Poisson mixture sampling: A family of designs for coordinated selection using permanent random numbers

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

by Hannu Kröger, Carl-Erik Särndal and Ismo Teikari

June 1999

Survey Methodology, June 1999
Vol. 25, No. 1, pp. 3-11
Statistics Canada, Catalogue No. 12-001

Poisson mixture sampling: A family of designs for coordinated selection using permanent random numbers

Hannu Kröger, Carl-Erik Särndal and Ismo Teikari [1]

Abstract

This paper introduces Poisson Mixture sampling, a family of sampling designs so named because each member of the family is a mixture of two Poisson sampling designs, Poisson πps sampling and Bernoulli sampling. These two designs are at opposite ends of a continuous spectrum, indexed by a continuous parameter. Poisson Mixture sampling is conceived for use with the highly skewed populations often arising in business surveys. It gives the statistician a range of different options for the extent of the sample coordination and the control of response burden. Some Poisson Mixture sampling designs give considerably more precise estimates than the usual Poisson πps sampling. This result is noteworthy, because Poisson πps is in itself highly efficient, assuming it is based on a strong measure of size.

Key Words: Business surveys; Skewed populations; Response burden; Regression estimators.

1. The objectives of Poisson mixture sampling

Poisson Mixture (Pomix) sampling is a family of sampling designs suitable for business surveys with its often highly skewed populations. The Pomix family contains the traditional Bernoulli sampling and Poisson πps sampling designs as two special cases, situated at the two extremes of a range of possibilities indexed by a continuous parameter. This parameter, called the Bernoulli width and denoted B, satisfies $0 \leq B \leq f_R$, where f_R is the predetermined expected sampling fraction in the "take-some" portion of the population, that is, the portion where randomized selection is applied.

Random numbers, in the form of independent realizations of the Unif (0, 1) random variable, are commonly used in modern computerized sample selection. Fan, Muller and Rezucha (1962) introduced several sequential (unit by unit) drawing mechanisms based on random numbers. Now, Pomix sampling is based on the Permanent Random Number (PRN) technique, which calls for assigning at birth a random number to each unit in the frame (the business register, in the case of a business survey). The random number is permanent in the sense of remaining attached to the unit during its entire lifetime. The PRN technique makes it easy to achieve coordination of samples and control of response burden. Early references to sampling with the aid of PRN's are Brewer, Early and Joyce (1972) and Atmer, Thulin and Bäcklund (1975). A recent review of different PRN techniques, and important extensions, are given in Ohlsson (1995).

Poisson πps sampling has the desirable feature of selecting large units with relatively greater probability than small units, whose contribution to estimated population totals will in any case be relatively minor. Coordination of Poisson πps samples with the aid of PRN's was introduced by Brewer *et al.*, (1972) and is discussed subsequently by several authors, including Sunter (1977) and Ohlsson (1995).

Similarly as in Poisson πps, Pomix sampling allows control of the response burden, as explained in the next section. Larger enterprises will be selected relatively more often than smaller ones. The selection is controlled through rotation so as to distribute the response burden. Another objective of Pomix sampling is for all (or a substantial portion) of the population units to be included in sample (therefore observed, so that their basic data can be updated) with regularity over a period of time. The objective can be, for example, that every enterprise should be in sample at least once during a ten or twelve year period.

2. The selection procedure under Poisson mixture sampling

Denote the finite population as $U = \{1, ..., k, ..., N\}$, where the integer k represents the k^{th} population unit. Denote by y the variable of interest and by y_k its value for unit k; y_k is unknown before sample selection and observation. With the unit $k \in U$ is also associated a known positive size measure x_k. Its role in Pomix sampling is to bring about a more frequent selection of the larger units; in addition, the size variable should be used as an auxiliary variable at the estimation stage.

A sample, s, is realized from the population U. The size of s may be random; its expected size, denoted n, is a number fixed in advance. We allow s to consist of two nonoverlapping parts, $s = s_C \cup s_R$, where s_C is called the certainty part of s and s_R the randomization part of s. The part s_C, consisting of very large units selected with

[1] Hannu Kröger and Ismo Teikari, P.O. Box 3A, Fin-00022, Statistics Finland, Finland; Carl-Erik Särndal, 11th floor, R.H. Coats Bldg., Statistics Canada, Ottawa, Ontario, K1A 0T6, Canada.

probability one, is designated in a preliminary step, with the aid of the known size measures x_k. One procedure for this is given in section 3. Depending on the population characteristics, it could happen that no certainty part is designated, so that s_C is the empty set, but this eventuality is rather exceptional with the highly skewed populations usually occurring in business surveys.

A frequently used synonymous term for the certainty part is take-all stratum. If the take-all stratum is denoted U_C, a probabilistic description is to say that s_C is drawn from the take-all stratum U_C so that $s_C = U_C$ with probability one. We denote the size of $s_C = U_C$ by n_C.

Next, the randomization sample, s_R, is selected from the rest of the population, $U_R = U - U_C$, of size $N_R = N - n_C$. It consists of units with inclusion probability π_k strictly less than unity. In this paper, s_R is drawn by Pomix sampling (thus it uses the PRN technique). The size of s_R is random; its expected size, denoted n_R, is fixed by the equation $n_R = n - n_C$.

In this paper, we use the term Poisson sampling for selection corresponding to independent unit by unit Bernoulli trials with any inclusion probabilities π_k. More specifically, by Poisson πps we mean Poisson sampling with π_k directly proportional to a measure of size. Bernoulli sampling is the special case of Poisson sampling where all π_k are equal.

For Pomix sampling, we need some more notation. For unit $k \in U_R$, define the relative size measure

$$A_k = n_R x_k \Big/ \sum\nolimits_{U_R} x_k. \qquad (2.1)$$

We can from now on assume that $A_k < 1$ for all $k \in U_R$, because if $A_k < 1$ had not been true for certain units $k \in U_R$, then the procedure in section 3 for constructing the certainty part of the sample would in effect have assigned those units to the certainty part s_C.

We now define Pomix sampling with the aid of a two-dimensional diagram. On the horizontal axis, a unit's PRN is plotted. On the vertical axis, a size-related measure, Q_k, is plotted. At each survey occasion, a new sampling selection region is designated by rotation in this diagram, and sample coordination is realized in the manner that we now describe.

Pomix sampling is characterized by two parameters, B and f_R, where f_R such that $0 < f_R = n_R / N_R < 1$ denotes the fixed expected sampling rate in $U_R = U - U_C$. The parameter B, called the Bernoulli width, is such that $0 \le B \le f_R$. For every unit $k \in U_R$, define

$$Q_k = \frac{f_R - B}{1 - B}\frac{x_k}{\bar{x}_{U_R}} = \frac{1 - B/f_R}{1 - B}A_k \qquad (2.2)$$

where $\bar{x}_{U_R} = \sum_{U_R} x_k / N_R$. For $B = 0$, we have $Q_k = A_k$, which is the size measure for the usual Poisson πps sampling. At the other extreme, $B = f_R$, we have $Q_k = 0$ for all $k \in U_R$; in this case, size will play no role in the selection from U_R, which will be seen to reduce to Bernoulli sampling. The measures Q_k are used in Pomix selection of coordinated samples, as we now describe.

Start with a plot of the points (r_k, Q_k) for $k \in U_R$, where r_k denotes the PRN attached at birth to unit k, and Q_k is given by (2.2). With reference to Figure 1, Pomix sampling is defined as follows: Include in the randomization sample, s_R, all units having PRN's r_k falling in the $(0, B]$ interval, and also include some units having PRN's r_k in the $(B, 1]$ interval, namely, those for which Q_k is at least equal to a threshold value situated on the line joining the points $(B, 0)$ and $(1, 1)$. The selection area is thus the shaded part of Figure 1. Note that since $A_k < 1$ for all $k \in U_R$, we have by (2.2) that $Q_k < (1 - B/f_R)/(1 - B) \le 1$ for all $k \in U_R$, when $0 \le B < f_R$.

For Poisson πps sampling, the inclusion probability of unit k is $\pi_k = A_k$ given by (2.1). Coordination of PRN-based Poisson samples was introduced by Brewer *et al.*, (1972) using a graphical representation corresponding to $B = 0$ in Figures 1 and 2. At each occasion, the selection area is then a triangle; the unit's PRN on the horizontal axis is plotted against the unit's size measure, A_k, on the vertical axis. Coordination is obtained by "moving the selection area over" to the right. Coordination of Pomix samples is realized in a similar fashion.

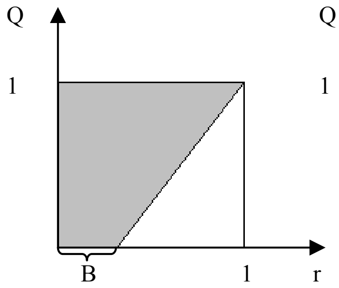

Figure 1 Sampling at time 1

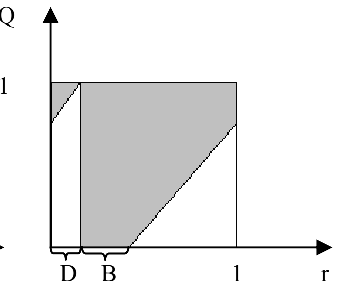

Figure 2 Sampling at time 2

Figures 1 and 2 illustrate how coordinated Pomix sampling from $U_R = U - U_C$ can be carried out at two consecutive survey occasions. In each of the two figures, the sample is defined as the set of units for which the point (r_k, Q_k) falls in the shaded area. The "starting point" on the PRN axis is the point to the right of which we start to count units for inclusion in the sample. At time 1 (Figure 1), the starting point is 0; at time 2 (Figure 2), the starting point is D. (In general, the starting point can be a randomly selected point in the unit interval; in other words, the sample identified in Figure 2 is also the one that would be selected at time $t = 1$ if at that time the randomly selected starting point on the PRN axis had been equal to D.)

A convenient way to achieve sample rotation is through the constant shift method, which implies that the starting point is moved over to the right by a fixed amount at every new occasion of sampling; see Ohlsson (1995). The constant D is called the constant shift. The starting point at time 3 would thus be $2D$, and so on.

In the following we examine Pomix sampling and estimation at a single occasion, and we can concentrate on Figure 1 (time 1), with starting point 0 on the PRN axis. The algorithm for Pomix sampling with parameters B and f_R, and starting point 0, is thus as follows: From Figure 1, unit k is included in the randomization part, s_R, (i) if $0 < r_k \le B$, or (ii) if $B < r_k \le 1$ and $Q_k \ge (r_k - B)/(1 - B)$.

Consequently, k is included in s_R if

$$0 < r_k \le B + Q_k(1 - B).$$

Because $r_k \sim \text{Unif}\,(0, 1)$, the first order inclusion probabilities under Pomix sampling are

$$\pi_k = \begin{cases} B + Q_k(1 - B) & \text{for } k \in U_R \\ 1 & \text{for } k \in U_C. \end{cases} \qquad (2.3)$$

It is easy to see that the inclusion probabilities satisfy the necessary requirement that their sum must equal the expected sample size fixed in advance:

$$\textstyle\sum_U \pi_k = \sum_{U_C} 1 + \sum_{U_R} \{B + Q_k(1 - B)\} = n_C + n_R = n.$$

We now note two extreme cases of the family of Pomix sampling schemes: Bernoulli sampling is obtained if $B = f_R$ in the Pomix algorithm, because then $Q_k = 0$ for all $k \in U_R$, and the algorithm becomes: Include unit $k \in U_R$ in s_R if $0 < r_k \le f_R$, which is Bernoulli sampling. Poisson πps sampling is obtained if $B = 0$ in the Pomix algorithm, because then the algorithm becomes: Include unit $k \in U_R$ in s_R if $0 < r_k \le A_k$, where A_k is given by (2.1). But this is Poisson πps sampling from U_R, the inclusion probability being $\pi_k = A_k$, that is, directly proportional to the size measure x_k.

Pomix sampling is a mixture of Poisson πps and Bernoulli in that the Pomix inclusion probability, $\pi_k = B + Q_k(1 - B)$, equals a linear combination of the inclusion probabilities that apply under the two extreme designs, weighted by the relative Bernoulli width, $\lambda = B/f_R$, such that $0 \le \lambda \le 1$. We have $\pi_k = \lambda\pi_k^{\text{BE}} + (1 - \lambda)\pi_k^{\pi\text{ps}}$, where $\pi_k^{\text{BE}} = f_R$ for all k (Bernoulli) and $\pi_k^{\pi\text{ps}} = A_k$ (Poisson πps).

The character of Pomix sampling is determined by its two parameters, B and f_R. To illustrate, we note from (2.1) and (2.3) that the inclusion probability of unit $k \in U_R$ is $\pi_k = B + (1 - B/f_R)A_k$. Thus, for a unit k that is large (but not large enough to qualify for $s_C = U_C$), so that A_k is near unity, we have, to close approximation, $\pi_k = B + 1 - B/f_R$. By contrast, for a unit that is small, so that its value A_k is very near zero, we have, to close approximation, $\pi_k = B$, independently of the size. For example, with $f_R = 10\%$, the following Table 1 shows how the inclusion probability π_k varies with B, where $B = 0$ is Poisson πps, and $B = f_R = 0.10$ is Bernoulli.

Table 1
Values of the inclusion probability π_k as a function of the parameter B and the relative size A_k when the fixed expected sampling rate, f_R, is 0.1

	Values of π_k				
Value of B	0	0.03	0.05	0.07	0.10
$A_k = 0$ (small unit)	0	0.03	0.05	0.07	0.10
$A_k = 1$ (large unit)	1	0.73	0.55	0.37	0.10

This illustrates that for a Pomix sampling design close to Bernoulli (B near 0.10), the inclusion probabilities of large and small units alike lie near the fixed expected sampling rate, 0.10. By contrast, in a Pomix design close to Poisson πps (B near 0), a small unit is practically certain not to be in sample, and a large unit is practically sure to be in sample. The table also illustrates how Pomix sampling with an intermediate value of B will modify the inclusion probabilities: a small unit's chances to appear in the sample are decreased somewhat compared to Bernoulli, and the selection of a large unit becomes less probable than under Poisson πps.

The implications for response burden are: The total response burden on the population rests the same for all values of B; an expected total of $n = n_R + n_C$ units are always asked to report. Compared to Poisson πps ($B = 0$), the fixing of a value B in the interior of the interval $[0,\ f_R]$ will have the effect of shifting some of the response burden from larger onto smaller units; at the same time the precision of the estimates is increased in many cases (see sections 5 and 6).

Finally, we need to mention the second order inclusion probabilities under Pomix sampling, because they are required for the design-based variance calculation. They are simple. If π_{kl} denotes the probability that units k and l are both in the sample, then

$$\pi_{k\ell} = \pi_k \pi_\ell \qquad (2.4)$$

for $k \ne l \in U = U_R \cup U_C$, because the PRN's r_k are independent realizations of the Unif(0, 1) random variable. For $k = l$, we have $\pi_{kl} = \pi_{kk} = \pi_k$. The multiplicative feature (2.4) of the π_{kl} greatly simplifies the design-based variance calculation. We get a simple, single-sum variance estimator, as in (4.2) below.

3. Determining the certainty part of the sample

If the population is highly skewed, a set of units (the certainty part of the sample, s_C) will be sampled with probability one, and Pomix sampling can be used for randomized selection in the remaining part of the population, U_R. Several procedures could be considered for the construction of the certainty set; here we give one that is reasonable (though not necessarily optimal) and used in the

Monte Carlo simulation reported later in section 5. The certainty set is designated with the aid of the known positive size measures x_k through the following procedure in one or more steps. An expected sample size, $n,$ is fixed for the whole sample, $s = s_R \cup s_C.$ In step one, compute the relative size measures $A_{k(1)} = nx_k / \sum_U x_k$ for $k \in U.$ Those units $k,$ if any, for which $A_{k(1)} \geq 1$ are assigned to the certainty part. They form a set denoted $U_{C(1)};$ let its size be $n_{C(1)}.$ The procedure is then repeated to see if additional units should be assigned to the certainty part. In step two, calculate the relative size measures $A_{k(2)} = (n - n_{C(1)})$ $x_k / \sum x_k$ where the summation extends over the set $U - U_{C(1)}.$ If $A_{k(2)} < 1$ for all $k \in U - U_{C(1)},$ the procedure stops, and the final certainty part is $s_C = U_{C(1)}.$ But if $A_{k(2)} \geq 1$ for some units, then these are also assigned to the certainty part, and so on until a step i is reached where all intermediate relative size measures $A_{k(i)}$ are less than unity. The ultimate certainty part s_C will contain, say, n_C units, and we then have $A_k < 1$ for all $k \in U - s_C,$ where $A_k = n_R x_k / \sum x_k$ with $n_R = n - n_C,$ and the sum extends over $U - s_C.$

4. Etimation following Pomix sampling

Although the auxiliary variable serves a useful purpose at the sampling stage, we advocate using it also at the estimation stage. To estimate the population total, $Y = \sum_U y_k,$ consider the generalized regression (GREG) estimator

$$\hat{Y}_{\text{GREG}} = \sum\nolimits_s a_k \, g_k \, y_k \qquad (4.1)$$

where $a_k = 1/\pi_k$ is the sampling weight and the second weight, the g -weight, is given by

$$g_k = 1 + (\boldsymbol{X} - \hat{\boldsymbol{X}})' \, \boldsymbol{T}_s^{-1} \boldsymbol{x}_k / c_k$$

where $\boldsymbol{x_k}$ denotes the auxiliary vector value for unit $k,$ $\hat{\boldsymbol{X}} = \sum_s a_k \boldsymbol{x_k},$ and $\boldsymbol{T_s} = \sum_s a_k \boldsymbol{x_k} \boldsymbol{x'_k} / c_k.$ The auxiliary information requirement is that the vector total $\boldsymbol{X} = \sum_U \boldsymbol{x_k}$ must be known from a reliable source. The unidimensional size variable x_k used for computing the Q_k in the Pomix sampling scheme can be one of the components of $\boldsymbol{x_k}$ or it can a linear combination of the components of $\boldsymbol{x_k}.$ In the empirical study reported in section 5, $\boldsymbol{x_k}$ is unidimensional, and $\boldsymbol{x_k} = x_k.$ The constants $c_k,$ specified by the user, provide a means of weighing the data, in addition to the survey weights $a_k.$ An often used choice is $c_k = 1$ for all $k.$

A commonly used estimator of the variance of the GREG estimator (see Särndal, Swensson and Wretman 1992, Chapter 6) is given as a quadratic form in $g_k e_k,$ where e_k is the regression residual $e_k = y_k - \boldsymbol{x'_k}\hat{\boldsymbol{b}},$ where $\hat{\boldsymbol{b}} = \boldsymbol{T_s^{-1}} \sum_s a_k \boldsymbol{x_k} y_k / c_k.$ One of the advantages of Pomix sampling is that the corresponding variance estimation is simple. This is because the PRN's r_k are independent realizations from the Unif(0, 1) distribution. Hence equation (2.4) applies, and all product terms of the quadratic form are zero. With only the squared terms left, the variance estimator becomes simply $\hat{V} = \sum_s a_k (a_k - 1) g_k^2 e_k^2.$ Finally, because $a_k - 1 = 0$ for all $k \in s_C,$ we get the variance estimator used in the Monte Carlo study reported later in section 5, namely,

$$\hat{V} = \sum\nolimits_{s_R} a_k (a_k - 1) g_k^2 e_k^2. \qquad (4.2)$$

5. A Monte Carlo study of Pomix sampling using finnish data

To illustrate various aspects of Pomix sampling, we conducted a Monte Carlo study involving four different estimators of the population total $Y.$ The experiment involved repeated draws of samples as well as repeated assignments of the set of N PRN's to the population units. Note that since every assignment of the N PRN's to the population units is a random outcome, a proper Monte Carlo study also requires repetitions of the PRN assignments. Therefore, after the first assignment of the N PRN's, we selected 100 Pomix samples, using a fixed value of the Bernoulli width $B.$ (Each sample was realized using a new, randomly selected starting point on the PRN axis). Then a new set of N PRN's were assigned, 100 additional samples were drawn, and so on, until we had reached $100 \times 100 =$ 10,000 PRN/sample pairs, for the given value of $B.$ For each of the 10,000 pairs, we computed the four point estimators, the corresponding four variance estimators, and the corresponding four confidence intervals. With 10,000 repetitions, we expect the Monte Carlo error to be rather small.

The four estimators used in the Monte Carlo study have the following expressions, where $a_k = 1/\pi_k$ is the sampling weight of unit $k,$ and π_k is given by (2.3):

1. The Horvitz-Thompson estimator,

$$\hat{Y}_1 = \sum\nolimits_s a_k y_k = \sum\nolimits_{U_C} y_k + \sum\nolimits_{s_R} a_k y_k.$$

2. The (combined) ratio estimator,

$$\hat{Y}_2 = X \, \hat{b}_2$$

where $X = \sum_U x_k$ and $\hat{b}_2 = \sum_s a_k y_k / \sum_s a_k x_k.$ It is a special case of (4.1) such that $\boldsymbol{x_k} = x_k = c_k.$

3. The GREG estimator

$$\hat{Y}_3 = \sum\nolimits_{U_C} y_k + \sum\nolimits_{s_R} a_k y_k + \hat{b}_3 \left(X_R - \sum\nolimits_{s_R} a_k x_k \right)$$

where $X_R = \sum_{U_R} x_k$ and $\hat{b}_3 = \sum_{s_R} a_k (a_k - 1)$ $y_k \, x_k / \sum_{S_R} a_k \, (a_k - 1) x_k^2.$ It is a special case of (4.1) such that $\boldsymbol{x_k} = x_k$ and $c_k = (a_k - 1)^{-1}.$

4. The (separate) ratio estimator,

$$\hat{Y}_4 = \sum_{U_C} y_k + X_R \hat{b}_4$$

where $X_R = \sum_{U_R} x_k$ and $\hat{b}_4 = \sum_{s_R} a_k y_k / \sum_{s_R} a_k x_k$.

For Poisson πps sampling ($B = 0$), we have $\hat{b}_4 = (\sum_{s_R} y_k / x_k) / n_{s_R}$, where n_{s_R} is the random size of s_R; the corresponding estimator $\hat{Y}_4$ was considered by Brewer *et al.*, (1972). Now $\hat{Y}_2$ and $\hat{Y}_4$ differ in that the regression slope is calculated in $\hat{Y}_2$ on the pooled sample s, but in $\hat{Y}_4$ the slope is calculated separately for the randomization sample s_R. Finally, $\hat{Y}_3$ differs from $\hat{Y}_2$ and $\hat{Y}_4$ in that it uses the weighting $a_k(a_k - 1)$, instead of just a_k. Note that all of $\hat{Y}_2$, and $\hat{Y}_3$ and $\hat{Y}_4$ are members of the GREG family of estimators given by (4.1). By equating $\hat{Y}_2$, $\hat{Y}_3$ and $\hat{Y}_4$ to (4.1), we find the g-weights implied by each of the three estimators. These weights are required for the variance estimation. We can expect the simulation to show that $\hat{Y}_2$, $\hat{Y}_3$ and $\hat{Y}_4$, which use the auxiliary variable both at the design stage and at the estimation stage, will improve on (have smaller variance than) the HT estimator $\hat{Y}_1$, which uses the auxiliary information only at the sampling stage, but the extent of the improvement is unpredictable and interesting to observe.

We used a real data population for the Monte Carlo simulation. This population consists of $N = 1{,}000$ Finnish enterprises. For enterprise k, $k = 1, ..., 1{,}000$, y_k is number of employees (full time equivalents) × 10, and x_k is the wages paid by the enterprise to its employees, in thousands of FIM (Finnish Marks). The auxiliary information (wages paid) comes from the Finnish tax authority's VAT register. The employment variable is the one requiring estimation. The 1,000 units were selected (in an essentially random manner) from an original larger population of Finnish enterprises. Units with a value of zero either on y_k or on x_k were eliminated so that the simulation results would not be disturbed by extraneous factors. Consequently, as for the values y_k and x_k, the population used in the simulation is a natural one, but because of the elimination of units, its features (mean, standard deviation, skewness, *etc.*) differ from those of the original larger population.

The population y-total to be estimated is $Y = \sum_U y_k = 169{,}168$. We fixed the expected sample size for the total sample, $s = s_C \cup s_R$, as $n = 100$. The procedure described in section 3 was used to determine the certainty part s_C of the sample. This resulted in a certainty part s_C consisting of the largest $29 = n_C$ units. The rest of the population, $U_R = U - s_C$ has the following descriptive characteristics: Its size is $N_R = 1{,}000 - 29 = 971$; the total of y is $Y_R = 46{,}138$ (which equals 27% of the entire population total $Y = 169{,}168$); the coefficient of variation (standard deviation divided by mean) is 1.78 for the variable y and 1.94 for the variable x; the coefficient of correlation between x and y is 0.965. The randomization part s_R, of expected size $n_R = 100 - 29 = 71$ units, is realized, in the simulation, by repeated Pomix sample selection from U_R. A plot of (x_k, y_k) for the units k in U_R is shown in the Appendix.

To see the effect of the Bernoulli width, we carried out the simulation for a range of different B-values: $B = 0$, 0.01, ..., 0.07, and, in addition, $B = f_R = n_R / N_R = 71/971 = 0.073$ (which gives Bernoulli). For each value of B, $100 \times 100 = 10{,}000$ PRN/sample pairs were realized, and the results were used to calculate, for each of the four point estimators, five Monte Carlo summary statistics. These are as follows, if $\hat{Y}$ denotes one of the four point estimators, $\hat{V}$ the corresponding variance estimator obtained from (4.2), and $(\hat{Y} - z_{1-\alpha/2}\sqrt{\hat{V}}, \hat{Y} + z_{1-\alpha/2}\sqrt{\hat{V}})$ the corresponding confidence interval for Y at the nominal confidence level $1 - \alpha$, where $z_{1-\alpha/2}$ is the standard normal score, $z_{1-\alpha/2} = 1.960$ for $\alpha = 5\%$, and $z_{1-\alpha/2} = 1.645$ for $\alpha = 10\%$:

(1) MCE $\hat{Y}$ = the Monte Carlo expectation of the point estimator $\hat{Y}$, that is, the arithmetic mean of the 10,000 point estimates;

(2) MCV $\hat{Y}$ = the Monte Carlo variance of the point estimator $\hat{Y}$, that is, the variance of the 10,000 point estimates;

(3) MCE $\hat{V}$ = the Monte Carlo expectation of the variance estimator $\hat{V}$, that is, the arithmetic mean of the 10,000 variance estimates;

(4) MCRTE95 = Monte Carlo coverage rate for nominal 95% confidence intervals, that is, the number of times that the target parameter Y is contained in the confidence interval, divided by 10,000, and expressed in per cent;

(5) MCRTE90 = Monte Carlo coverage rate for nominal 90% confidence intervals; its definition is analogous to that of MCRTE95.

The simulation results are shown in Table 2 (Average sample size, Monte Carlo variance; Monte Carlo expectation of variance estimator) and in Table 3 (Monte Carlo coverage rates). The tables do not show MCE $\hat{Y}$, because in all cases this quantity was very close to the target parameter value $Y = 169{,}168$, confirming that all estimators are essentially unbiased. The deviation of MCE $\hat{Y}$ from Y was in all cases less than 0.14%, in most cases considerably less. The average sample size over the 10,000 repetitions is seen to be very close to $n = 100$, as it should.

Table 2
Results of simulation study for different Bernoulli widths B: Average sample size, MCV$\hat{Y}$ and MCE$\hat{V}$; est. j refers to estimator $\hat{Y}_j$; $j = 1, \ldots, 4$. (Values in the last eight columns to be multiplied by 10^6)

Bernoulli	Average	MCV$\hat{Y}$				MCE$\hat{V}$			
width B	sample size	est. 1	est. 2	est. 3	est. 4	est. 1	est. 2	est. 3	est. 4
0.000	99.95	24.56	3.63	3.43	3.46	24.92	3.67	3.43	3.46
0.010	100.05	22.74	1.96	1.84	1.85	23.53	1.98	1.86	1.87
0.020	100.04	24.75	1.82	1.77	1.78	25.37	1.85	1.77	1.78
0.025	100.09	25.51	1.82	1.78	1.79	26.86	1.87	1.81	1.82
0.030	100.06	28.03	1.83	1.80	1.81	28.58	1.91	1.87	1.88
0.040	99.86	35.17	2.01	2.03	2.06	33.54	2.11	2.11	2.15
0.050	100.02	42.25	2.56	2.64	2.67	41.42	2.48	2.51	2.59
0.060	99.99	56.08	3.42	3.65	3.67	55.70	3.20	3.24	3.44
0.070	100.05	90.73	4.80	5.47	5.59	91.28	4.89	4.72	5.37
0.073	100.02	119.13	6.06	7.09	7.43	116.27	6.00	5.34	6.49

Table 3
Results of simulation study for different Bernoulli widths B: Coverage rates in % of nominal 95% and 90% confidence intervals, MCRTE95 and MCRTE90; est. j refers to estimator $\hat{Y}_j$; $j = 1, \ldots, 4$

Bernoulli	nominal 95% confidence level				nominal 90% confidence level			
width B	est. 1	est. 2	est. 3	est. 4	est. 1	est. 2	est. 3	est. 4
0.000	94.50	92.03	92.75	92.48	89.70	86.36	87.35	86.74
0.010	95.20	93.13	93.47	93.52	90.43	88.06	87.93	88.02
0.020	95.06	94.23	93.88	93.88	90.36	89.36	88.49	88.55
0.025	95.06	93.73	94.56	94.70	90.64	89.15	89.73	89.72
0.030	94.63	94.12	94.09	94.19	89.85	89.06	88.70	88.86
0.040	93.84	94.44	94.47	94.64	88.77	89.60	89.41	89.60
0.050	93.97	94.38	93.76	93.82	88.67	88.67	88.08	88.53
0.060	93.54	93.57	92.12	92.69	89.10	87.57	85.99	87.27
0.070	92.93	94.99	90.67	92.03	88.40	88.62	84.27	86.11
0.073	91.03	95.02	88.03	90.46	86.53	88.62	81.26	83.86

Tables 2 and 3 generate these comments:

1. Let us begin the examination of Table 2 by a comparison of Monte Carlo variances across estimators, for a fixed Bernoulli width B. This shows that, for every value of B, there is little to choose between $\hat{Y}_2$, $\hat{Y}_3$ and $\hat{Y}_4$, in terms of variance. By contrast, the HT estimator $\hat{Y}_1$ has considerably greater variance. To illustrate, the ratio MCV $\hat{Y}_3$/MCV $\hat{Y}_1$ equals 3.43/24.56 = 0.140 for B = 0 (Poisson πps), 1.80/28.03 = 0.064 for $B = 0.03$, and 7.09/119.13 = 0.060 for $B = 0.073$ (Bernoulli). This confirms that the HT estimator is a poor choice compared to an alternative that uses the strongly correlated auxiliary variable. This is true, not surprisingly, for Bernoulli, but also for B-values near the lower end of the $[0, f_R]$ interval, which shows that the sampling design alone does not extract all the power of the auxiliary variable, even though with B near zero, we are close to a strict πps selection (thus supposedly highly efficient). Part of the reason that the HT estimator has a comparatively large variance is that the randomness of the sample size under Pomix sampling penalizes the HT estimator (but not the GREG estimators). Since the HT estimator is inefficient, we do not further discuss it.

2. Examining the small differences between $\hat{Y}_2$, $\hat{Y}_3$ and $\hat{Y}_4$, we note in Table 2: As measured by the Monte Carlo variance, $\hat{Y}_3$ is better than $\hat{Y}_4$ for all Bernoulli widths B, but only marginally so. Also, $\hat{Y}_3$ and $\hat{Y}_4$ are better than $\hat{Y}_2$ at the lower end of the range of B-values, possibly because of the fact that in $\hat{Y}_2$ we allow the certainty part the sample to contribute to the slope estimate, somewhat inappropriately, since there is only an estimation problem for the randomization part. But at the upper end, the relation is reversed and for the upper extreme $B = 0.073$ (Bernoulli), $\hat{Y}_2$ is clearly better than $\hat{Y}_3$. That the differences between $\hat{Y}_2$, $\hat{Y}_3$ and $\hat{Y}_4$ are so small is not surprising, because all are varieties of the GREG estimator (4.1) using essentially the same auxiliary information.

3. Table 2 confirms that the proposed variance estimator $\hat{V}$ works well, as we would expect; $\text{MCE}\hat{V}$ is with few exceptions very close to the target that $\hat{V}$ aims at estimating, that is, the variance of $\hat{Y}$, measured here by $\text{MCV}\hat{Y}$. This holds for all estimators and all values of B, with a few notable exceptions, namely, in the case of $\hat{Y}_3$ and $\hat{Y}_4$ when B is close to the upper extreme (Bernoulli). Then the variance estimator underestimates the variance.

4. The most interesting result in Table 2 we consider to be the fact that the variance of $\hat{Y}_2$ or $\hat{Y}_3$ or $\hat{Y}_4$, when viewed as a function of the Bernoulli width B, does not attain its minimum at $B = 0$ (Poisson πps), as one might have initially guessed, but rather for a value of B somewhere between 0.02 and 0.03. Moreover, the improvement of the case $B = 0.02$ over the case $B = 0$ is substantial for all of $\hat{Y}_2$, $\hat{Y}_3$ and $\hat{Y}_4$. Measuring this improvement by $\text{MCV}(\hat{Y}|B = 0.02)$ divided by $\text{MCV}(\hat{Y}|B = 0)$, we find that this ratio is only about 50% for all of $\hat{Y}_2$, $\hat{Y}_3$ and $\hat{Y}_4$. More precisely, for $\hat{Y}_2$ the ratio is $1.82/3.63 = 0.501$, for $\hat{Y}_3$ it is $1.77/3.43 = 0.516$, and for $\hat{Y}_4$ it is $1.78/3.46 = 0.514$. In view of these results, we added a simulation for $B = 0.025$, a value not examined in the original round of simulations. The results, also displayed in Table 1, confirm that a minimum variance is obtained, for all three estimators, $\hat{Y}_2$, $\hat{Y}_3$ and $\hat{Y}_4$, at a point in the vicinity of $B = 0.025$. One possible explanation of why it is considerably better to take B to be a value distinctly greater than $B = 0$ (which gives Poisson πps) is the following: When B is 0 (or very near 0), the units with the smallest x-values, when selected, will have unduly large weights, which induces high variability. This is avoided by choosing B clearly away from zero.

5. The Monte Carlo results in Table 3 concerning the coverage rates show that the variance estimation and the confidence interval procedure function to satisfaction. As theory leads us to expect, MCRTE95 and MCRTE90 are close, for all four estimators, to their theoretical values, 95% and 90%, respectively. Only for and $\hat{Y}_3$ and $\hat{Y}_4$, when B gets close to the upper extreme (Bernoulli), do we notice any marked tendency for the MCRTE to drop below the nominal value, resulting in part from the underestimation of variance mentioned earlier.

6. Further evidence that Pomix sampling is more efficient than Poisson πps sampling

Initially we had no strong reason to believe that Pomix sampling combined with a GREG estimator would be more efficient for some Bernoulli widths B in the interior of $[0, f_R]$ than for Poisson πps $(B = 0)$. The strong improvement – a variance reduction of around 50% for our particular population – was rather surprising. For other populations, the variance reduction can be more or less than the 50% we found. Because our finding is data dependent, it is desirable to provide some more general evidence in support of the proposition that Pomix sampling with a B-value well into the interior of $[0, f_R]$ is better than Poisson πps sampling $(B = 0)$. We now present some evidence of this kind.

We examined the Taylor variance of $\hat{Y}$ (that is, the variance of the Taylor linearized statistic). It is given by (see Särndal *et al.*, 1992, Chapter 6) $V_{\text{TAY}} = \sum_{U_R}(a_k - 1)E_k^2$, where E_k is the population analogue of the sample based residual e_k used in the variance estimator (4.2). For example, for the estimator $\hat{Y}_3$, the residual in question is $E_k = y_k - b_3 x_k$ with $b_3 = \sum_{U_R}(a_k - 1)\,y_k x_k / \sum_{U_R}(a_k - 1)\,x_k^2$; for $\hat{Y}_4$, $b_4 = \sum_{U_R} y_k / \sum_{U_R} x_k$ replaces b_3.

It is reasonable to model the squared residual as $E_k^2 = \sigma^2 x_k^p(1 + \delta_k)$, where p satisfies $0 \le p \le 2$, and δ_k is near zero. This corresponds to assuming a superpopulation model $y_k = \mathbf{x}_k'\boldsymbol{\beta} + \boldsymbol{\varepsilon}_k$, where the ε_k are independent errors with model expected value zero for every k, and ε_k has model variance $\sigma^2 x_k^p$. Using the approximation $E_k^2 \approx \sigma^2 x_k^p$, and that $a_k = 1/\pi_k$ with π_k given by (2.3), we have

$$V_{\text{TAY}} \approx \sigma^2 \sum_{U_R}(a_k - 1)\,x_k^p = \sigma^2\{H(B,\ p) - T(p)\}$$

where $H(B,\ p) = \bar{x}_{U_R}\sum_{U_R} x_k^p\,[B\bar{x}_{U_R} + (f_R - B)\,x_k]^{-1}$ and $T(p) = \sum_{U_R} x_k^p$. Now consider a fixed value of p such that $0 \le p \le 2$. We want to find out if $H(B,\ p)$ has a smaller value for some B in the interior of the interval $[0,\ f_R]$, compared to its value at $B = 0$, which is $H(0,\ p)$. To this end, let us examine if the derivative $H'(B,\ p) = \partial H(B, p)/\partial B$ is negative at $B = 0$. We find

$$H'(B,\ p) = \bar{x}_{U_R}\sum_{U_R} x_k^p(x_k - \bar{x}_{U_R})[B\bar{x}_{U_R} + (f_R - B)x_k]^{-2}.$$

Its value at $B = 0$ is $H'(0, p) = (\bar{x}_{U_R}/f_R^2)\sum_{U_R} x_k^{p-2}$ $(x_k - \bar{x}_{U_R})$. The sign of $H'(0,\ p)$ is the same as that of $\sum_{U_R} x_k^{p-2}(x_k - \bar{x}_{U_R})$. But this quantity equals, apart from the factor $1/(N_R - 1)$, the covariance in U_R between x_k^{p-2} and $x_k - \bar{x}_{U_R}$ (note that $x_k - \bar{x}_{U_R}$ has zero mean). When p satisfies $0 \le p < 2$, this covariance is negative: when x_k increases, $x_k - \bar{x}_{U_R}$ increases steadily, and x_k^{p-2} decreases steadily (and remains always positive). The sign of $H'(0,\ p)$ is therefore negative; consequently, it is not at $B = 0$ that $H(B,\ p)$ attains its minimum value, but for some B in the interior of $[0,\ f_R]$. Now for $p = 2$, $H'(0,\ p) = 0$, and $H(B,\ p)$ has a minimum at $B = 0$.

These considerations raise the question whether the population used for our simulation in section 5 corresponds to a value of $p \in [0, 2]$, but distinctly less than 2, so that we can expect significant gains from Pomix sampling. To obtain an answer, we estimated p by fitting the logarithmic version of the model $E_k^2 = \sigma^2 x_k^p(1 + \delta_k)$ to the data available for $U_R = U - U_C$. That is, we fitted $w_k = a + pz_k$, where $w_k = \log E_k^2$; $E_k = y_k - b_4 x_k$ with $b_4 = \Sigma_{U_R} y_k / \Sigma_{U_R} x_k$; $z_k = \log x_k$, and a is an intercept term. We obtained the value $p = 1.45$, by treating p as a linear regression slope estimated as $p = \Sigma_{U_R} (w_k - \bar{w}_{U_R})\,(z_k - \bar{z}_{U_R}) / \Sigma_{U_R} (z_k - \bar{z}_{U_R})^2$. Since this p-value is considerably less than 2, our Monte Carlo population is indeed one where one can expect significant gains from the use of Pomix sampling with a value of B in the interior of $[0,\ f_R]$.

7. Concluding discussion and tentative recommendations

The survey sampler will ask: If I consider using Pomix sampling for my survey, combined with a GREG estimator, what is an appropriate choice for B? Recall that in this paper we found, for one particular population, that a large efficiency gain (roughly 50% variance reduction compared to Poisson πps) is realized by fixing the Pomix parameter B at around 30% of f_R. We were led to suspect that the variance gain is related to residual error characteristics, and this was confirmed in Section 6 which presented evidence that when the squared residual pattern conforms to $E_k^2 = \sigma^2 x_k^p$, where p satisfies $0 \le p < 2$, as is the case in many business survey populations, then Pomix sampling with B in the interior of the interval $[0,\ f_R]$ may well be advantageous. However, the present paper does not address the question of the optimal choice of B. A difficulty is that in practice the value of B must be fixed at the design stage, and that the optimal B depends on unknown population characteristics. Prior knowledge of the population, notably about its residual variance structure, can guide the choice of B.

Our tentative recommendations based on this paper are: If prior information suggests a squared residual pattern conforming to $\sigma^2 x_k^p$ with $p < 2$, then use Pomix sampling with $B = 0.3\ f_R$. On the other hand, if in reality the unknown p is such that $p > 2$, then, although the best choice in this case might be $B = 0$ (Poisson πps), little harm would probably be done to use $B = 0.3\ f_R$, because the variance viewed as a function of B is likely to increase at a gentle rate. Therefore, $B = 0.3\ f_R$, seems a reasonable all-purpose suggestion. These recommendations are tentative; the question merits a further study that lies beyond the scope of this paper.

Acknowledgements

The authors gratefully acknowledge the work of a referee, whose suggestions led to valuable improvements in the original manuscript.

Appendix

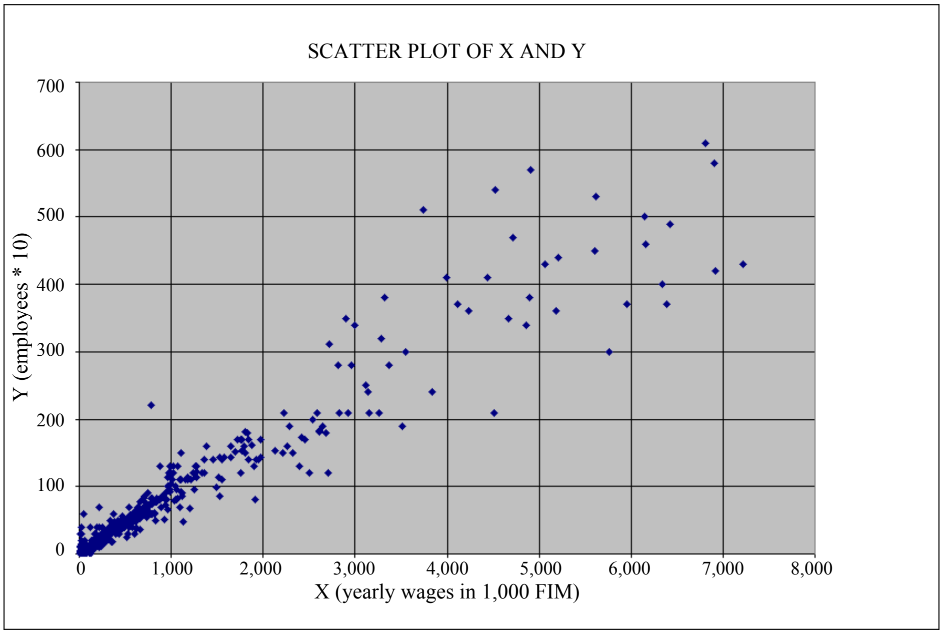

Figure 3 Scatter plot of x (yearly wages) against y (employment) for the portion U_R of the Monte Carlo population of 1,000 Finnish enterprises

References

Atmer, J., Thulin, G. and Bäcklund, S. (1975). Coordination of samples with the JALES technique. *Statistisk Tidskrift*, 13, 443-450.

Brewer, K.R.W., Early, L.J. and Joyce, S.F. (1972). Selecting several samples from a single population. *Australian Journal of Statistics*, 14, 231-239.

Fan, C.T., Muller, M.E. and Rezucha, I. (1962). Development of sampling plans by using sequential (item by item) techniques and digital computers. *Journal of the American Statistical Association*, 57, 387-402.

Ohlsson, E. (1995). Coordination of samples using permanent random numbers. In *Business Survey Methods* (Eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott) New York: John Wiley & Sons, Inc., 153-169.

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

Sunter, A.B. (1977). Response burden, sample rotation, and classification renewal in economic surveys. *International Statistical Review*, 45, 209-222.