Coordination of Conditional Poisson Samples

Anton Grafström [1] & Alina Matei [2]

[1] *Department of Forest Resource Management, Swedish University of Agricultural Sciences, SE-90183, Umea, Sweden, Anton.Grafstrom@slu.se*
[2] *Institute of Statistics, University of Neuchâtel, Rue A.-L. Bréguet 2, Neuchâtel and Institute of Pedagogical Research and Documentation Neuchâtel, Switzerland, alina.matei@unine.ch*

Résumé. La coordination des échantillons a comme but de créer une dépendance probabiliste entre deux ou plusieurs échantillons tirés dans des populations finies qui se chevauchent. Cette dépendance maximise ou minimise la taille de l'échantillon commun, de façon à ce que les échantillons d'enquêtes différentes se recouvrent le plus possible ou le moins possible. Dans le premier cas on parle d'une coordination positive, et dans le deuxième cas d'une coordination négative. La coordination positive est surtout utilisée pour améliorer la qualité des estimations. A son tour, la coordination négative est employée principalement pour diminuer la charge de réponse des unités sélectionnées dans plusieurs échantillons. Nous montrons deux méthodes pour coordonner des échantillons de type Poisson conditionnels à la taille. Le plan de Poisson conditionnel à la taille a des propriétés théoriques importantes comme le fait de maximiser l'entropie dans la classe des plans de sondages ayant les mêmes probabilités d'inclusion. Les méthodes proposées sont évaluées en utilisant des simulations de Monte Carlo et sont comparées avec d'autres méthodes existantes dans la littérature.

Mots-clés. coordination des échantillons, taux de recouvrement, nombres aléatoires permanents, plans à probabilités inégales

1 Introduction

Consider two finite overlapping populations U_1 and U_2. Two sampling designs p_1 and p_2 of fixed size n_1 and n_2 are defined on these two populations, respectively. Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be the sets of all possible samples defined by p_1 and p_2 on U_1 and U_2, respectively. Samples defined on $\mathcal{S}_1$ are denoted s_{1i}, $i = 1, 2, \ldots, m$, while samples defined on $\mathcal{S}_2$ are denoted s_{2j}, $j = 1, 2, \ldots, q$. Our general notation for samples is $s_1 \in \mathcal{S}_1$ and $s_2 \in \mathcal{S}_2$. We note $\pi_{1k} = \sum_{s_1 \ni k, s_1 \in \mathcal{S}_1} p_1(s_1), k \in U_1$ and $\pi_{2k} = \sum_{s_2 \ni k, s_2 \in \mathcal{S}_2} p_2(s_2), k \in U_2$ the first-order inclusion probabilities of unit k in the two samples, respectively. For simplicity, let U be the union of U_1 and U_2. Thus, for units $k \in U \setminus U_1$, we set $\pi_{1k} = 0$, while for $k \in U \setminus U_2$, we set $\pi_{2k} = 0$. An overall sampling design p is defined on $\mathcal{S}_1 \times \mathcal{S}_2$, with marginals p_1 and p_2. The overall sampling design is said to be coordinated if $p(s_{1i}, s_{2j}) = p_{ij} \neq p_1(s_{1i})p_2(s_{2j})$,

i.e. if the two samples are not selected independently. The joint inclusion probability of unit k in s_1 and s_2 is denoted

$$\pi_k^{1,2} = P(k \in s_1, k \in s_2) = \sum_{\substack{s_{1i} \cap s_{2j} \ni k \\ s_{1i} \in \mathcal{S}_1, s_{2j} \in \mathcal{S}_2}} p_{ij}.$$

Let c_{ij} be the overlap size of samples s_{1i} and s_{2j} $c_{ij} = |s_{1i} \cap s_{2j}|$, where $|A|$ denotes the cardinality of a set A. In general, the overlap size c_{ij} is random. Let c denote the random variable called 'overlap size'. A measure of the coordination degree between two samples is given by $E(c) = \sum_{i=1}^{m} \sum_{j=1}^{q} c_{ij} p_{ij} = \sum_{k \in U} \pi_k^{1,2}$. In positive coordination, the goal is to maximize $E(c)$, while in negative coordination, we want to minimize it. Bounds for $E(c)$ exist. They are determined by the Fréchet bounds of the joint inclusion probabilities $\pi_k^{1,2}$

$$\sum_{k \in U} \max(0, \pi_{1k} + \pi_{2k} - 1) \leq E(c) \leq \sum_{k \in U} \min(\pi_{1k}, \pi_{2k}). \tag{1}$$

The the left side-part in (1) is called the Absolute Lower Bound (ALB) and the right side-part in (1) the Absolute Upper Bound (AUB). Ideally, in positive coordination we want to achieve the AUB, while in negative coordination the ALB. Few methods achieve these bounds. In positive coordination, Poisson sampling with permanent random numbers (PRN; Brewer, Early, and Joyce, 1972) applied in both selections provides an important property: $\pi_k^{1,2} = \min(\pi_{1k}, \pi_{2k})$, and thus the AUB is reached. The sample sizes are, however, random for s_1, s_2 and $s_1 \cap s_2$.

While all sample coordination methods seek to increase or decrease the sample overlap, there are different ways to measure the effectiveness of the positive or negative coordination (e.g. the size of the expected overlap or the expected load of a unit which is defined as the sum of its selection probabilities in the surveys). Consequently, there is not a unique definition of optimality in sample coordination. We focus on methods which try to reach the AUB.

We consider the coordination of Conditional Poisson (CP) samples (or maximum fixed-size entropy samples) over time or simultaneously. CP-sampling has an important property: it maximizes the entropy in the class of fixed-size πps designs with the same inclusion probabilities. This property has important consequences on the sample selection randomness, on the variance estimation and on the convergence to a normal distribution of the Horvitz-Thompson estimator. Methods to coordinate CP-samples have not yet been introduced in the literature. We proposed two methods. The methods are evaluated using the size of the expected sample overlap, and are compared with their competitors. We focus on positive coordination, but negative coordination is also possible using the proposed methods.

2 Conditional Poisson sampling

Conditional Poisson sampling is a fixed-size πps sampling design. It was introduced by Hájek (1964) as a modification of the classical Poisson sampling. Different implementations of CP-sampling are available (see e.g. Tillé, 2006 and Bondesson, Traat, and Lundqvist, 2006). The initial implementation of CP-sampling given by (Hájek, 1964, 1981) uses a rejective algorithm to obtain a sample of size n as follows. Draw Poisson samples with parameters $0 \leq p_k \leq 1$, $k = 1, 2, \ldots, N$ until we get a sample of size n, i.e. we condition the Poisson design on the fixed sample size n. Usually it is assumed that $\sum_{k=1}^{N} p_k = n$ because it maximizes the probability of obtaining samples of size n. The assumption $\sum_{k=1}^{N} p_k = n$ is, however, not restrictive. If it is not satisfied, the p_ks can be transformed to satisfy that condition, see e.g. Broström and Nilsson (2000) or Tillé (2006), p. 89. Assume that $\sum_{k=1}^{N} p_k \neq n$, then transformed parameters p'_k, $k = 1, 2, \ldots, N$, with sum n can be calculated. As long as $\frac{p'_k}{1-p'_k} \propto \frac{p_k}{1-p_k}$, the design remains unchanged. We can let $p'_k/(1-p'_k) = dp_k/(1-p_k)$, which imply $p'_k = \frac{dp_k}{1-p_k+dp_k}$, and then find d such that $\sum_{k=1}^{N} p'_k = n$.

When implementing CP-sampling of size n with parameters p_k, $\sum_{k=1}^{N} p_k = n$, the true inclusion probabilities will only approximately equal the p_ks. The first and second-order inclusion probabilities for CP-sampling can be calculated recursively from p_k.

It is also possible to adjust the p_ks to obtain desired inclusion probabilities using an iterative algorithm (Aires, 2000). Let $\pi_k^{CP(n,t)}$ be the achieved inclusion probabilities with the parameters p_k^t for CP-sampling with sample size n, where t denotes the current iteration of the algorithm, and let $p_k^0 = \pi_k$. Then, practically, only a few iterations of

$$p_k^t = p_k^{t-1} + (\pi_k - \pi_k^{CP(n,t-1)}), \tag{2}$$

is enough to find parameters p_k^t that yield inclusion probabilities π_k.

3 Coordination of CP-samples using list-sequential implementation

List-sequential implementations of CP-sampling can be found in e.g. Chen and Liu (1997), Traat, Bondesson, and Meister (2004) and Tillé (2006). The units are sampled list-sequentially with start from unit 1. Unit k should be selected in the sample with an updated probability, here denoted by $\pi_k^{(k-1)}$. Thus, we select the unit k in the sample if $r_k \leq \pi_k^{(k-1)}$, where r_k is a random number from $U(0,1)$. The random number r_k may be a permanent random number for unit k (and it will be used in all coordination process). We assume that $r_1, \ldots, r_k, \ldots, r_N$ are independent.

Let $I_k \sim Bin(1, p_k)$, $k = 1, 2, \ldots, N$ be independent random variables, where p_ks are the Poisson parameters and $\sum_{k \in U} p_k = n$. The updated probabilities can be calculated

as follows $\pi_k^{(k-1)} = P\left(I_k = 1 \,|S_k = n - n_{k-1}\right)$, where $S_k = \sum_{\ell=k}^{N} I_\ell$, $n_k = \sum_{\ell=1}^{k} I_\ell$, and $n_0 = 0$. The updated probabilities can be rewritten as $\pi_k^{(k-1)} = p_k \cdot \frac{P(S_{k+1}=n-n_{k-1}-1)}{P(S_k=n-n_{k-1})}$, where $S_{N+1} = 0$. The probabilities $P(S_k = a)$ for given k and a can easily be calculated recursively. The start is given by $P(S_N = 0) = 1 - p_N$ and $P(S_N = 1) = p_N$. Then, for $k = N-1, N-2, \ldots, 1$ and $a = 0, 1, \ldots, N-k+1$, we have $P(S_k = a) = p_k P(S_{k+1} = a-1) + (1-p_k)P(S_{k+1} = a)$, if $a > 0$, and $P(S_k = a) = (1-p_k)P(S_{k+1} = a)$, if $a = 0$. If the population is very large, the recursions may take some time. Using this method we can calculate the updated probabilities $\pi_k^{(k-1)}$, for $k = 1, 2, \ldots, N$, and directly get a sample.

To coordinate two CP-samples with inclusion probabilities π_{1k} and π_{2k}, $k = 1, 2, ..., N$, we use the algorithm given by Expression (2) to find the corresponding Poisson parameters p_{1k} and p_{2k}, respectively. We then apply the list-sequential method with the permanent random numbers r_k in each selection. Even though it is logical to try to coordinate CP-samples in this manner, the approach seems to be new. In fact, any design with a list-sequential implementation can easily be coordinated by the use of PRN.

Remark 1 *Negative coordination can be achieved using the list-sequential method. For negative coordination of two samples, antithetic random numbers $r_k^* = 1 - r_k$ can be used in the second selection. For $\beta > 2$ samples, new random numbers can be constructed by shifting the PRN an amount α to the right before the selection of each sample different from the first one: $r_k + \alpha$. A possible choice of α is the inverse of the number of samples to coordinate (see Ohlsson, 2000). If $r_k + \alpha$ is larger than 1, we can replace it by $(r_k + \alpha)$* mod 1, *where mod is the modulo operator.*

4 An approximative method to coordinate CP-samples

Here we suggest a new approximative method to coordinate two CP-samples that does not use permanent random numbers, but instead uses updated parameters for the second selection. In the first selection, we select a CP-sample s_1 of size n_1 with inclusion probabilities π_{1k}, $k = 1, 2, ..., N_1$, using any suitable method to obtain a CP-sample.

In the second selection we select a CP-sample s_2 of size n_2 with updated parameters $p_{2k|s_1}$, $k = 1, 2, ..., N_2$. If $\pi_{1k} \leq \pi_{2k}$, then

$$p_{2k|s_1} = \begin{cases} 1 & \text{if } k \in s_1 \\ (\pi_{2k} - \pi_{1k})/(1 - \pi_{1k}) & \text{if } k \notin s_1 \end{cases},$$

and if $\pi_{1k} > \pi_{2k}$, then

$$p_{2k|s_1} = \begin{cases} \pi_{2k}/\pi_{1k} & \text{if } k \in s_1 \\ 0 & \text{if } k \notin s_1 \end{cases}.$$

The updated parameters are only used for units $k \in U_1$; for new units $k \notin U_1$, we let $p_{2k|s_1} = \pi_{2k}$.

If we could achieve inclusion probabilities equal to these parameters, we get the prescribed inclusion probabilities π_{2k}. Then it also follows that $\pi_k^{1,2} = \min(\pi_{1k}, \pi_{2k})$ and the expected overlap is maximized (the AUB is achieved). However, the parameters $p_{2k|s_1}$ cannot be used as inclusion probabilities for a fixed size design because they do not in general sum to n_2 for a given s_1. Only the sum of the expected value of the $p_{2k|s_1}$ equals n_2. Thus it is impossible to achieve inclusion probabilities equal to these parameters for a given s_1 if only samples of size n_2 are accepted. If a rejective implementation of CP-sampling is used, we can maximize the probability of obtaining a sample of size n_2 by using transformed parameters with sum n_2 (see Section 2).

Some situations may arise where it is impossible to draw a sample s_2 using the parameters $p_{2k|s_1}$. Consider, for example, the case where $N = 6, n_1 = n_2 = 3$, $\boldsymbol{\pi}_1 = (0.3, 0.3, 0.3, 0.7, 0.7, 0.7)'$, $\boldsymbol{\pi}_2 = (0.4, 0.4, 0.4, 0.4, 0.4, 1)'$ and the sample s_1 is $\{1, 2, 3\}$. The parameters $p_{2k|s_1}$ are $(1, 1, 1, 0, 0, 1)'$, and they do not allow the selection of a sample s_2 of size 3. In these unlikely situations, it will be impossible to achieve a CP-sample of size n_2 using the parameters $p_{2k|s_1}$ because either more than n_2 of the parameters equal 1 or more than $N_2 - n_2$ equal 0. In such cases, we suggest the following modification to the parameters. If there are more than n_2 of the $p_{2k|s_1}$ that equal 1, we use

$$p^*_{2k|s_1} = \begin{cases} 0 & \text{if } p_{2k|s_1} < 1 \\ 1 & \text{if } \pi_{2k} = 1 \\ \frac{n_2 - |\{j:\pi_{2j}=1\}|}{|\{j:p_{2j|s_1}=1,\, \pi_{2j}<1\}|} & \text{otherwise} \end{cases}.$$

If there are more than $N_2 - n_2$ of the $p_{2k|s_1}$ that equal 0, we use

$$p^*_{2k|s_1} = \begin{cases} 1 & \text{if } p_{2k|s_1} > 0 \\ 0 & \text{if } \pi_{2k} = 0 \\ \frac{n_2 - |\{j:p_{2j|s_1}>0\}|}{|\{j:p_{2j|s_1}=0,\, \pi_{2j}>0\}|} & \text{otherwise} \end{cases},$$

where $|\{\cdot\}|$ is the size of $\{\cdot\}$. The $p^*_{2k|s_1}$ sum to n_2.

Remark 2 *The second proposed method is an approximative one because the second sampling design is not exactly respected. For small populations and samples, there are differences between the inclusion probabilities provided by the proposed sampling design in the second occasion and those of the corresponding CP-sampling. The first and second-order inclusion probabilities of the proposed design are denoted by $\widetilde{\pi}_{2k}$ and $\widetilde{\pi}_{2k\ell}$, respectively. However, they cannot be directly computed. The probabilities $\widetilde{\pi}_{2k}$ and $\widetilde{\pi}_{2k\ell}$ are estimated through simulation and are denoted by $\widehat{\widetilde{\pi}}_{2k}$ and $\widehat{\widetilde{\pi}}_{2k\ell}$, respectively. Simulation not shown here suggest that the differences between the inclusion probabilities of the proposed design and the prescribed inclusion probabilities π_{2k} and $\pi_{2k\ell}$ seem to vanish as the population and sample size increase. For example, for a population of size $N = 5$ and $n_1 = n_2 = 2$, the highest absolute difference between $\widehat{\widetilde{\pi}}_{2k}$ and π_{2k} is about 0.02, while for the second-order*

inclusion probabilities $\widehat{\widehat{\pi}}_{2k\ell}$ *and* $\pi_{2k\ell}$ *is about 0.07. For a population of size* $N = 1000$ *and* $n_1 = 100, n_2 = 250$, *the highest absolute difference between* $\widehat{\widehat{\pi}}_{2k}$ *and* π_{2k} *is about 0.0015; the highest absolute difference between* $\widehat{\widehat{\pi}}_{2k\ell}$ *and* $\pi_{2k\ell}$ *is about 0.0016.*

Performed simulations not shown here indicate that the estimators based on the proposed design in the second scheme and using the prescribed inclusion probabilities π_{2k} *and* $\pi_{2k\ell}$ *do not suffer from much larger variance than those computed using CP-sampling. A slight bias is present in estimations for small populations and samples, but its values were small in our simulations.*

5 Examples

To check the coordination performance of the two proposed methods, we also consider Poisson sampling with PRN (Brewer, Early, and Joyce, 1972) and Pareto sampling with PRN (Rosén, 1997a, 1997b). Two simulation studies are shown below using the following five different sampling schemes: a) two CP-samples are drawn independently (IND) (using the rejective method for both); b) two Poisson samples are drawn using Poisson sampling (POI) with PRN; c) two Pareto samples are drawn using Pareto sampling (PAR) with PRN; d) two CP-samples are drawn using the list-sequential method (SEQ) with PRN; e) the first sample is a CP one drawn using the rejective method; the second one is selected using the rejective method with updated parameters as described in Section 4. We call this method the mixed one (MIX).

For the methods a), d) and e) (only for the first sample) the parameters p_1 and p_2 were computed from π_{1k} and π_{2k} respectively and used in sample selection. A number of 10^5 simulation runs was used to compute the expected overlap of two samples drawn using the five methods. N random numbers from $U(0,1)$ distribution were generated in each simulation, and used as PRN in each method. The expected overlap for each method was computed using the formula $E_{sim}(c) = \frac{1}{m}\sum_{\ell=1}^{m} c_\ell^{1,2}$, where $m = 10^5$ is the number of runs, $c_\ell^{1,2} = |s_{1\ell} \cap s_{2\ell}|$, and $s_{1\ell}$, $s_{2\ell}$, are the samples drawn in the ℓ^{th} run of the simulation. The Monte Carlo variance of the overlap between samples was also reported in the tables below $V_{sim}(c) = \frac{1}{m-1}\sum_{\ell=1}^{m}(c_\ell^{1,2} - E_{sim}(c))^2$.

Example 1: We consider the well known MU284 population. Changes in population are assumed. We consider the region 2 from the MU284 population as current stratum, where 50% of the units are new in the second occasion (births), and 50% of the units change the stratum (deaths). We have considered that the births units were initially in the third stratum (the third region). Thus, 24 units have been randomly drawn from the third stratum using simple random sampling without replacement; these units represent the births for the second stratum. The number of persistent units in the two occasions is 24. The overall population is formed by the persistents, births and deaths; its size is 72. Samples of expected sizes $n_1 = 10, n_2 = 6$ respectively are drawn from this population of size $N = 72$. Table 1 shows the expected overlap and variance for each method. The

mixed method shows an expected overlap equal to the theoretical AUB, and provides better performance than Pareto sampling with PRN. The sequential method performs worse than Pareto sampling concerning the expected overlap. The differences for $V_{sim}(c)$ in Table 1 (except IND and POI) may seem too small to be practically significant.

Example 2: We consider an extreme situation for the mixed method. It is a case where it is not always possible to directly draw a sample s_2 using the parameters $p_{2k|s_1}$. Instead a sample s_2 is selected using the parameters $p^*_{2k|s_1}$. We have $N = 6, n_1 = n_2 = 3$, $\boldsymbol{\pi}_1 = (0.3, 0.3, 0.3, 0.7, 0.7, 0.7)'$, $\boldsymbol{\pi}_2 = (0.4, 0.4, 0.4, 0.4, 0.4, 1)'$. Table 1 gives the expected overlap and variance for each method. The mixed method shows an expected overlap larger than the AUB because the first-order inclusion probabilities are not exactly respected for the second design. The estimated first-order inclusion probabilities ($\widehat{\widetilde{\pi}}_{2k}$) for the second design are the following: 0.3722, 0.3729, 0.3727, 0.4417, 0.4405, 1.0000. By computing $\sum_{k\in U} \min(\pi_{1k}, \widehat{\widetilde{\pi}}_{2k})$ we obtain the expected overlap of the mixed method to be 2.4822, matching the value of $E_{sim}(c)$ for the mixed method given in Table 1.

Table 1: Expected overlap and variance based on 10^5 simulation runs

	Example 1		Example 2	
Method	$E_{sim}(c)$	$V_{sim}(c)$	$E_{sim}(c)$	$V_{sim}(c)$
IND	1.55	0.78	1.62	0.44
POI	2.79	1.94	2.40	1.33
PAR	2.76	1.04	2.33	0.32
SEQ	2.55	1.00	2.32	0.35
MIX	2.79	0.99	2.48	0.30
AUB	2.79		2.40	

6 Conclusions

The first method is based on the list-sequential implementation of CP-sampling. It is a PRN method and has the advantage to preserve exactly the second sampling design (both samples are CP). It provides a good level of expected overlap as shown in our examples, but smaller than the AUB. This is mainly due to the differences between selection and inclusion probabilities.

The second method is an approximative one because the second sampling design is not exactly respected. For small populations and samples, there are differences between the inclusion probabilities provided by the proposed sampling design in the second occasion and those of the corresponding CP-sampling. In our examples, these differences seem to vanish as the population and sample size increase.

The mixed method shows high performance comparable to Poisson sampling with PRN. It has the advantage of allowing fixed sample sizes comparing to Poisson sampling with PRN. Due to this fact, the mixed method provides smaller overlap variance than

Poisson sampling with PRN, as also shown in our simulations. Compared to Pareto sampling with PRN, the mixed method performs better in simulations, but it has the disadvantage of providing only an approximate CP-sample in the second selection. On the other hand, Pareto sampling does not possess the maximum entropy property for given first-order inclusion probabilities.

Based on the criterion to achieve the AUB, the second sampling in the mixed method is an optimal sampling design for the first one. In our paper, the first sample is a CP-sample. It is possible to apply the mixed method for any type of fixed-size sampling design used in the first selection. Hence, the method allows to use e.g. a balanced sample in the first selection.

Bibliography

[1] Aires, N. (2000). *Techniques to calculate exact inclusion probabilities for conditional Poisson sampling and Pareto πps sampling designs.* Doctoral thesis, Chalmers University of technology and Göteborg University, Göteborg, Sweden.

[2] Bondesson, L., Traat, I., and Lundqvist, A. (2006). Pareto sampling versus Sampford and conditional Poisson sampling. *Scandinavian Journal of Statistics*, 33:699–720.

[3] Brewer, K., Early, L., and Joyce, S. (1972). Selecting several samples from a single population. *Australian Journal of Statistics*, 3:231–239.

[4] Broström, G. and Nilsson, L. (2000). Acceptance-rejection sampling from the conditional distribution of independent discrete random variables, given their sum. *Statistics*, 34:247–257.

[5] Chen, S. and Liu, J. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica*, 7:875–892.

[6] Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35:1491–1523.

[7] Hájek, J. (1981). *Sampling from a Finite Population.* Marcel Dekker, New York.

[8] Ohlsson, E. (2000). Coordination of pps samples over time. In *Proceedings of the Second International Conference on Establishment Surveys*, pages 255–264. American Statistical Association.

[9] Rosén, B. (1997a). Asymptotic theory for order sampling. *Journal of Statistical Planning and Inference*, 62:135–158.

[10] Rosén, B. (1997b). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, 62:159–191.

[11] Traat, I., Bondesson, L., and Meister, K. (2004). Sampling design and sample selection through distribution theory. *Journal of Statistical Planning and Inference*, 123:395–413.

[12] Tillé, Y. (2006). *Sampling algorithms.* Springer series in statistics, Springer science + Business media, Inc., New York.