

Sampling and Survey methodology

**EMOS Learning Materials
Service Contract n. 2019.0249
between Eurostat and the
University of Pisa, Italy**

Definition

- A survey refers to any form of data collection.
- A **sample survey** is more restricted in scope: the data collection is based on a sample, a subset of the **target population** (*Eurostat (2008). Survey sampling reference guidelines*)
- The aim of a sample survey is to **make inferences** about the entire population

Target and sampled population

Target population is the population we theoretically are interested in. It is assumed to be fixed (and finite).

Sampled population is the collection of all possible observation units that might have been chosen in a sample; the population from which the sample was taken.

In an ideal survey, the sampled population will be identical to the target population, but this ideal is rarely met exactly. In surveys of people, the sampled population is usually smaller than the target population: some persons in the target population are missing from the sampling frame, and some will not respond to the survey.

Sampling Rare Populations

Rare Populations

A population can be *rare* in several ways:

- The number of individuals belonging to the rare population may be very small. For instance, relatively few people are victims of violent crime in a given year. The size of the population (N) is very small.
- There may be many individuals, but they are a small fraction of the population. Rare here refers to the rarity of the sub-population (M out of N) displaying the trait of interest, such as such as genetic disorders that occur very infrequently in live births

Rare Populations

- The elements are not necessarily rare but are cryptic or hidden.
- The proportion of sampling units containing elements from the population is very small. It is common when the population or the trait within a population is highly clustered in space or time, and the sampling units are spatial regions.

Nonresponse is a real concern in surveys of rare populations: *if population members with the rare characteristic are more likely to be nonrespondents than members without the rare characteristic, estimates of prevalence will be biased.*

Sampling frames are often nonexistent or incomplete for most rare populations

Sampling of Rare Populations

- How to design a survey to sample units that belong to a rare population?
- Rare populations require sampling designs that provide high observation rates while also controlling sample sizes
- The aim of here is to obtain a sufficiently large probability sample of the rare population for the desired accuracy while controlling costs
- We describe survey designs that have been proposed for estimating the prevalence of a rare characteristic or estimating quantities of interest for a rare populations with a particular focus on multiple frame surveys

Disproportional stratified sampling

- Defining strata such that the rare elements are concentrated in one (or a few strata) and that those strata with rare elements be oversampled to obtain sufficient elements for precise estimation
- Suppose you are interested in sampling millionaires, you may divide census block groups into strata by the estimated 90th percentile of income, and then oversample the strata where the percentile is high
- Disproportional stratified sampling may work well when the allocation is efficient for all items of interest

Network or Multiplicity Sampling

- In a network sampling, when a sampling unit (household) is selected, information is obtained both on the individuals within the household as well as on individuals in other households who are linked to those in the sampled household
- For instance, in a survey on crime victimization, the sampled household can provide information on units linked to it (the network for that household). For instance, the network of a household might be the adult siblings of adult household members.

Network or Multiplicity Sampling (2)

- Let ζ_i the network for unit i in the sample.
- The multiplicity of individual k is the number of links leading to that individual
- $\omega_k = 1/(\text{multiplicity of person } k)$ is the multiplicity weight for person k in the population of interest
- Let y_k be an indicator for whether person k was a victim of crime. Estimate the total number of crime victims by:

$$\hat{t}_{y,net} = \sum_{i \in \zeta} w_i \sum_{k \in \zeta_i} w_k y_k$$

Adaptive cluster sampling

- Select an initial probability sample of primary sampling units (PSUs)
- For each PSU in the initial sample, measure the response y
- If y in PSU i exceeds a predetermined value c , then add neighbors of PSU i to the sample
- Continue the procedure until none of the neighbors has $y > c$

Snowball Sampling

- Snowball sampling is based on the premise that members of the rare population know one another
- To take a snowball sample of homeless, you would locate an initial sample of drug-addicted persons. Ask each of those persons to identify other homeless who could be included in your sample, then ask the new persons in your sample to identify additional homeless, and so on.
- It is like the network sampling, however, in the network sampling the initial sample is a probability sample.
- In snowball sampling, the initial sample is usually chosen conveniently: snowball sample is a convenience sample, where the selection probabilities are unknown

Reading

- For more details and other sampling strategies for sampling rare population please refer to
 - Lohr, S. L. (2022). *Sampling: design and analysis*. Chapman and Hall/CRC (chapter 14)
 - Christman, M. C. (2009). Sampling of rare populations. In *Handbook of statistics* (Vol. 29, pp. 109-124). Elsevier.

Multiple Frame Surveys

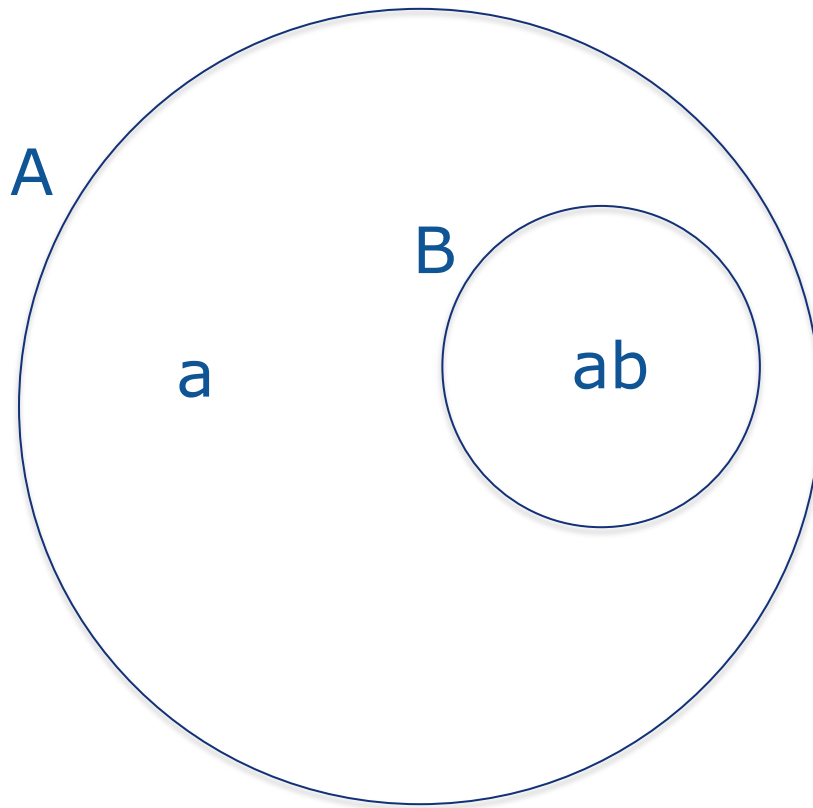
A single frame

- Usually, in the classical design-based sampling theory, we take a probability sample from a single sampling frame, containing all the units in the target population: inclusion probabilities in the sampling design are used to make inferences about the population
- Let y_i a measurement on unit i in the population of N units; let s denote the units in the sample and π_i the probability of inclusion. The the Horvitz-Thompson estimator of the population total is:

$$\hat{Y} = \sum_{i \in s} w_i y_i$$

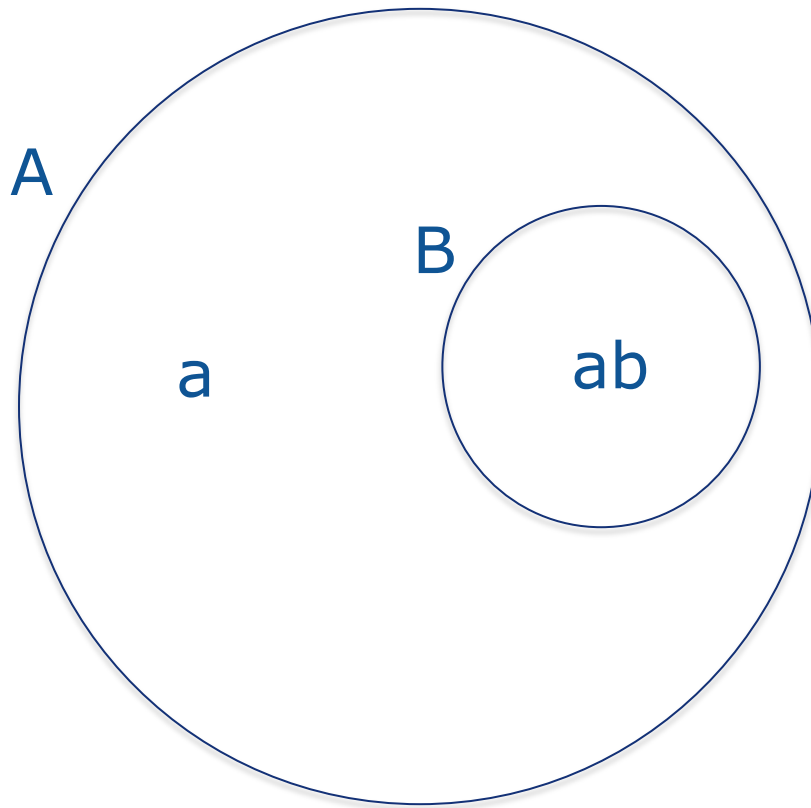
where $w_i = 1/\pi_i$ is the sampling weight

Dual frame survey



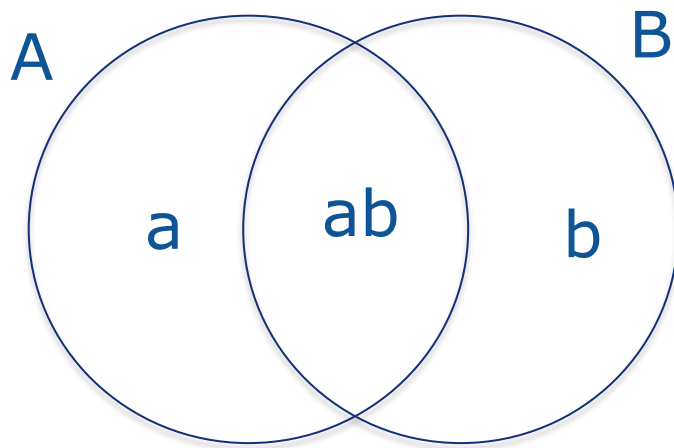
In an overlapping dual frame survey, independent probability samples are taken from frame A (the area frame) and frame B (the list frame)

Dual frame survey: example



In an epidemiology study, for example, frame A might be that used for a general population health survey, while frame B might be a list frame of clinics specializing in a certain disease. The sample from frame B is expected to yield a high percentage of persons with the disease of interest, so that sampling will be efficient; the sample from frame A, though more expensive, leads to complete coverage of the population (Lohr, 2011)

Incomplete overlapping frames

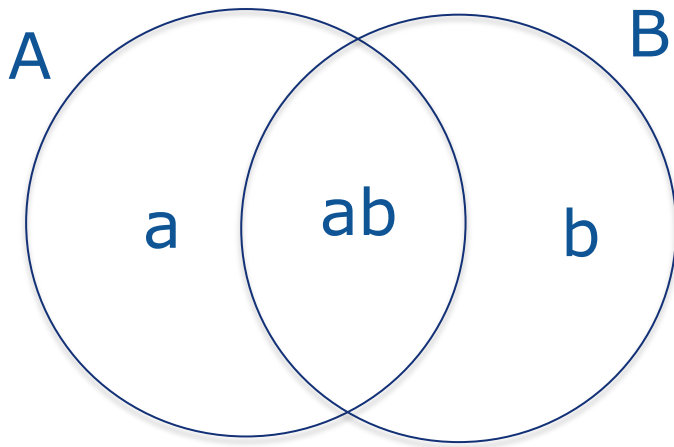


It could be situation where all frames are incomplete.

There are three domains:

- domain *a* consists of units in frame *A* but not in frame *B*
- domain *b* consists of units in frame *B* but not in frame *A*
- domain *ab* consists of units in both frames

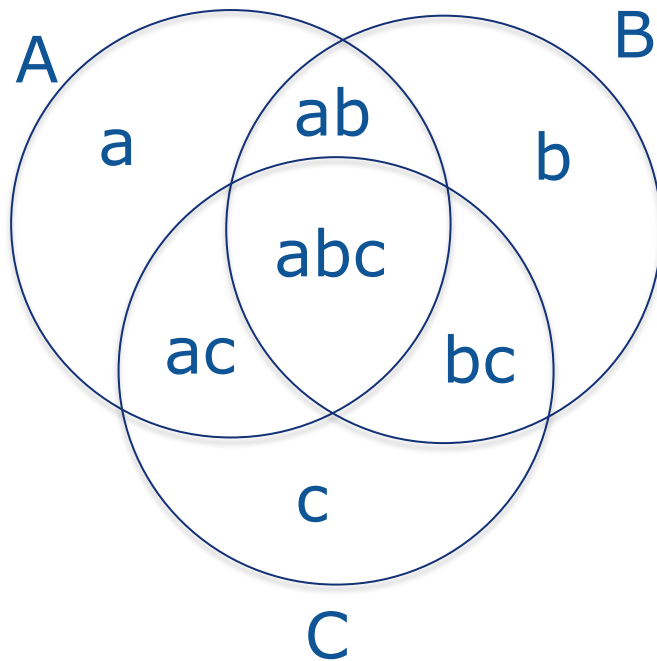
Incomplete overlapping frames: example



Frame A is a frame of landline telephones and frame B consists of cellular telephone numbers.

It is unknown in advance whether a household member sampled using one frame also belongs to the other frame (Lohr, 2011)

Three-frames

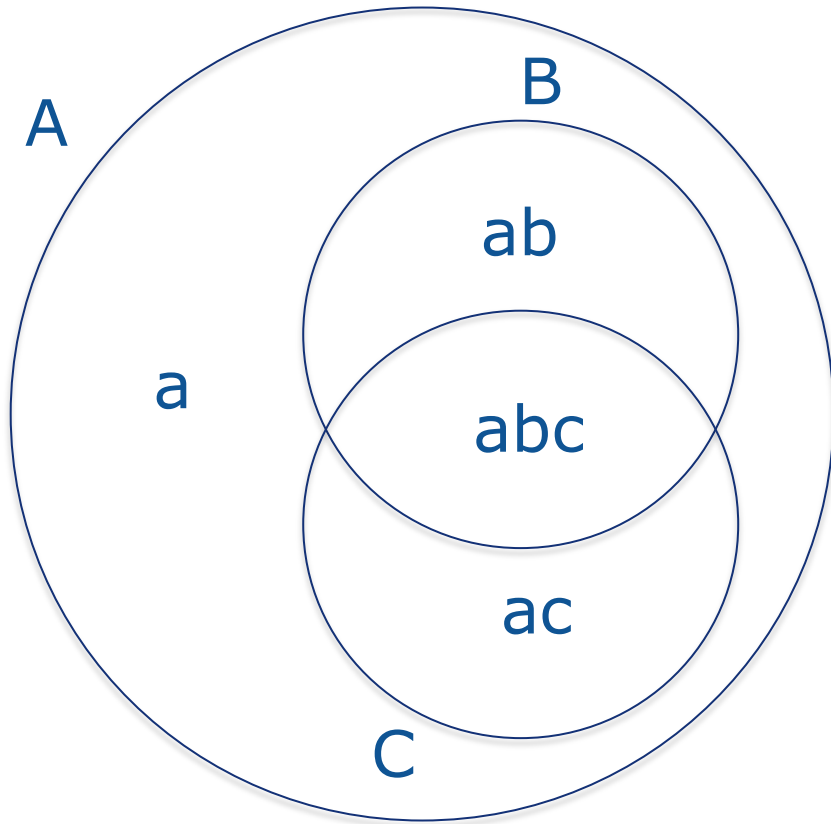


Three-frame survey in which all frames are incomplete.

There are seven domains

Example to sample the homeless population (Lohr, 2011): frame *A* is a list of soup kitchens, frame *B* is a list of shelters, and frame *C* consists of street locations

Three-frames (2)

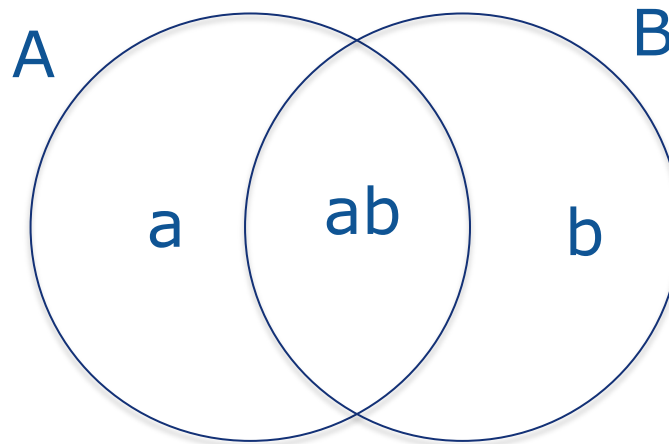


3-frame survey in which frame A has complete coverage, while overlapping frames B and C are both incomplete but are less expensive to sample.

This design might be used when A is the frame for a general population survey, B is a landline telephone survey, and C is a cell phone survey

Estimation problem

Aim: estimate the population total Y from overlapping multiple frame surveys. Formally, we consider an overlapping dual frame survey where the domain ab is nonempty.



Estimation problem (2)

- A probability sample $s(A)$ of size n_A is drawn from the N_A units in frame
- An independent probability sample $s(B)$ of size n_B is drawn from the N_B units in frame B .
- Unit i in sample $s(A)$ has probability of inclusion π_i^A and weight w_i^A
- Unit i in sample $s(B)$ has probability of inclusion π_i^B and weight w_i^B

The weights may be the inverses of the inclusion probabilities, or they may be poststratified to agree with population counts; it is assumed that estimators of population totals are approximately unbiased.

Estimation problem (3)

- $E\left[\sum_{i \in S(A)} w_i^A y_i\right] \approx Y_a + Y_{ab}$
- $E\left[\sum_{i \in S(B)} w_i^B y_i\right] \approx Y_b + Y_{ab}$

The estimator $\hat{Y} = \sum_{i \in S(A)} w_i^A y_i + \sum_{i \in S(B)} w_i^B y_i$ that combines the observations from both surveys with the original weights is biased for the population total.

It is necessary to modify weights

Estimation problem (4)

The population total is then estimated:

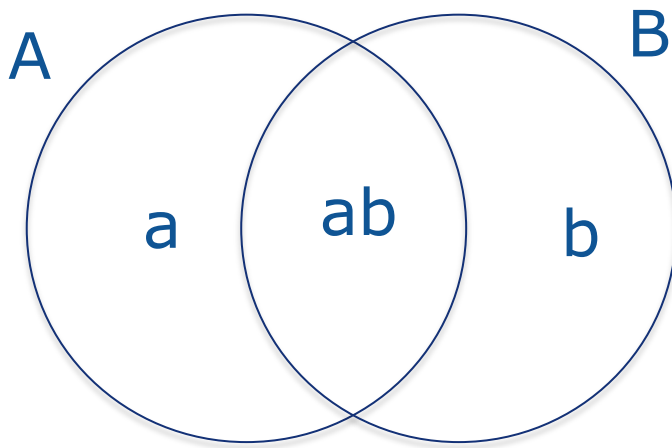
$$\hat{Y} = \sum_{i \in s(A)} \tilde{w}_i^A y_i + \sum_{i \in s(B)} \tilde{w}_i^B y_i$$

\tilde{w}_i^A are the the modified weights in the form $\tilde{w}_i^A = m_i^A w_i^A$ and $\tilde{w}_i^B = m_i^B w_i^B$.

The estimator will be approximately unbiased if $m_i^A \approx 1$ for $i \in a$, $m_i^B \approx 1$ for $i \in b$ and $m_i^A + m_i^B \approx 1$ for $i \in ab$

Point estimation

Estimator for the general dual-frame situation (Hartley, 1962)



$$\hat{Y}_H(\theta) = \hat{Y}_a^A + \theta \hat{Y}_{ab}^A + (1 - \theta) \hat{Y}_{ab}^B + \hat{Y}_b^B$$

where \hat{Y}_a^A is the estimated population total for units in domain a , \hat{Y}_{ab}^A is the estimated population total in domain ab using the sample from frame A , \hat{Y}_{ab}^B is the estimated population total in domain ab using the sample from frame B , \hat{Y}_b^B is the estimated population total for domain b , and $0 \leq \theta \leq 1$

Point estimation

To preserve unbiasedness of the proposed estimator Hartley (1962) propose the following sampling weights modification:

$$\tilde{w}_i^A = m_{i,\theta}^A w_i^A \quad \text{and} \quad \tilde{w}_i^B = m_{i,\theta}^B w_i^B$$

$$m_{i,\theta}^A = \begin{cases} 1 & \text{if } i \in a \\ \theta & \text{if } i \in ab \end{cases} \quad m_{i,\theta}^B = \begin{cases} 1 & \text{if } i \in b \\ 1 - \theta & \text{if } i \in ab \end{cases}$$

Point estimation

The estimator takes the form:

$$\begin{aligned}\hat{Y}_H(\theta) &= \sum_{i \in S(A), i \in a} m_{i,\theta}^A w_{i,\theta}^A y_i + \sum_{i \in S(A), i \in a} m_{i,\theta}^B w_{i,\theta}^B y_i \\ &= \hat{Y}_a^A + \theta \hat{Y}_{ab}^A + (1 - \theta) \hat{Y}_{ab}^B + \hat{Y}_b^B\end{aligned}$$

$$\hat{Y}_a^A = \sum_{i \in S(A), i \in a} w_i^A y_i$$

$$\hat{Y}_{ab}^A = \sum_{i \in S(A), i \in ab} w_i^A y_i$$

$$\hat{Y}_{ab}^B = \sum_{i \in S(B), i \in a} w_i^B y_i$$

$$\hat{Y}_b^B = \sum_{i \in S(B), i \in b} w_i^B y_i$$

Variance estimation

Since frames A and B are sampled independently and θ is fixed, the variance of the estimator is:

$$V[\hat{Y}_H(\theta)] = V[\hat{Y}_a^A + \theta \hat{Y}_{ab}^A] + V[(1 - \theta) \hat{Y}_{ab}^B + \hat{Y}_b^B]$$

Hartley's estimator

In the previous estimator, θ is chosen by minimizing the variance of $\hat{Y}_H(\theta)$

For a general survey-design:

$$\theta_{opt} = \frac{V(\hat{Y}_{ab}^B) + Cov(\hat{Y}_b^B, \hat{Y}_{ab}^B) - Cov(\hat{Y}_a^A, \hat{Y}_{ab}^A)}{V(\hat{Y}_{ab}^A) + V(\hat{Y}_{ab}^B)}$$

The Fuller–Burmeister estimator

To take into account the additional information regarding the estimation of N_{ab} , Fuller and Burmeister (1972) proposed modifying Hartley's estimator:

$$\hat{Y}_{FB}(\beta) = \hat{Y}_a^A + \hat{Y}_b^B + \beta_1 \hat{Y}_{ab}^A + (1 - \beta_1) \hat{Y}_{ab}^B + \beta_2 (\hat{N}_{ab}^A + \hat{N}_{ab}^B)$$

β_1 and β_2 are chosen by minimizing the variance of $\hat{Y}_{FB}(\beta)$

When a SRS is taken in each frame \hat{Y}_{FB} can be obtained from ML principles

Pseudo-ML Estimation

Skinner and Rao (1996) proposed modifying the simple random sample estimator to obtain a pseudo-maximum-likelihood (PML) estimator for a complex design.

$$\begin{aligned}\hat{Y}_{PML}(\theta) = & \frac{N_A - \hat{N}_{ab}^{PML}(\theta)}{\hat{N}_a^A} \hat{Y}_a^A + \frac{N_B - \hat{N}_{ab}^{PML}(\theta)}{\hat{N}_b^B} \hat{Y}_b^B + \\ & + \frac{\hat{N}_{ab}^{PML}(\theta)}{\theta \hat{N}_{ab}^A + (1 - \theta) \hat{N}_{ab}^B} [\theta \hat{Y}_{ab}^A + (1 - \theta) \hat{Y}_{ab}^B]\end{aligned}$$

Pseudo-ML Estimation (2)

\hat{N}_{ab}^{PML} is the smaller roots of the quadratic equation:

$$\left[\frac{\theta}{N_B} + \frac{1 - \theta}{N_A} \right] x^2 + \theta \hat{N}_{ab}^A + (1 - \theta) \hat{N}_{ab}^B - \left[1 + \frac{\theta \hat{N}_{ab}^A}{N_B} + \frac{(1 - \theta) \hat{N}_{ab}^B}{N_A} \right] x$$

The value θ_p for θ that minimizes the asymptotic variance of \hat{N}_{ab}^{PML} is used.

Pseudo-ML Estimation (3)

The adjusted weights are:

$$\tilde{w}_{i,P}^A = \begin{cases} \frac{N_A - \hat{N}_{ab}^{PML}(\hat{\theta}_P)}{\hat{N}_a^A} w_i^A & \text{if } i \in A \\ \frac{\hat{N}_{ab}^{PML}(\hat{\theta}_P)}{\hat{\theta}_P \hat{N}_{ab}^A + (1 - \hat{\theta}_P) \hat{N}_{ab}^B} \hat{\theta}_P w_i^A & \text{if } i \in ab \end{cases}$$

$$\tilde{w}_{i,P}^B = \begin{cases} \frac{N_B - \hat{N}_{ab}^{PML}(\hat{\theta}_P)}{\hat{N}_b^B} w_i^B & \text{if } i \in B \\ \frac{\hat{N}_{ab}^{PML}(\hat{\theta}_P)}{\hat{\theta}_P \hat{N}_{ab}^A + (1 - \hat{\theta}_P) \hat{N}_{ab}^B} (1 - \hat{\theta}_P) w_i^B & \text{if } i \in ab \end{cases}$$

Comparison of the estimators

- The Fuller–Burmeister estimator has the greatest asymptotic efficiency.
- The Fuller–Burmeister estimator has the greatest asymptotic efficiency among all linear estimators
- The Fuller–Burmeister and Hartley estimators both result in a different set of weights for each response variable considered. In addition, with more than two frames, these two estimators can be unstable
- This comparison has been done by Lohr and Rao (2000, 2006)

Variance estimation

- Variance estimation can be more complicated for multiple-frame surveys than for a single-frame survey
- Several methods can be used to estimate variances of estimated population quantities in general multiple frame surveys.
- These methods include Taylor linearization techniques, jackknife, and bootstrap
- See Lohr (2007, 2009) for more details

Challenges for multiple-frame surveys

- Internet can provide an inexpensive method of data collection, but a frame of internet users rarely includes the entire population of interest. This opens many possibilities for using multiple-frame surveys.
- In some cases, multiple frame may also mean **multiple mode** (the surveys from the different frames are taken using different modes)

Multi-mode surveys

- Using different data collection techniques in the same survey (face-to-face, telephone, mail, Interactive Voice Response, web interviews...).
- Trade-offs between the strong and weak points of each mode
- Advantages of Mixed mode (MM)
 - Contrast declining response and coverage rates (give respondents the option to respond in the survey mode they prefer)
 - Reduce the cost of the surveys
- A significant drawback with mixing modes in one study is that the survey mode may have an effect on the data that are collected

Mode effect in mixed mode surveys

- **Mode effect** refers to the introduction of bias effects on the survey estimates
- Mode effect has two components:
- Differential non-observation error or *mode-selection-effect* (different coverage errors and total nonresponse in each technique: desirable aspect of MM strategy)
- Differential observation error or *mode-measurement-effect* (influence of a survey mode on the answers of the respondents)
- Mode effect is net effect of non-observation and measurement error differences by mode

Mode effect in mixed mode surveys (2)

- Consider a survey with 2 modes, referred to as mode m_1 and mode m_2
- The total bias of estimator $\hat{\bar{y}}_{m_1}$ can be expressed:

$$TB = MB_{m_1} + SB_{m_1}$$

- MB: Measurement error conditional on response
- SB: Selection error respect to the population mean

Mode effect in mixed mode surveys (3)

- If we consider the true value to be the measurement obtained with a reference mode (or benchmark), m_1 for example, we can write

$$ME = E(y_{m_2} | R_{m_2} = 1) - E(y_{m_1} | R_{m_2} = 1) = ME_{m_2}$$
$$SE = E(y_{m_1} | R_{m_2} = 1) - E(y_{m_1} | R_{m_1} = 1) = SE(y_{m_1})$$

- ME_{m_2} conditional on respondents with m_2 can be view as a variation or bias due to measurement error in m_2 .
- $SE(y_{m_1})$ respect to the variable measured with m_1 is a variation of bias due to the selection error generated by the use of the mode m_2 , instead of the m_1 mode for measuring y_{m_1}
- $E(y_{m_1} | R_{m_2} = 1)$ is a "counterfactual" quantity

Distinguish selection and measurement effects (3)

Method	Analysis	Conditions	Context
Weighting - Propensity score (PS) - Calibration - Post-stratification	Analysis based on response model to control for respondent characteristics (comparable samples)	MAR assumption Mode-insensitive auxiliary variables Balancing assumption in PS	Observational studies

Distinguish selection and measurement effects (2)

Method	Analysis	Conditions	Context
Regression model (Kolenikov, Kennedy, 2014)	Model analysis to estimate measurement and selection errors	Mode-insensitive auxiliary variables to control selection effect	Observational studies
Other methods - Use of outcome regression with a propensity score model	Model to estimate causal effect	Appropriate statistical models	Observational studies

Distinguish selection and measurement effects (3)

Method	Analysis	Conditions	Context
Re-interview (Biemer, 2001) Re-interview data combined with administrative data and paradata.	Estimate Measurement effect - as remaining difference between modes. Estimate Selection effect - using mix of re-interview data, administrative data and paradata.	Re-interview does not affect measurement behavior of respondent. Nonresponse to re-interview is unrelated to variables of interest given administrative data and paradata.	Re-interview of subset of mixed-mode respondents

Multi-mode surveys

- More detail on mixed mode surveys, the formalization of the mode effects and the methods of adjusting the mode effect have been discussed during two EMOS Webinars:
 - [Mixed-mode surveys, Edith de Leeuw & Anne Elevelt, Utrecht University](#)
 - [The mode effect in mixed-mode surveys Claudia De Vitiis & Francesca Inglese, ISTAT](#)

**...there's much more
about Sampling and
Survey methodology!**

References

- Cochran, W.G. (1977). *Sampling Techniques*, 3rd Ed. New York: John Wiley & Sons,
- de Leeuw, E. D. (2005). To mix or not to mix data collection modes in surveys. *Journal of official statistics*, 21(5), 233-255.
- de Leeuw, E.D., Hox, J., & Dillman, D. (Eds.). (2008). *International Handbook of Survey Methodology* (1st ed.). Routledge.
- Hartley, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section*, American Statistical Association, 203-206
- Fuller, W.A., and Burmeister, L.F. (1972). Estimators for samples selected from two overlapping frames. *Proceedings of the Social Statistics Section, American Statistical Association*, 245-249.

References

- Kalton, G. (2003). Practical methods for sampling rare and mobile populations. *Statistics in Transition* 6, 491–501
- Lohr, S.L. (2007). Recent developments in multiple frame surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3257-3264.
- Lohr, S.L. (2009). Multiple frame surveys. In *Handbook of Statistics, Sample Surveys: Design, Methods and Applications*, (Eds., D. Pfeffermann and C.R. Rao). Amsterdam: North Holland, Vol. 29A, 71-88.
- Lohr, S. L. (2022). Sampling: design and analysis. Chapman and Hall/CRC (chapter 14)

References

- Lohr, S.L., and Rao, J.N.K. (2000). Inference in dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280
- Pfeffermann D. and Rao C.R. (Eds.). (2009). Handbook of statistics (Vol. 29A). Chapter 4 and 6. Elsevier.
- Skinner, C.J., and Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356
- Tourangeau, R., B. Edwards, T. P. Johnson, K. M. Wolter, and N. Bates (Eds.) (2014). *Hard-to- Survey Populations*. Cambridge University Press.