European Establishment Statistics Workshop
Efficient methodology for producing high quality establishment statistics

Stockholm, Sweden, 6-9 September 2009

MEASURING THE PERFORMANCE OF A SAMPLE COORDINATION SYSTEM

Paul-André Salamin
Statistical methods unit
Swiss Federal Statistical Office

1 Introduction

A general program of enterprise statistics is one component of the modernization strategy being currently implemented at the Federal Statistical Office (FSO). This program aims at reforming the business statistics and at developing a coherent and integrated system of business surveys.

Burden reduction is one of the goals of this program. Burden is defined here as the number of surveys in which an enterprise has to participate over a given period, e.g. one calendar year. As a rough measure of burden we can consider the overall sample size of the surveys carried out during one year, divided by the total number of enterprises. This quantity can be interpreted as the *mean expected burden per enterprise*. More refined measures of burden would be obtained by considering enterprises in different types of economic activity, sizes classes, etc. Burden reduction can be achieved through a better use of administrative data. Indeed, given the surveys which have to be conducted during one year, burden reduction can only be achieved through reduced sample sizes. This entails a lost of efficiency, which may possibly be compensated by more efficient sampling plans and estimation procedures, provided that strong auxiliary information is available. Due to their importance, the large enterprises have to participate in all surveys. For large enterprises burden reduction cannot be achieved through sample size reduction but through profiling, e.g. personalized contacts with the enterprises, centralized administration of all the surveys in which they have to participate, etc.

Once all measures have been taken to reduce the burden, a National Statistical Institute has the further possibility of spreading this burden evenly over the units in the population. This is the goal of a sample coordination system. This paper addresses the issue of how to measure the performance of a sample coordination system. Indeed, the development of a sample coordination system is a fairly complex endeavour and one may legitimately ask what advantages such a system brings over independent selections of samples. The paper is organized as follows. In Section 2 we define the concepts of negative and positive coordination. In Section 3 we describe the selection algorithm chosen by the FSO for sample coordination. In Section 4 we address the issue of how to measure the performance of a sample coordination system. In Section 5 finally we discuss some of the impacts that a sample coordination system can have on the surveys.

2 Sample coordination

The current practice is to design and carry out each survey independently of the other surveys. Each survey has its own sampling frame and the sample is selected according to an optimized sampling plan, based on a detailed stratification. For each survey, the sample is selected independently of the samples already selected for other surveys. Thus the overlaps of the samples for different surveys are not controlled.

The problem of sample coordination is to select several samples according to given sampling plans, while controlling the sizes of the intersections of the samples. In practice the samples are selected sequentially. Thus, when selecting a sample at a given occasion, according to a given sampling plan, we want to control the sizes of its overlaps with the already selected samples.

Spreading the response burden corresponds to negative coordination, in which case the sizes of the intersections have to be minimized. If the total size of the samples is less than the size of the population, then it is possible to select disjoint random samples. Otherwise some overlap is unavoidable.

Most of the enterprise surveys are repeated surveys, where the same units have to be observed on several occasions. In this case a certain degree of overlap between samples is required. This can be achieved through a rotating panel: at each new selection of a sample some of the units are kept in the sample, some are eliminated and replaced by new units. In the most extreme case of positive coordination, as many units as possible are retained.

3 Algorithm for sample coordination

There are a number of methods for sample coordination which have been proposed in the literature, see e.g. Ohlsson [2] and Nedyalkova et al. [1]. The algorithm used at the FSO for coordinated sample selection is based on Poisson sampling and permanent random numbers. This method has been chosen for its theoretical simplicity, its ease of implementation and also for its ability to handle several rotating panels, negatively coordinated among them. The algorithm was developed, within the framework of a research convention, in collaboration with the Institute of Statistics of the University of Neuchâtel, see Qualité [3].

Each unit of the sampling frame receives a random number uniformly distributed between 0 and 1. This random number stays associated to the unit as long as it exists. For *each survey*, one defines for *each unit* a zone of selection. In the simplest case the zone of selection is an interval. In more complex situations the zone of selection can be the union of disjoint intervals. The total length of the zone of selection corresponds to the inclusion probability for that unit. Finally, a unit is selected in the sample if its permanent random number falls within its zone of selection. Different types of coordination (negative, positive, rotation) can be achieved by an appropriate choice of the zones of selection.

We consider as an example the selection of a unit in 6 samples. Table 1 gives the sampling fractions (f) for that unit in the 6 samples, as well as the types of coordination desired. In this example we have two panels: the samples 1, 3 and 6 are three waves of the panel 1 and the samples 2 and 5 are two waves of the panel 2. The sample 4 is for a survey conducted only once.

Here we consider only negative (N) or positive (P) coordination. Globally the two panels and the sample 4 have to be coordinated negatively. Thus, for example, sample 4 has to be coordinated negatively with the samples 1, 2 and 3. For a panel, the current wave has to be positively coordinated with the earlier waves. Thus, for example, the second wave of the panel 2 (sample 5), is positively coordinated with the first wave of the panel (sample 2), and negatively coordinated with all the other samples (samples 1, 3 and 4). Figure 1 shows the zones of selection associated to the 6 samples. Assuming that the permanent random number is equal to 0.42 for the unit under consideration, we see that the unit is selected in the waves 1 and 2 of the panel 2, and not selected in the samples 1, 3, 4 and 6.

				Coordination with				
Sample	f	Panel	Wave	1	2	3	4	5
1	0.30	1	1					
2	0.20	2	1	N				
3	0.40	1	2	P	N			
4	0.20			N	N	N		
5	0.30	2	2	N	P	N	N	
6	0.45	1	3	P	N	P	N	N

Table 1: Sampling fractions and types of coordination

4 Performance of sample coordination

The performance of a sample coordination algorithm can be measured by comparing the current practice of independent selections of samples with coordinated selections. We consider in this paper only the special case of global negative coordination in a population which is not changing.

It is important to note that there are characteristics of repeated sampling procedures which do not depend on the selection algorithm. These are the properties which depend on the first order inclusion probabilities only. One important example of such a property is the *mean expected burden per unit*. Specifically, we consider a population U of size N in which the samples $(S_t, t \in T)$ are selected. The burden for unit $k \in U$ is defined as $\beta_k = \sum_{t\in T} I_{kt}$ where $I_{kt} = 1$ if $k \in S_t$ and $I_{kt} = 0$ otherwise. Burden is thus a random variable which can take the values 0 to T. The expected burden for unit $k \in U$ depends only on the first order inclusion probabilites: $b_k = E\left(\beta_k\right) = \sum_{t\in T} \pi_{kt}$. The mean expected burden per unit is given by

$$\bar{b} = \frac{1}{N}\sum_{k\in U} b_k = \frac{1}{N}\sum_{t\in T}\sum_{k\in U}\pi_{kt}.$$

As $\sum_{k\in U}\pi_{kt} = E\left|S_t\right|$, the mean expected burden per unit can be seen as the overall expected sample size divided by the population size:

$$\bar{b} = \frac{1}{N}\sum_{t\in T} E\left|S_t\right|.$$

Thus, mean expected burden per unit stays the same, whatever method of coordination is chosen. This means that a sample coordination system cannot reduce the burden; it can only spread it as equitably as possible among the units in the population. Another parameter of repeated sampling, essentially equivalent to mean expected burden per unit, is mean expected total time out of sample per unit $T - \bar{b}$. As for total burden, this parameter stays the same, whatever method of coordination is chosen.

The effect of coordinating the samples is on the distribution of burden. In particular the variance of burden does depend on the coordination algorithm. For negative coordination with Poisson sampling, burden can take only two consecutive values, which depend on the mean expected burden per unit. One can show that the variance of burden is then bounded by 0.25. On the other hand, for independent selections of T samples, burden can take any of the values from 0 to T, and the variance of burden in this case is unbounded. Thus, the effect of negative coordination is to drastically reduce the spread of the distribution of burden. Indeed, consider the expected burden $b_k = \sum_{t\in T}\pi_{kt}$ for unit $k \in U$ and write

$$b_k = \lfloor b_k \rfloor + r_k = c_k + r_k \text{ where } c_k \in \mathbb{N} \text{ and } 0 \leq r_k < 1.$$

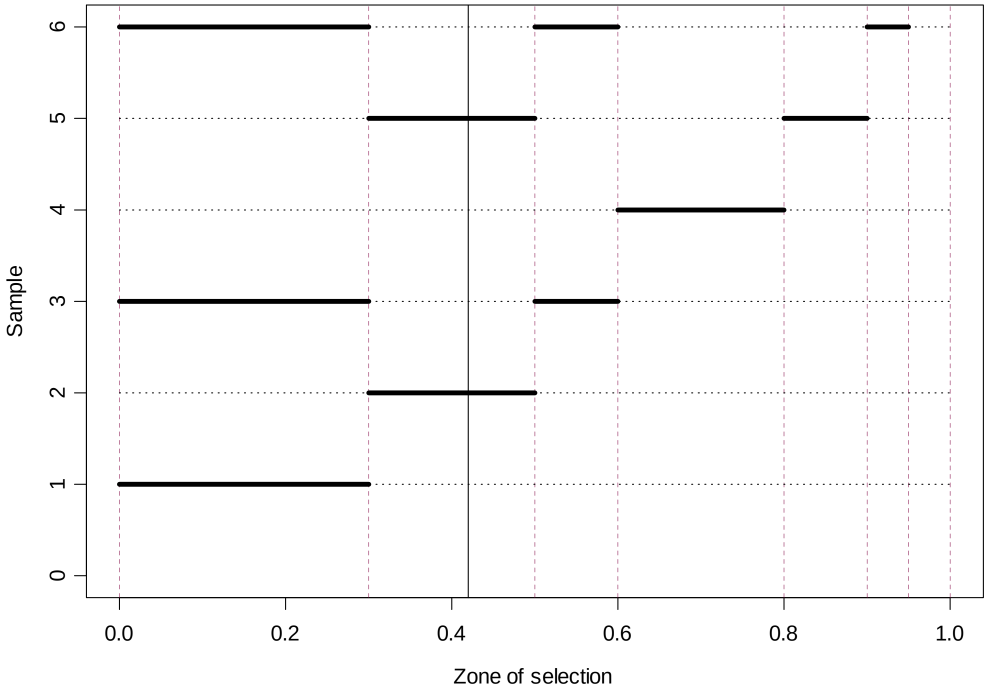

Figure 1: Zones of selection for sample coordination

It follows then that $\beta_k \in \{c_k, c_k + 1\}$ with

$$\beta_k = \begin{cases} c_k & \text{with probability } 1 - r_k \\ c_k + 1 & \text{with probability } r_k \end{cases}$$

from which we obtain

$$\mathrm{var}_{neg.coord}\left(\beta_k\right) = r_k\left(1 - r_k\right) \leq \frac{1}{4}.$$

We illustrate these properties by an example, see Table 2. We assume that at each occasion the sampling fraction is the same for all units in the population (Bernoulli sampling). In this case, the mean expected burden per unit is simply the sum of the sampling fractions.

With a mean expected burden per unit of 2.3, burden under negative coordination can only take the values 2 or 3, whereas for independent selections the burden can take the values from 0 to 5. Table 3 compares the distribution of burden under independent and coordinated

Sample	Sampling fraction
1	0.4
2	0.3
3	0.5
4	0.5
5	0.6
Mean expected burden per unit	2.3

Table 2: Sampling fractions for negative coordination

	Probability of Burden	
Burden	Independent selection	Negative coordination
0	0.04	
1	0.19	
2	0.34	0.70
3	0.29	0.30
4	0.12	
5	0.02	
Mean	2.3	2.3
Variance	1.19	0.21

Table 3: Probability distribution of burden

sampling. For both selection algorithms, the mean expected burden per unit is equal to 2.3, as was to be expected. The effect of spreading the burden is to concentrate the distribution of burden on the values 2 and 3. It follows that no unit has burden 4 or 5, but also that no unit has burden 0 or 1. Thus the mean expected burden of 2.3, which can be seen as an external parameter reflecting the selection of 5 samples with the indicated sampling fractions, is spread as equitably as possible. The concentration of burden distribution is also apparent in the variances: 1.19 for independent selection vs. 0.21 for negative coordination.

Total burden is still a rather coarse measure. One is often interested in the time between two selections in a sample. One has then to examine the different patterns of being in or out of sample. For the selection of 5 samples there are 32 such patterns, see Table 4. For example, pattern 00101 means that a unit is not selected in the samples 1, 2 and 4 and is selected in the samples 3 and 5. In the case considered here, see Table 2, global negative coordination is implemented by the following zones of selection

Sample										
1	X	X	X	X						
2					X	X	X			
3	X	X						X	X	X
4			X	X	X	X	X			
5	X	X	X					X	X	X
Pattern	1	1	2	3	4	4	4	5	5	5

For example, the zone of selection for the sample 3 is given $[0, 0.2) \cup [0.7, 1)$. Note that the length of the zone of selection is equal to the sampling fraction $f_3 = 0.5$ for the sample 3. The zones of selection define a decomposition of the interval $[0, 1)$ into disjoint subintervals, which can be indexed by patterns $a \in \{0, 1\}^T$. The length of the subintervals give the probabilities of the profiles. For the example we are considering we obtain the following patterns and probabilities, see also Table 4,

Pattern	a	$P(a)$
1	10101	0.2
2	10011	0.1
3	10010	0.1
4	01010	0.3
5	00101	0.3

Here also, all patterns can occur under independent selections, while under negative coordination we have a concentration on a small number of patterns.

It would be possible to define characteristics summarizing some properties of the profiles and to compute their distributions under independent or coordinated selections. For example

		Prob. Pattern				Prob. Pattern	
Pattern	Burden	Indep	Neg	Pattern	Burden	Indep	Neg
00000	0	0.042		11111	5	0.018	
00001	1	0.063		01111	4	0.027	
00010	1	0.042		10111	4	0.042	
00100	1	0.042		11011	4	0.018	
01000	1	0.018		11101	4	0.018	
10000	1	0.028		11110	4	0.012	
00011	2	0.063		00111	3	0.063	
00101	2	0.063	0.30	01011	3	0.027	
00110	2	0.042		01101	3	0.027	
01001	2	0.027		01110	3	0.018	
01010	2	0.018	0.30	10011	3	0.042	0.10
01100	2	0.018		10101	3	0.042	0.20
10001	2	0.042		10110	3	0.028	
10010	2	0.028	0.10	11001	3	0.018	
10100	2	0.028		11010	3	0.012	
11000	2	0.012		11100	3	0.012	

Table 4: Probability distribution of selection patterns

one could compute the distribution of the lengths of time between two selections in a sample, considering that longer periods out of sample are desirable.

It is important to realize that a sample coordination system cannot guarantee that a unit will stay out of sample for a given length of time. Given a number of surveys, if the sampling fractions are too high, the unit will have to be selected again, this being for example the case for the pattern 10011, occurring with probability 0.1 even under optimal negative coordination, see Table 4. All one can say is that small sampling fractions are desirable to make sample coordination worthwhile. With small sampling fractions one can have, for example, patterns with long periods out of sample, and the effect of coordination will be to concentrate the distribution on these good patterns.

5 Impact on surveys

The introduction of a sample coordination system has consequences on the sampling designs of surveys. We argued in the preceding Section that small sampling fractions are desirable. This can be achieved by small sample sizes, i.e. by designing efficient sampling plans making optimal use of the available auxiliary information. Once the sample size is fixed, one can still try to make the sampling fractions as small as possible by defining large strata. Of course, small sampling fractions can only be used for the population of small and medium size enterprises. Since we use Poisson sampling, sample size is random and cannot be strictly controlled. It must also be noted that techniques for coordinating samples of different types of units, e.g. enterprises and establishment, are at present not available. For the sample coordination system being implemented at the FSO, it has been decided to spread the burden at the enterprise level only. Thus, it appears that a sample coordination system imposes some constraints on the definition of the strata, the overall sample size and the sample sizes in the strata, the rotation rate for panels and the choice of the sampling unit.

One further aspect of the introduction of a sample coordination system is a shift in perspective from the planning of individual surveys to a global planning of several surveys. For example, one may consider parameters like mean expected burden per unit or mean expected time out of sample per unit as instruments for planning and designing an integrated system of enterprises surveys.

References

[1] Nedyalkova, D., Pea, J. and Tillé, Y. (2008a). Sampling Procedures for Coordinating Stratified Samples: Methods Based on Microstrata. Int. Stat. Rev., 76, 368-386.

[2] Ohlsson, E. (1995). Business Survey Methods. Vol. 1, chap. 9, "Coordination of samples using permanent random numbers", pp. 153-169. New York: Wiley, Inc.

[3] Qualité, L. (2009). Unequal Probability Sampling and Repeated Surveys. Thèse. Institut de Statistique. Université de Neuchâtel.