6.1. Models used to generate the variables

In the dataset delivered with this article, the count variable $X_{ik}$ is randomly generated by a Poisson distribution with parameter $P_{ik}$, generated according to the model

$$200 + \sigma_1 U_i + \sigma_2 V_k + \sigma_3 W_{ik} \quad (9)$$

where $U_i$, $V_k$ and $W_{ik}$ are independently generated with a distribution $N(0,1)$ and with $\sigma_1 = 2$ and $\sigma_2 = \sigma_3 = 0.2$.
Conditionally to the value of $x_{ik}$, the variable $Y_{ik}$ (respectively $Z_{ik}$) is a binomial variable of parameters $x_{ik}$ and $p_{ik}^Y$ (respectively $p_{ik}^Z$). The probabilities $p_{ik}^Y$ and respectively $p_{ik}^Z$ are dependent on $i$ and $k$:

$$p_{ik}^Y = \frac{e^{\beta A_{ik}}}{1+e^{\beta A_{ik}}}$$
$$p_{ik}^Z = \frac{e^{\beta B_{ik}}}{1+e^{\beta B_{ik}}}$$

where the variable $A_{ik}$ (respectively $B_{ik}$) is generated according to the model (9) with $\sigma_1 = \sigma_2 = \sigma_3 = 0.2$ (respectively $\sigma_2 = 2$, $\sigma_1 = \sigma_3 = 0.2$) and $\beta$ is chosen in order to the average probability is 0.3.

### 6.2. Term by term variance estimation for two-stage sampling

The variance in (1) may be unbiasedly estimated term by term by

$$\hat{\mathbf{V}}_{2d}\left(\hat{t}_Y\right) = \hat{\mathbf{V}}_{PSU}\left(\hat{t}_Y\right) + \hat{\mathbf{V}}_{SSU}\left(\hat{t}_Y\right)$$

where

$$\begin{aligned}
\hat{\mathbf{V}}_{PSU}\left(\hat{t}_Y\right) &= \hat{\mathbf{V}}^1_{PSU}\left(\hat{t}_Y\right) - \hat{\mathbf{V}}^2_{PSU}\left(\hat{t}_Y\right), \\
\hat{\mathbf{V}}_{SSU}\left(\hat{t}_Y\right) &= \left(\frac{N_M}{n_M}\right)^2 \sum_{u_i\in S_M} N_i^2\left(\frac{1}{n_i}-\frac{1}{N_i}\right) s^2_{Y_{i\circ}} \\
\hat{\mathbf{V}}^1_{PSU}\left(\hat{t}_Y\right) &= N_M^2\left(\frac{1}{n_M}-\frac{1}{N_M}\right) s^2_{\hat{Y}_{\circ\bullet}}, \\
\hat{\mathbf{V}}^2_{PSU}\left(\hat{t}_Y\right) &= \frac{N_M^2}{n_M}\left(\frac{1}{n_M}-\frac{1}{N_M}\right) \\
&\quad \sum_{u_i\in S_M} N_i^2\left(\frac{1}{n_i}-\frac{1}{N_i}\right) s^2_{Y_{i\circ}}.
\end{aligned}$$

### 6.3. Analogy between two-stage sampling and one-way ANOVA: formula details

Analysis of variance (ANOVA) uses the partitioning of sums of squared deviations. For one-way ANOVA, the total sum of squares $SS_T = \sum_{u_i\in U_M}\sum_{k\in u_i}\left(Y_{ik}-\bar{Y}_{\bullet\bullet}\right)^2$ may be written as

$$SS_T = SS_M + SS_E.$$

We have

$$SS_M = \sum_{u_i\in U_M}\sum_{k\in u_i}\left(\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet}\right)^2$$

the explained sum of squares (a.k.a. the sum of squares between classes), where $\bar{Y}_{\bullet\bullet} = N^{-1}\sum_{u_i\in U_M}\sum_{k\in u_i} Y_{ik}$ is the population mean and $\bar{Y}_{i\bullet} = N_i^{-1}\sum_{k\in u_i} Y_{ik}$ is the mean inside the Primary Sampling Unit $u_i$. Also,

$$SS_E = \sum_{u_i\in U_M}\sum_{k\in u_i}\left(Y_{ik}-\bar{Y}_{i\bullet}\right)^2$$

denotes the residual sum of squares (a.k.a. sum of squares within classes).

In what follows, the factor, which is the categorical variable used to explain $Y$, is the belonging to one particular PSU $u_i$ ($N_M$ modalities). The total number of cases is $N = \sum_{u_i\in U_M} N_i$. We consider the {SI,SI} sampling case, and assume for simplicity that all the PSUs are of the same size $N_i = N_D$, and that the same sample size $n_i = n_D$ is used inside any selected PSU. In this case, we have

$$\begin{aligned}
SS_M &= \frac{N_M-1}{N_D}\, S^2_{Y_{\circ\bullet}} \\
SS_E &= (N_D-1)\sum_{u_i\in U_M} S^2_{Y_{i\bullet}}.
\end{aligned}$$

Now, we use ANOVA on the sample $S_M \times S_D$. The total number of cases is $n = n_M \times n_D$, and we denote

$$\begin{aligned}
ss_T &= \sum_{u_i\in S_M}\sum_{k\in S_i}\left(Y_{ik}-\hat{\bar{Y}}_{\bullet\bullet}\right)^2 \\
&= ss_M + ss_E,
\end{aligned}$$

where

$$\begin{aligned} ss_M &= \sum_{u_i \in S_M} \sum_{k \in S_i} \left(\hat{\hat{Y}}_{i\bullet} - \hat{\hat{Y}}_{\bullet\bullet}\right)^2 \\ ss_E &= \sum_{u_i \in S_M} \sum_{k \in S_i} \left(Y_{ik} - \hat{\hat{Y}}_{i\bullet}\right)^2 \end{aligned}$$

with $\hat{\hat{Y}}_{i\bullet} = \frac{1}{n_D} \sum_{u_i \in S_D} Y_{ik}$ the estimated population mean inside $u_i$ and $\hat{\hat{Y}}_{\bullet\bullet} = \frac{1}{n} \sum_{u_i \in S_M} \sum_{k \in S_i} Y_{ik}$ the estimated population mean.

### 6.4. *Analogy between CCS and two-way ANOVA: formula details*

For a two-way ANOVA without replication, the total sum of squares $SS_T = \sum_{u_i \in U_M} \sum_{k \in u_i} \left(Y_{ik} - \bar{Y}_{\bullet\bullet}\right)^2$ may be written as

$$SS_T = SS_M + SS_D + SS_E.$$

The total number of cases is $N = N_M \times N_D$. We have

$$SS_M = N_D \sum_{i \in U_M} \left(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}\right)^2$$

the sum of squares explained by the belonging to one particular unit $i$ ($N_M$ modalities), where $\bar{Y}_{\bullet\bullet} = \frac{1}{N} \sum_{i \in U_M} \sum_{k \in U_D} Y_{ik}$ is the population mean and $\bar{Y}_{i\bullet} = \frac{1}{N_D} \sum_{k \in U_D} Y_{ik}$ is the mean inside the unit $i$. Then, we have

$$SS_D = N_M \sum_{k \in U_D} \left(\bar{Y}_{\bullet k} - \bar{Y}_{\bullet\bullet}\right)^2$$

the sum of squares explained by the belonging to one particular unit $k$ ($N_D$ modalities), where $\bar{Y}_{\bullet k} = \frac{1}{N_M} \sum_{i \in U_M} Y_{ik}$ is the mean inside the unit $k$. Also,

$$SS_E = \sum_{i \in U_M} \sum_{k \in U_D} \left(Y_{ik} - \bar{Y}_{i\bullet} - \bar{Y}_{\bullet k} + \bar{Y}_{\bullet\bullet}\right)^2$$

denotes the residual sum of squares.

Now, we use ANOVA on the sample $S_M \times S_D$. The total number of cases is $n = n_M \times n_D$, and we denote

$$\begin{aligned} ss_T &= \sum_{i \in S_M} \sum_{k \in S_D} \left(Y_{ik} - \hat{\hat{Y}}_{\bullet\bullet}\right)^2 \\ &= ss_M + ss_D + ss_E, \end{aligned}$$

where

$$\begin{aligned} ss_M &= \sum_{i \in S_M} \sum_{k \in S_D} \left(\hat{\hat{Y}}_{i\bullet} - \hat{\hat{Y}}_{\bullet\bullet}\right)^2 \\ ss_D &= \sum_{i \in S_M} \sum_{k \in S_D} \left(\hat{\hat{Y}}_{\bullet k} - \hat{\hat{Y}}_{\bullet\bullet}\right)^2 \\ ss_E &= \sum_{i \in S_M} \sum_{k \in S_D} \left(Y_{ik} - \hat{\hat{Y}}_{i\bullet} - \hat{\hat{Y}}_{\bullet k} + \hat{\hat{Y}}_{\bullet\bullet}\right)^2 \end{aligned}$$

with $\hat{\hat{Y}}_{\bullet k} = \frac{1}{n_M} \sum_{i \in S_M} Y_{ik}$ the estimated population mean inside the unit $k$, $\hat{\hat{Y}}_{i\bullet} = \frac{1}{n_D} \sum_{i \in S_D} Y_{ik}$ the estimated population mean inside the unit $i$ and $\hat{\hat{Y}}_{\bullet\bullet} = \frac{1}{n} \sum_{i \in S_M} \sum_{k \in S_D} Y_{ik}$ the estimated population mean.

### 6.5. *Term by term variance estimation for CCS*

A term by term unbiased estimator of the variance of $\hat{t}_Y$ in formula (7) is

$$\hat{\mathbf{V}}_{CCS}\left(\hat{t}_Y\right) = \hat{\mathbf{V}}_1\left(\hat{t}_Y\right) + \hat{\mathbf{V}}_2\left(\hat{t}_Y\right) + \hat{\mathbf{V}}_3\left(\hat{t}_Y\right)$$

with

$$\begin{aligned} \hat{\mathbf{V}}_1\left(\hat{t}_Y\right) &= \hat{\mathbf{V}}_D\left(\hat{t}_Y\right) - \hat{\mathbf{V}}_E\left(\hat{t}_Y\right) \\ \hat{\mathbf{V}}_2\left(\hat{t}_Y\right) &= \hat{\mathbf{V}}_M\left(\hat{t}_Y\right) - \hat{\mathbf{V}}_E\left(\hat{t}_Y\right) \\ \hat{\mathbf{V}}_3\left(\hat{t}_Y\right) &= \hat{\mathbf{V}}_E\left(\hat{t}_Y\right) \end{aligned}$$

where $\hat{\mathbf{V}}_D\left(\hat{t}_Y\right)$, $\hat{\mathbf{V}}_M\left(\hat{t}_Y\right)$ and $\hat{\mathbf{V}}_E\left(\hat{t}_Y\right)$ are done in formula (8).