Developments on Coordinated Poisson Sampling

Lionel, Qualité
University of Neuchâtel, Institute of Statistics
Pierre-à-Mazel 7
2000 Neuchâtel,Switzerland
E-mail: lionel.qualité@unine.ch

1 Introduction

The Swiss Federal Office of Statistics (SFSO) uses a coordinated sampling system developed in Qualité (2009) that extends the method proposed in Brewer et al. (1972). Each transversal sample selected through this system stems from a Poisson sampling design. This procedure, with its inherent size-variability, calls for updated planification methods of target sample sizes within domains, to replace allocation optimization techniques that were used for stratified designs. One particular aspect, introduced with these Poisson designs, and that did not exist with the stratified sampling designs that were commonly used before the introduction of the coordination system, is the risk of selecting a sample that is well below the expected size in some domains. This risk is also present when non-response is modeled as a second-phase Bernoulli sampling within domains. Another problem for which we give our best yet found solution, is the unit-level selection dependence required in some surveys (e.g. surveys where at most one selection may occur in any household), and unobtainable through a simple use of our coordinated sampling system. In order to satisfy these requirements, we propose to use coordinated selection as one phase in a multi-stage sampling design. The computation of correct (conditional-)inclusion probabilities at each phase is however non-trivial in the general case.

2 Coordinated Poisson Sampling

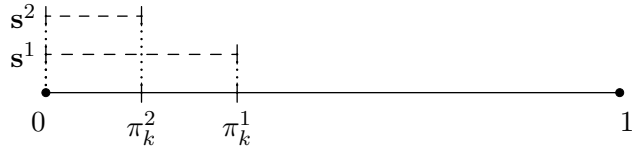
In Qualité (2009), we proposed a coordinated sampling method that allows to obtain, a minimal or maximal correlation between selection indicators I_k^t of unit k in different samples s^t for all k in a population U. It is a natural extension of Brewer et al. (1972)'s sampling design for two surveys. The method of Brewer et al. (1972) consists in generating a permanent uniform random number u_k in $[0,1]$, and defining selection zones as subsets of $[0,1]$ for each unit k in such a way that,

1. the length of the selection zone for sample s^1 (resp. s^2) is equal to the desired inclusion probability π_k^1 (resp. π_k^2),

2. the overlap between selection zones is minimal if negative coordination is desired and maximal if positive coordination is desired.

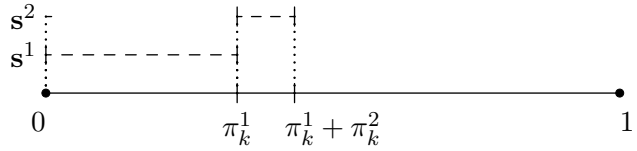
For positive coordination, it amounts to define the selection zone of k in s^1 as $[0,\pi_k^1)$ and for the selection of k in s^2 as $[0,\pi_k^2)$. Thus $[0,1]$ is effectively split into three intervals as in Figure 1. Each of these intervals corresponds to a possible value of the couple (I_k^1, I_k^2).

Figure 1: positive coordination when $\pi_k^2 \leq \pi_k^1$



For negative coordination the selection zones in s^1 and s^2 are typically equal respectively to $[0, \pi_k^1)$ and $[\pi_k^1, \pi_k^1 + \pi_k^2)$, if the sum of inclusion probabilities does not exceed 1, as in Figure 2. In the

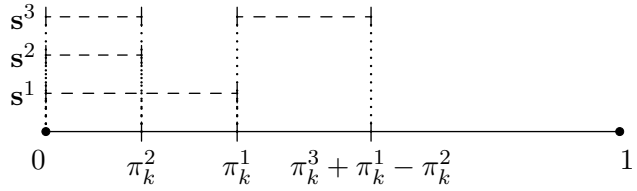
Figure 2: negative coordination when $\pi_k^1 + \pi_k^2 \leq 1$



general case of negative coordination, $[0, 1]$ is split into three intervals, the boundaries of which are given by 0, π_k^1, $(\pi_k^1 + \pi_k^2)$ mod1 and 1.

We extend this idea to the selection of any number of samples by defining recursively for any new survey a selection zone for each unit. The principle is easily understood on an example: say that, after s^1 and s^2 have been selected with positive coordination, and that the situation is similar to that of Figure 1. Suppose that one wants to select a third sample s^3 positively coordinated with s^2 but negatively coordinated with s^1, and that, for example, $\pi_k^3 > \pi_k^2$. Then, the selection zone for s^3 will contain the selection zone for s^2 and an other bit of $[0, 1]$ that respects the desired coordination rules in the best possible way. In this case, typically, we will add $[\pi_k^1, \pi_k^1 + \pi_k^3 - \pi_k^2)$ to obtain the selection zone of s^3, as in Figure 3.

Figure 3: Coordination of a third sample



More formally, at time t and for a unit k, the interval $[0, 1]$ is split into a collection of at most $t + 1$ sub-intervals. Each of these sub-intervals is associated to a possible history of selections of unit k. The addition of a new survey s^{t+1} is obtained by including into the selection zone for s^{t+1} the intervals that correspond to the most desirable history of selections, and usually splitting one of these sub-intervals into two parts so that the total length of the selection zone is equal to the desired inclusion probability π_k^{t+1}. A total order on the sub-intervals is necessary to make this operation. The one we use is obtained by asking of the user to specify the type of coordination that he would like to have with each previous survey, and to give an order of priority for these coordinations.

The transversal sampling designs are Poisson sampling designs, and hence are random size. If coordinations are all negative and respect the order of selection in time, the longitudinal design for all units is systematic, which is arguably (see for example Nedyalkova et al., 2009) the best design regarding burden repartition.

3 Developments and real life adaptation

The method presented in section 2 is flexible enough to draw all types of samples currently used in the SFSO: one occasion surveys, panels updated every other years and rotating panels. The latest are in fact selected as collections of subsamples. For example, if the expected rotation rate is 20%, we select five subsamples that constitute the initial sample. The following year, five other subsamples

are selected with coordination rules that ensure that four out of the five initial samples are simply updated for births and deaths in the population, while the fifth is replaced by a new and negatively coordinated sample. Births and deaths in the population do not jeopardize the system as units are treated independently, and even mergers and splits can be dealt with by transmitting the history of one former unit to a new one when it seems to make sense.

This sampling program has been used in the SFSO since October 2009 for business surveys, and since November 2010 for population surveys. It has admittedly a modest impact on business surveys burden repartition. Indeed, in business surveys, most units are either in "take-all strata" or have very small inclusion probabilities. In both cases, sample coordination does not bring much compared to independent selections. Still, it provides a simple method with solid theoretical foundations to update panel samples, and to draw rotating panels in a dynamic population, as well as the assurance that we did the best we could to avoid unnecessarily frequent selections of the same units. For population surveys, the need for a coordination method has been made pregnant by the introduction of an annual "structural survey" with a sampling fraction close to 7%.

Up to now, two limitations of the method have had to be accounted for. Both are related to the fact that transversal designs are Poisson designs. The simplicity of this sampling design is the reason we are able to implement a flexible coordination sampling program, but it does not completely suit every needs of the statistician who is only concerned with his sole upcoming one-occasion survey. One aspect of Poisson sampling that may be problematic is its random size. As we see in section 3.1, this has little to no effect on the expected accuracy of the sampling strategy, but the risk to select a sample smaller than anticipated exists, and some measures have to be taken to account for that risk. The other problematic aspect is that, with Poisson sampling, the selection of a unit is independent from the selection of another one. However, for business surveys as well as for population surveys, two kinds of units are of interest: companies and establishments in the first case, individuals and households in the second. In the case of population surveys, the usual procedure at the SFSO before a population register became available, was to select phone numbers in an available sampling frame, followed by a selection of one respondent unit in households corresponding to selected phone numbers. We discuss in section 3.2 how we managed to select samples of one unit per household through our sampling system.

3.1 Planification with Poisson sampling

When we introduced our coordination system, with its Poisson transversal designs, the concern most frequently expressed by our partners was with the loss of precision anticipated due to its random size. And it is true that, for some variables strongly correlated to the inclusion probabilities, a random sampling design used in conjunction with the Horvitz-Thompson estimator (Horvitz & Thompson, 1952) has higher variance than a fixed size design also used with the Horvitz-Thompson estimator. However, in practice, it is never the Horvitz-Thompson estimator that is used for estimation, but rather the Hájek estimator (Hájek, 1971) or better a calibrated estimator (see Deville & Särndal, 1992). Then, if the inclusion probabilities are among the calibration variables, sample size randomness is almost entirely irrelevant for the precision of the sampling strategy, as is shown in the following widely applicable example.

Consider a population of size N, an interest variable y with corrected variance S_y^2 and the simplest possible example of Bernoulli sampling used with Hájek's estimator, noted $\widehat{Y}_{Hj}(s)$, and simple random sampling without replacement, with the same inclusion probabilities $p = n/N$, used with Horvitz-Thompson's estimator noted $\widehat{Y}_{HT}(s)$ (see for example Särndal et al., 1992). Conditional on

size $n(s)$ of a sample and provided that $n(s) \neq 0$, we get that

$$(1) \quad \text{var}\left(\widehat{Y}_{Hj}|n(s)\right) = N^2\left(1-\frac{n(s)}{N}\right)\frac{S_y^2}{n(s)}, \text{ and } \text{E}\left(\widehat{Y}_{Hj}|n(s)\right) = Y,$$

where Y is the true population total of y. In order to carry out computations, we need to extend $\widehat{Y}_{Hj}(s)$ to the null sample and choose a value $\widehat{Y}_{Hj}(\emptyset)$. Estimator $\widehat{Y}_{Hj}(s)$'s bias is equal to

$$(2) \quad \text{B}(\widehat{Y}_{Hj}) = (1-p)^N\left(\widehat{Y}_{Hj}(\emptyset) - Y\right),$$

and is of the order of $\exp(-n)$ if N is large enough. In most applications, $\exp(-n) \ll 1/n$ and we will neglect this bias. In order to simplify the variance computation, suppose that $\widehat{Y}_{Hj}(\emptyset) = Y$. Then,

$$(3) \quad \text{var}(\widehat{Y}_{Hj}) = \text{var}\left\{\text{E}\left[\widehat{Y}_{Hj}|n(s)\right]\right\} + \text{E}\left\{\text{var}\left[\widehat{Y}_{Hj}|n(s)\right]\right\},$$

simplifies to

$$\begin{aligned}
\text{var}(\widehat{Y}_{Hj}) &= \text{E}\left\{\text{var}\left[\widehat{Y}_{Hj}|n(s)\right]\right\}, \\
&= \sum_{m=1}^{N} N^2\left(1-\frac{m}{N}\right)\frac{S_y^2}{m}\binom{N}{m}p^m(1-p)^{N-m}, \\
&= N^2 S_y^2\left[1-(1-p)^N\right]\left[\frac{1}{1-(1-p)^N}\sum_{m=1}^{N}\frac{1}{m}\binom{N}{m}p^m(1-p)^{N-m} - \frac{1}{N}\right].
\end{aligned}$$

Approximations for the summation in the last expression are available in Thionet (1963); Marciniak & Wesolowski (1999); Grab & Savage (1954); David & Johnson (1956). They all lead to conclude that

$$(4) \quad \text{var}(\widehat{Y}_{Hj}) = \text{var}(\widehat{Y}_{HT}) + \mathcal{O}\left(n^{-2}\right).$$

The real problem is that, in small domains, the selected sample can have a smaller size than what is deemed acceptable, even before the non-response phase. Variances conditional to size will then be uncomfortably large. In order to limit that risk, we may choose to modify the initial allocation of the sample between domains and increase the sampling size in certain domains. When inclusion probabilities are equal within domains, we can easily compute the probability of obtaining a sample size below a given value $P(n(s) < n_{min})$, that is a function of the sampling rate. We then invert this function and determine sampling fractions such that $P(n(s) < n_{min}) = \alpha$ where α is the accepted risk of obtaining a sample that is too small. This leads us to modify our allocation algorithms, and accept a result that is less than optimal for the estimation of a total on the whole population. Also, when there is a large number of such small domains, it becomes very costly, in terms of precision or of expected sample size, to use a parameter α small enough so that the probability of having one or more unwanted domain sample sizes remains small. While this is a serious problem, it is in fact inherent to all sampling operations with non-response when one models the non-response phase by a Bernoulli or multinomial sampling design. Cost added by the random size Poisson selection is then relatively small compared to that of controlling risks of an unlucky non-response phase result.

3.2 Introducing unit-level dependence

In order to limit the survey burden within households, and hopefully obtain better response rates, it is common practice in population surveys not to select more than one individual per household. Most SFSO samples were, up to now, designed that way. Before 2010, the only available and practical sampling frame was a phone number register. Selecting one unit per phone number after having listed all the members of the contacted household also allowed to precisely control the sample size. The new

population register, however, is obtained by collecting population lists of all municipalities (for all people living in Switzerland, it is mandatory to register with their commune of residence authorities, a foreign workers office or an immigration office). The first objective of this administrative process is to keep track of individuals, not of households. Thus it was considered safer to build a sampling frame of individuals, with their social security numbers as identifiers than to build a sampling frame of households. However, for most people a household number is also available, and the inclusion of this number will become mandatory by the end of 2013.

While samples of one unit per household cannot be obtained directly with our coordinated sampling system, a two-phase operation allows to select such samples. Computations are relatively easy when inclusion probabilities within any household are equal or null, and become complex otherwise. Also, when such a sample has to be selected on multiple different occasions while taking care of not selecting units in households already surveyed, the evolution of households composition makes the operation computationally difficult. This is the case of one SFSO survey for which a third of the sample was selected on the register created at the end of 2010 and the two other thirds are to be selected on the March 2011 and June 2011 registers.

We adopted the following procedure:

1. first a sample is selected through the coordination algorithm, with a set of selection probabilities that has to be computed so as to obtain the correct final inclusion probabilities,

2. then, in households with multiple selections, one of the selected units is kept at random and the others are eliminated from the sample,

3. if applicable, selections in households that were part of previous samples are filtered.

With this procedure, some selections are erroneously recorded in the coordination system, as some units selected by it will not be part of the survey. But the inclusion probabilities used at the coordinated sampling phase are close to the true final inclusion probabilities, except in very large households, and so, the system retains an acceptable performance.

When only steps 1 and 2 have to be performed, parameters p_k of the first phase selection of units k in a household $\mathcal{M}$ of size m_k can be computed as functions of the final inclusion probabilities π_k by noting that

$$\pi_k = \frac{1}{m_k}\left[1-(1-p_k)^{m_k}\right]. \tag{5}$$

We then get that we need to use selection probabilities $p_k = 1-(1-m_k\pi_k)^{\frac{1}{m_k}}$. In the general case of unequal inclusion probabilities, we should solve equations (6) in p_k, $k \in \mathcal{M}$:

$$\pi_k = p_k \sum_{n=0}^{m_k-1} \sum_{i_1\neq\cdots\neq i_n\neq k\in\mathcal{M}} \frac{1}{n+1} \prod_{j\in 1,\ldots,n} p_{i_j} \prod_{\ell\notin i_1,\ldots,i_n,k} (1-p_\ell)\,,\ k\in\mathcal{M}. \tag{6}$$

Unfortunately, there is usually no closed form solution for these equations. Numerical procedures seem to work on some toy examples, but it is unsure whether they are fast and convenient enough to be used on a population of about eight million units.

If step 3 has to be performed, an equation similar to 6 is relatively easy to obtain. However, it depends on selection probabilities of units of the household at former sampling occasions. If the composition of the household has changed, there is a strong possibility that these probabilities are not equal among all members of the household. In that case, we do not have closed form solutions to select the sample with prescribed inclusion probabilities. Luckily enough the household structure is not evolving so quickly. Indeed, when we compare the September and December 2010 registers, we

observe that this mixing of inclusion probabilities within households has occurred for less than 2% of the households. For the other 98% we are able to select a sample that exactly respects planified inclusion probabilities.

REFERENCES

Brewer, K. R. W., Early, L. J. & Joyce, S. F. (1972). Selecting several samples from a single population. *Australian Journal of Statistics* **3**, 231–239.

David, F. & Johnson, N. (1956). Reciprocal bernoulli and poisson variables. *Metron* **18**, 77–81.

Deville, J.-C. & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376–382.

Grab, E. & Savage, I. (1954). Tables of the expected value of 1/x for positive bernoulli and poisson variables. *Journal of the American Statistical Association* **49**, 169–177.

Hájek, J. (1971). Discussion of an essay on the logical foundations of survey sampling, part on by d. basu. In *Foundations of Statistical Inference*, V. P. Godambe & D. A. Sprott, eds. Toronto, Canada: Holt, Rinehart, Winston.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

Marciniak, E. & Wesolowski, J. (1999). Asymptotic eulerian expansions for binomial and negative binomial reciprocals. *Proceedings of the American Mathematical Society* **127**, 3329–3338.

Nedyalkova, D., Qualité, L. & Tillé, Y. (2009). Tirages coordonnés d'échantillons à entropie maximale. Technical report, University of Neuchâtel.

Qualité, L. (2009). *Unequal probability sampling and repeated surveys.* Thèse de doctorat, Université de Neuchâtel, Neuchâtel, Suisse.

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

Thionet, P. (1963). Sur le moment d'ordre (-1) de la distribution tronquée. application à l'échantillonnage de hájek. *Publ. Inst. Statist. Univ. Paris 12* **31:827**, 93–102.

RÉSUMÉ

L'Office Fédéral de la Statistique (Suisse), utilise un système de sélection d'enquêtes coordonnées développé par Qualité (2009). Il s'agit d'une extension de la méthode de Brewer et al. (1972). au cas de plus de deux enquêtes. Chaque échantillon transversal sélectionné avec ce système provient d'un plan de Poisson. La variabilité de la taille des échantillons qui en découle nécessite de nouvelles procédures d'allocation des tailles d'échantillons visées dans des domaines pour remplacer celles jusque là utilisées pour des plans stratifiés. Un problème particulier introduit avec ces plans de Poisson, et qui n'était pas présent avec les plans stratifiés de taille fixe, est le risque de tirer un échantillon dont la taille est bien en dessous de ce qui avait été visé dans certains domaines. Ce risque est en fait toujours présent lorsque la non-réponse est considérée comme une deuxième phase de sondage par un plan bernoullien dans des domaines. Un autre problème pour lequel nous présentons notre solution actuelle est l'adaptation du système aux enquêtes pour lesquelles une dépendance dans la sélection des unités est exigée (par exemple pour les enquêtes où l'on ne veut pas sélectionner plus d'une unité par ménage). Ces enquêtes ne peuvent être directement implémentées avec notre système, mais nous montrons comment nous l'utilisons comme une des phases de tirage dans un plan en plusieurs phases qui satisfait les contraintes voulues. Le calcul des probabilités d'inclusion conditionnelles à chaque phase pour obtenir les probabilités finales voulues est cependant en général non trivial.