Proceedings of Statistics Canada Symposium 2018
Combine to Conquer: Innovations in the Use of Multiple Sources of Data

Indirect sampling applied to capture-recapture models with dependence between sources

Herménégilde Nkurunziza, and Ayi Ajavon[1]

Abstract

Capture-recapture is a method that is widely used to estimate the total number of units in a population of unknown size. It involves drawing two independent samples from the target population. The Petersen estimator of population size is one that is used frequently, and depends on the size and overlap between the two samples. Lavallée and Rivest (2012) looked at the case where the samples come from *indirect sampling* and introduced a generalization of the Petersen estimator based on *the generalized weight share method* (GWSM). In practice, the assumption of independence on which the estimator is based is often not verified (Brenner, 1995). In this article, we will focus on the capture-recapture models with dependence between the sources and propose an extension of the Lavallée and Rivest (2012) estimator. We analyze the properties of the obtained estimator and provide an example of the method using simulated data.

Key words: Indirect sampling; Generalized method of weight sharing; Sampling frame; Estimator; Dependence.

1. Background and objectives

1.1 Indirect sampling

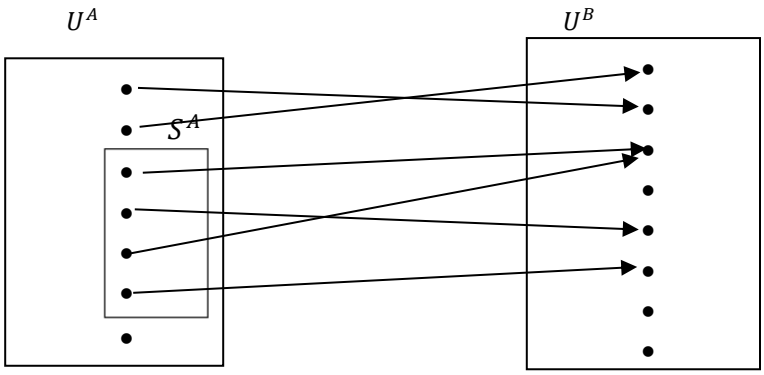
Under normal sampling conditions, there is a sampling frame for the population of interest. From this frame, a sample is drawn to produce estimates. However, in some situations there is no sampling frame for the population of interest, but there is one for another population that is related to the population of interest in some way. Thus, indirect sampling is used to survey populations that are difficult to study, find or reach because there are few individuals in the target population or because no reliable sampling frame exists.

Indirect sampling is also used with populations for which measurements or interviews are difficult to obtain (for example, if prohibited by law) (Lavallée, 2016; Kiesl, 2016). In this case, you would use a different sampling frame that is still related in some way to the difficult-to-sample population (Lavallée, 2016). For example, with a list of agencies that provide services to individuals with no fixed address (e.g., meals, housing), the total number of these individuals could be estimated for a given city using indirect sampling. Alternatively, a survey on children could be conducted if all that is available is a list of parents.

Formally, there are two related populations: U^A and U^B (to keep the same notations as Lavallée, 2002). We want to produce an estimate for U^B, but a sampling frame is only available for U^A. A sample of U^A is drawn and then used to produce an estimate for U^B.

Figure 1.1-1
Graphic representation of indirect sampling

1. Herménégilde Nkurunziza, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (hermenegilde.nkurunziza@canada.ca); Ayi Ajavon, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (Ayi.Ajavon@canada.ca)



The major challenge with indirect sampling is defining the link between U^A and U^B (Kiesl, 2016). Once this relationship has been defined, the next challenge is to associate a probability of selection or sampling weight to the surveyed units of the target population U^B (Deville and Lavallée, 2006). To resolve the latter problem, Lavallée (1995) proposed the Generalized Weight Share Method (GWSM), which makes it possible to associate weights from *k* units of U^B that are related to the sampled units *i* of U^A, as follows:

-A sample S^A is selected from U^A, $\pi_i > 0$, which is the selection probability of unit *i*. The sampling weight is thus $\frac{1}{\pi_i}$, provided there are no other adjustments. The sampling weight of unit *k* of U^B is given by (Lavallée, 2002):

$$\widehat{w}_k = \sum_{i \in S^A} \frac{1}{\pi_i} \frac{l_{i,k}}{L_k^B} \quad (1)$$

where $l_{i,k} = 1$ if unit *i of* U^A is related to unit *k* of U^B, otherwise $l_{i,k} = 0$

$$L_k^B = \sum_{i \in U^A} l_{i,k} \quad (2)$$

This weight can be considered an average of the sampling weights of the units *i* of population U^A that are related to *k* units of U^B (Lavallée, 2002). Once the weights of the sampled units of S^A have been established, the estimation for U^B is acquired in the usual way. Non-responses of the units of S^A are also dealt with in the usual survey manner.

As previously mentioned, the major difficulty with the GWSM is determining whether an element *i* of U^A is related to unit *k* of U^B, because doing so can create biases in the estimates. In practice, for example, the links between elements *i* and units *k* can be established from the interviews of the units selected from the sample S^A (Deville and Lavallée, 2006) or by matching.

1.2 Capture-recapture

The capture-recapture method has long been used to estimate the size of an unknown population. One classic example of this method is to estimate the number of fish in a lake (Lavallée and Rivest, 2012). It involves drawing a sample size (n_1) of the population being estimated, labelling the sampled units, and returning them back into the population. Next, a second sample of this population is taken (n_2), and an estimate of the total population size is produced based on the size of these two samples and the number of units common to both samples (n_{12}). The Peterson estimator is then used.

$$\widehat{N} = \frac{n_1 n_2}{n_{12}} \quad (3)$$

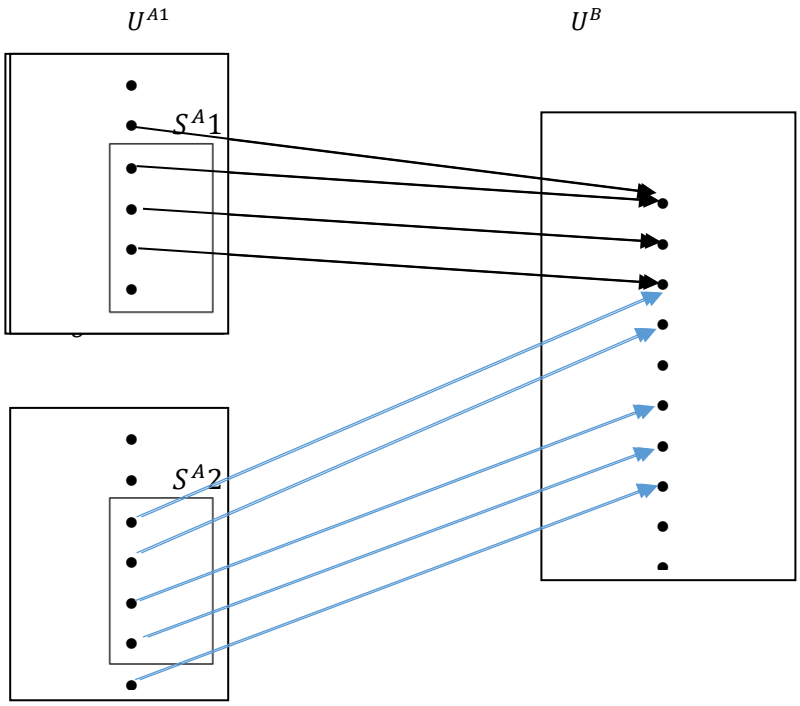
This method is currently used more in biology and epidemiology to estimate the size of populations that are difficult to reach or count, such as the number of individuals affected by a given disease (Lavallée and Rivest, 2012; Tilling, 1999), but also in other areas of study (Corrao et al., 2000; Kiesl, 2016). The Peterson estimator is also used to estimate the size of a population that is partially covered by two files. In this case, the Peterson estimator is given by:

$$\widehat{N}_{pet} = \frac{N_1 N_2}{N_{12}} \quad (4)$$

where N_1 and N_2 are the respective registration numbers for the two files, and N_{12} is the number of records common to both files.

A more detailed description of the Peterson estimator is given in Lavallée and Rivest (2012). They were interested in the capture-recapture method in a context where the samples originate from *indirect sampling*, which is the case where both files (A1 and A2) do not represent the population of interest, but other populations are somehow related to the population of interest. Graphically, the situation is shown below.

Figure 1.2-1
Capture-recapture and indirect sampling



They presented a generalization of the Petersen estimated based on the GWSM, referred to as the generalized capture–recapture estimator (GCRE), given by:

$$\widehat{N}^{B}_{GCRe} = \frac{\widehat{N}^{B}_{A1}\widehat{N}^{B}_{A2}}{\widehat{N}^{B}_{A1,A2}} \quad (5)$$

where

$\widehat{N}^{B}_{A1}$ (Respectively $\widehat{N}^{B}_{A2}$) is the estimate based on the units related to those of A1 (respectively A2), with survey weights given by (1).

$\widehat{N}^{B}_{A1,A2}$ is the estimate based on the units related to units of A1 and A2, and whose weights are given by $w_k^{A1,A2} = \left(\frac{1}{L_k^{A1}}\sum_{i\in S^{A1}}\frac{l_{ik}}{\pi_i^{A1}}\right)\left(\frac{1}{L_k^{A2}}\sum_{i\in S^{A2}}\frac{l_{ik}}{\pi_i^{A2}}\right)$ (6).

They then showed how the method could be applied to estimating the total of any variable of interest Y.

2. Indirect sampling applied to capture-recapture models with dependence between sources

As previously mentioned, the estimator (5) is based on the theory of independence between sources. In practice, however, this theory is not always confirmed (Brenner, 1995) because it states that dependence between sources leads to bias in the estimators (Chao, 2001; Corrao et al., 2000). Dependence can occur in situations such as the following examples: in the capture-recapture of animals, the first capture could create a feeling of fear/panic in the animals, thereby creating a negative correlation among the captures; and, in epidemiology, the lists used might be dependent (Chao, 2001).

When there is a positive dependence between the sources, the probability of finding cases on one file increases the probability of finding these cases on the other file. Alternatively, when there is a negative dependence between the sources, the probability of finding cases on one file reduces the probability of finding these cases on the other file (Brenner, 1995).

Very few studies have looked at cases where the sources are dependant. In this article, we are specifically looking at indirect sampling applied to capture-recapture models with dependence between the sources, and are proposing an extension of the estimator (5).

In the case of negative dependence between the 2 sources, we propose an extension of the GCRE as follows:

$$\widehat{N}^B_{\text{GCRed}} = \frac{\widehat{N}^B_{A1}\widehat{N}^B_{A2}}{\widehat{N}^B_{A1,A2}+\frac{|cov(X_{A1},X_{A2})|}{\min(p_1-p_{12},\ p_2-p_{12})}(\min(\widehat{N}^B_{A2},\widehat{N}^B_{A1})-\widehat{N}^B_{A1,A2})} \quad (7)$$

where:

$\widehat{N}^B_{\text{GCRed}}$ is the generalized capture-recapture estimator with dependence between sources.

$\widehat{N}^B_{A1}$ and $\widehat{N}^B_{A2}$ are defined as in (5) above. $\widehat{N}^B_{A1,A2}$ is the estimate based on the units related to units of A1 and of A2, using weights defined as in (1) by

$$w_k^{A1,A2} = \frac{1}{(L_k^{A1}+L_k^{A2})}\left(\sum_{i\in S^{A1}}\frac{l_{ik}}{\pi_i^{A1}}+\sum_{i\in S^{A2}}\frac{l_{ik}}{\pi_i^{A2}}\right) \quad (8)$$

$cov(X_{A1},\ X_{A2})$ is the covariance between the fact that it is in file A1 and the fact of being in file A2, and we get $-1 \le cov(X_{A1},\ X_{A2}) \le 1$ (Brenner, 1995)
$X_{A1i} = 1$ if element *i* is in file A1, and $X_{A1i} = 0$ otherwise
$X_{A2i} = 1$ if element *i* is in file A2, and $X_{A2i} = 0$ otherwise
p_1 is the probability of being in field A_1, p_2 the probability of being in field A_2, and p_{12} the probability of being in field $A_1 \cap A_2$. Quantity $p_1 - p_{12}$ is equal to the expectation of the part related to A1 minus the intersection, and $p_2 - p_{12}$ is equal to the expectation of the part related to A2 minus the intersection.

The term $\frac{|cov(X_{A1},X_{A2})|}{\min(p_1-p_{12},\ p_2-p_{12})}\left(\min(\widehat{N}^B_{A2},\widehat{N}^B_{A1})-\widehat{N}^B_{A1,A2}\right)$ can be seen as the term correcting the bias that results from the dependence between the sources.

Note: In the equation (7),

1) If both sources are independent then $cov(X_{A1},\ X_{A2}) = 0$ and we find the expression (5) of $\widehat{N}^B_{GCRe}$.
2) If both sources are highly dependent, with a negative correlation then $cov(X_{A1},\ X_{A2}) \cong \min((1-p_1)(1-p_2)\ ,\ p_1p_2\ )$.
3) If $p_1 + p_2 > 1$ then $(1-p_1)(1-p_2) <\ p_1p_2$ (Brenner, 1995). We get

$$\widehat{N}^B_{\text{GCRed}} \cong \frac{\max(N^B_{A1},N^B_{A2})\min(p_1-p_{12},\ p_2-p_{12})}{(1-p_1)(1-p_2)} \cong \frac{\max(N^B_{A1},N^B_{A2})}{\max((1-p_1),(1-p_2))}\ .$$

In the event of positive dependence, we get an estimator similar to (7) through symmetry.

2.1 Property of proposed estimator

Lemme: low convergence.

If we consider the estimator $\widehat{N}^B_{\text{GCRed}} = \frac{\hat{N}^B_{A1}\hat{N}^B_{A2}}{\hat{N}^B_{A1,A2}+\frac{|cov(X_{A1},X_{A2})|}{\min(p_1-p_{12},\ p_2-p_{12})}(\min(\hat{N}^B_{A2},\hat{N}^B_{A1})-\hat{N}^B_{A1,A2})}$.

The negative dependence between the sources implies that there is $\theta \geq 0$ such that $p_1p_2 = (1-\theta)p_{12} + \theta min(p_1, p_2)$ (Kimeldorf and Sampson, 1989).

Therefore, estimator $\hat{N}^B_{GCRed}$ converges in probability toward the size of the population.

Proof: The proof is identical to the independent case (Lavallée and Rivest, 2012), showing simply that $\hat{N}^B_{GCRed}/N^B$ converges slightly toward 1 due to the fact that the numerator $\hat{N}^B_{A1}\hat{N}^B_{A2}$ and the denominator converge toward the same quantity.

3. Simulation

We want to estimate the number of cell phone users in a city, based on files A1 and A2 supplied by the only two phone providers. A1 contains 1,000 numbers and A2 contains 800 numbers. Using SRS, a 500-number sample is taken from each file (the probabilities of inclusion are p_1=1/2 and p_2=5/8). We call each number selected and obtain the owner's information. This creates the link between the two lists and the person's files.

Assuming that each company does not give a person more than one number, but that a person can have more than one number from different providers.

After completing the calls, we fill in the table, as shown in the following example (fictitious data).

Number	X_{A1} (A1's number)	X_{A2} (A2's number)	Owner
613 000 6644	1	0	Jean
819 333 9999	0	1	Alice
613 777 0000	1	0	Peter
613 777 8888	1	0	Alice
613 999 0000	0	1	Jean
613 000 2222	1	0	
819 000 5555	0	1	Smith

In this example, Jean and Alice each have two numbers—one from each of the two providers.
In this simulation, we see that there is a strong negative dependence, so we can take $cov(X_{A1}, X_{A2}) \cong (1-p_1)(1-p_2)$.

Assuming also that:

- 400 of the 500 numbers in sample S^{A1} are related to owners;
- 450 of the 500 numbers in sample S^{A2} are related to owners;
- In both samples, 30 individuals have one number for each of the two providers.

Using the expressions (1) and (6), we arrive at the following estimates:

$\widehat{N}^B_{A1}$	$\widehat{N}^B_{A2}$	$\widehat{N}^B_{A1,A2}$
800	810	54

Since the dependence is negative and strong, and since $p_1 + p_2 > 1$ then
$|\text{cov}(X_{A1},\ X_{A2})| \cong (1-p_1)(1-p_2)$ (Brenner, 1995).

This gives $\widehat{\mathrm{N}}^{\mathrm{B}}_{\mathrm{GCRed}} \cong \frac{\max(\widehat{\mathrm{N}}^{\mathrm{B}}_{\mathrm{A1}}, \widehat{\mathrm{N}}^{\mathrm{B}}_{\mathrm{A2}})}{\max((1-\mathrm{p}_1),(1-\mathrm{p}_2))} = \frac{810}{1/2} = 1620$.

If we ignore the dependence $\widehat{\mathrm{N}}^{\mathrm{B}}_{\mathrm{GCRe}} = \frac{800 \times 810}{96} = 6750$.

This would overestimate the total population owning a cellular telephone.

4. Conclusion

We have proposed an estimator of the total size for a given population using indirect sampling applied to the capture-recapture method, where there is dependence between the sources. The estimator presented here is only valid for a negative dependence. For a positive dependence, a similar estimator is obtained through symmetry. It is assumed that the links between the units have been established correctly.

References

Brenner, H. (1995), "Use and Limitations of the Capture-Recapture Method in Disease Monitoring with Two Dependent Sources", *Epidemiology*, 6(1), pp. 42-48.

Chao, A. (2001), "An Overview of Closed Capture–Recapture Models", *Journal of Agricultural, Biological, and Environmental Statistics*, 6(2), pp. 158–175.

Deville, J.-C., and P. Lavallée (2006), "Indirect Sampling: the Foundations of the Generalised Weight Share Method", *Survey Methodology*, 32, pp. 165–176.

Giovanni, C. G. et al. (2000), "Capture-recapture methods to size alcohol related problems in a population", *J Epidemiol Community Health*, 54, pp. 603–610.

Kiesl, H. (2016), "Indirect Sampling: A Review of Theory and Recent Applications", *German Statistical Society*, 10(4), pp. 289-303.

Kimeldorf, G., and A. R. Sampson (1989), "A framework for positive dependence", *Ann. Inst. Statist. Math*, 41(1), pp. 31-45.

Lavallée, P. (1995), "Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households using the Weight Share Method", *Survey Methodology*, 21, pp. 25–32.

Lavallée, P. (2002), *Le sondage indirect ou la méthode généralisée du partage de poids*, Bruxelles: Éditions de l'Université de Bruxelles.

Lavallée, P. (2016), "Le sondage indirect pour les populations difficiles à joindre", Course offered at Statistics Canada, Canada.

Lavallée, P., and L. P. Rivest (2012), "Capture-recapture sampling and indirect sampling", *Journal of Official Statistics*, 28(1), pp. 1-27.

Tilling, K., and J. A. C. Sterne (1999), "Capture-Recapture Models Including Covariate Effects", *American Journal of Epidemiology*, 149(4), pp. 392-400.