Developments on Coordinated Poisson Sampling

Lionel Qualité

E-mail: lionel.qualite@bfs.admin.ch

Swiss Federal Statistical Office, Neuchâtel, Switzerland

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. Up to now, the effect of this coordinated sampling system on the number of repeated selection of businesses has been modest. Still, it has permitted to easily select a new sample after a major redesign in the survey of value added, which is a rotating panel type of survey, and to have a unified frame and rapid access to selection history of businesses.

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).

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 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.

Fig. 1: positive coordination when $\pi_k^2 \leq \pi_k^1$

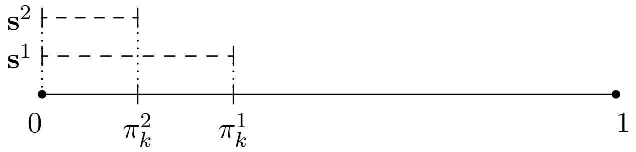
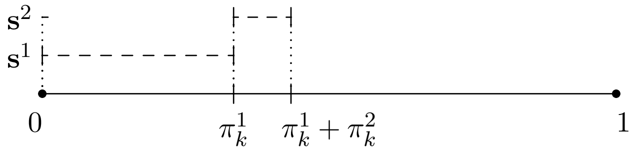
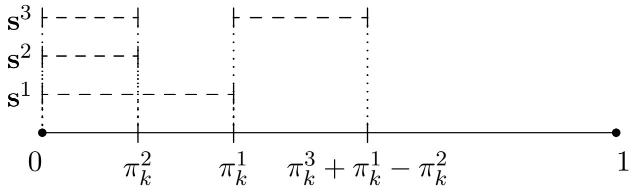


Fig. 2: negative coordination when $\pi_k^1 + \pi_k^2 \leq 1$



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.

Fig. 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: businesses and local units in the first case, individuals and households in the second. The independence between selections prevents us from selecting coordinated surveys at both levels with coordination between both kinds of surveys.

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

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

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

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

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,

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

simplifies to

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

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

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

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. Up to now, this risk has been taken care of by bluntly and arbitrarily raising the minimum target size in the intersection of activity domain and size class.

3.2 Effects of coordinated sampling on units burden

As positive coordinations are required as well as negative negative coordinations, monitoring the performance of our coordination system is not an easy task. Moreover, the presence of take-all strata in some surveys implies that the absolute number of repeated selections may not be sufficient to evaluate the system, even if only negative coordinations were required. Indeed, some of these multiple selections may be forced, and when it is not the case, coordination is ineffective in a stratum that is exhaustively selected at one survey and not at the other: a sample slightly larger than expected in such strata induces a larger number of repeated selections than expected under independent sampling. Tables 1 and 2 illustrate these problems. In table 1, we find the number of units by number of actual selections and size class after four sampling occasions, all negatively coordinated, as well as by the number sampling occasions were these units received an inclusion probability equal to one. We can see in table 1 that the effect of coordination on businesses with up to 49 employees was to increase the number of units that are selected once and decrease the number of businesses that are selected more than once, if we compare it with independent sampling. For businesses of 50 employees or more, the system had no impact whatsoever on the burden spread. There were actually slightly few more triple selections than would be expected with independent sampling, probably due to the size variability of the Poisson sampling design. In table 2, we find the situation after seven sampling occasions, one of which is the renewed value-added rotating panel. Here we cannot evaluate anymore the system performance by comparing coordinated and independent sampling since some of the repeated selections are actually sought after. And indeed, for businesses of less than 50 employees, the number of repeated selections has been larger than under independent sampling. For large businesses, we cannot conclude anything. Only a careful examination of our two value-added samples permitted us to confirm that the system did what we wanted: it selected overlapping samples as well as possible in spite of the major redesign that occurred between their selection.

		Number of selections (actual/exp. independent/forced)				
Nb. employ.	Total	0	1	2	3	4
0-2	305344	305284/305268/305330	54/71/12	6/5/2	0/0/0	0/0/0
3-9	147076	137710/137910/146601	9335/9026/473	30/138/2	1/1/0	0/0/0
10-19	29231	26298/26300/28672	2905/2866/552	27/64/6	1/1/1	0/0/0
20-49	16305	9752/10121/13755	5220/4582/2195	1327/1536/354	6/67/1	0/0/0
50-99	5494	1419/1415/1458	1905/1914/2037	1900/1901/1992	270/263/7	0/0/0
100+	4959	1011/1018/1272	1142/1126/1366	2388/2399/1941	417/415/379	1/1/1
Total	508409	481474/482033/497088	20561/19585/6635	5678/6043/4297	695/746/388	1/1/1

Tab. 1: Number of selections after 4 surveys: actual/expected under independent sampling/forced.

	Number of selections (actual/exp. independent)							
Nb. employ.	0	1	2	3	4	5	6	7
0-2	305284/305268	54/71	6/5	0/0	0/0	0/0	0/0	0/0
3-9	134616/132257	7605/12610	4794/2064	60/142	1/3	0/0	0/0	0/0
10-19	21978/21743	5488/5982	1532/1194	225/274	8/37	0/1	0/0	0/0
20-49	7770/8013	4324/4227	2808/2422	1033/1148	299/401	71/94	0/0	0/0
50-99	1179/1189	245/232	1682/1700	1597/1607	687/655	103/111	1/1	0/0
100+	859/895	409/351	834/810	952/1024	1170/1133	717/723	18/22	0/0
Total	471686/469365	18125/23472	11656/8195	3867/4195	2165/2230	891/929	19/24	0/0

Tab. 2: Number of selections after 7 surveys: actual/expected under independent sampling.

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.