Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

On the efficiency of randomized probability proportional to size sampling

by Paul Knottnerus

June 2011

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 2011
Vol. 37, No. 1, pp. 95-102
Statistics Canada, Catalogue No. 12-001-X

On the efficiency of randomized probability proportional to size sampling

Paul Knottnerus [1]

Abstract

This paper examines the efficiency of the Horvitz-Thompson estimator from a systematic probability proportional to size (PPS) sample drawn from a randomly ordered list. In particular, the efficiency is compared with that of an ordinary ratio estimator. The theoretical results are confirmed empirically with of a simulation study using Dutch data from the Producer Price Index.

Key Words: Horvitz-Thompson estimator; Producer Price Index; Ratio estimator; Sampling autocorrelation coefficient.

1. Introduction

When the study variable y in a population of N units is more or less proportional to a size variable x, one may use the ratio estimator from a simple random sample of size n without replacement (SRS). An alternative estimator in such a situation is the Horvitz-Thompson (HT) estimator in combination with a systematic probability proportional to size sample from a randomly ordered list, henceforth called a randomized PPS sample.

In recent years several authors investigated variance estimation procedures for the HT estimator from a randomized PPS sample. See, among others, Brewer and Donadio (2003), Cumberland and Royall (1981), Deville (1999), Knottnerus (2003), Kott (1988 and 2005), Rosén (1997) and Stehman and Overton (1994). For a comparison between the efficiencies of the ratio estimator and the randomized PPS estimator, the reader is referred to Foreman and Brewer (1971), Cochran (1977) and the references given therein. A drawback of these comparisons is that finite populations corrections are ignored. Hartley and Rao (1962) take the finite population correction into account but without an explicit formula for the efficiency. Elaborating on the results of Gabler (1984), Qualité (2008) shows that the related HT estimator from a rejective Poisson sample of size n is more efficient than the Hansen-Hurwitz estimator for a sampling scheme with replacement. No formula for the increased efficiency is given, however.

The main aim of this paper is to derive formulas for the efficiency of the randomized PPS estimator relative to the ratio estimator. To this end, we present a simple formula for the change in the sample size required to maintain the same variance when a randomized PPS estimator is replaced by a ratio estimator. From the design based point of view these formulas are valid when $n = o(N)$ as $N \to \infty$. This condition suggests that the finite population correction can be neglected for this kind of sampling design. Surprisingly, as we will see in an example in section 4, the randomized PPS sampling can reduce variance by more than 30% compared to PPS sampling *with* replacement even when the sampling fraction n/N is much smaller than 30%; see also Kott (2005, page 436). Furthermore, the formulas remain appropriate from a model assisted point of view when n and N are of the same order, provided that N is large and that the hypothetical model for the observations Y_i ($i = 1, ..., N$) satisfies mild conditions.

The outline of the paper is as follows. Section 2 describes an alternative expression for the variance of the HT estimator based on the sampling autocorrelation coefficient. The corresponding variance estimator for randomized PPS sampling is shown to be nonnegative with probability 1. Section 3 presents the formulas for the efficiency of the randomized PPS estimator relative to the ratio estimator for various data patterns often met in practice. Section 4 features an example with data on the Producer Price Index in The Netherlands illustrating the substantial efficiency gains obtainable in practice. A counterexample shows that randomized PPS sampling is not *always* advantageous. The paper concludes with a summary.

2. An alternative variance expression for randomized PPS sampling

Consider a population $U = \{1, ..., N\}$, and let s be a sample of fixed size n drawn from U without replacement according to a given sampling design with first order inclusion probabilities π_i and second order inclusion probabilities π_{ij} ($i, j = 1, ..., N$). The HT estimator of the population total, $Y = \Sigma_{i\in U} Y_i$, is defined by $\hat{Y}_{\mathrm{HT}} = \Sigma_{i\in s} Y_i/\pi_i$. Suppose there is a measure of relative size X_i (*i.e.*, $X = \Sigma_{i\in U} X_i = 1$) such that all $X_i \le 1/n$. In fact, it is assumed here that units with $X_i > 1/n$ are put together in a separate certainty-stratum. When the π_i are proportional to these size measures, $\pi_i = nX_i$. Defining $Z_i = Y_i/X_i$, we can write Y as a weighted mean of the Z_i, that is, $Y = \mu_z = \Sigma_{i\in U} X_i Z_i$. Likewise, we can write the HT estimator of Y in randomized PPS sampling as $\hat{Y}_{\mathrm{HT}} = \hat{Y}_{\mathrm{PPS}} = \bar{z}_s$, where $\bar{z}_s$ is sample mean of the Z_i.

1. Paul Knottnerus, Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands. E-mail: pkts@cbs.nl.

The variance of the randomized PPS estimator $\hat{Y}_{\text{PPS}}$ is

$$\text{var}(\hat{Y}_{\text{PPS}}) = \frac{1}{n^2}\sum_{i\in U}\sum_{j\in U}(\pi_{ij} - \pi_i\pi_j)Z_iZ_j \quad (1)$$

$$= -\frac{1}{2n^2}\sum_{i\in U}\sum_{j\in U}(\pi_{ij} - \pi_i\pi_j)(Z_i - Z_j)^2 \quad (2)$$

with $\pi_{ii} = \pi_i$. The former is attributed to Horvitz and Thompson (1952) and the latter is due to Sen (1953) and Yates and Grundy (1953). The following alternative expression for the variance is more convenient for our purposes:

$$\text{var}(\hat{Y}_{\text{PPS}}) = \text{var}(\bar{z}_s) = \{1 + (n-1)\rho_z\}\frac{\sigma_z^2}{n}, \quad (3)$$

where $\sigma_z^2 = \sum_{i\in U} X_i(Z_i - \mu_z)^2$, and

$$\rho_z = \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}\frac{\pi_{ij}}{n(n-1)}\left(\frac{Z_i - \mu_z}{\sigma_z}\right)\left(\frac{Z_j - \mu_z}{\sigma_z}\right). \quad (4)$$

For a proof of (3), see Knottnerus (2003, page 103). Note that σ_z^2/n would have been the variance if the sample had been drawn with replacement with drawing probabilities X_i.

The sampling autocorrelation coefficient ρ_z in (4) is a generalization of the more familiar intraclass correlation coefficient ρ in systematic sampling with equal probabilities; see, for instance, Cochran (1977, pages 209 and 240) and Särndal, Swensson and Wretman (1992, page 79). Note that ρ_z is a fixed population parameter. The phrase *sampling autocorrelation* is used because ρ_z refers to the autocorrelation between two randomly chosen observations, say z_{s1} and z_{s2}, from s. Consequently, the value of ρ_z depends on the sampling design. In particular, when sampling with replacement, $\rho_z = 0$, while under SRS sampling, $\rho_z = -1/(N-1)$.

Although exact expressions for the π_{ij} under randomized PPS sampling are available, they can be cumbersome when N is large. For an exact expression, see Connor (1966) and for a modification Hidiroglou and Gray (1980). Here we use an approximation proposed by Knottnerus (2003, page 197):

$$\pi_{ijK} = n(n-1)\frac{X_iX_j(1 - X_i - X_j)}{\gamma(1-2X_i)(1-2X_j)} \quad (5)$$

$$\gamma = \frac{1}{2} + \frac{1}{2}\sum_{i\in U}\frac{X_i}{1-2X_i}.$$

These π_{ijK} have been shown to satisfy the second-order restrictions for the π_{ij}:

$$\sum_{i,\, j\in U(j\neq i)}\pi_{ij} = n(n-1),$$

and

$$\sum_{j\in U(j\neq i)}\pi_{ij} = (n-1)\pi_i.$$

Furthermore, (5) is correct for SRS sampling for any $n \leq N$, while π_{ijK} coincide with the π_{ijBD} from the special designs proposed by Brewer (1963a) and Durbin (1967) for PPS samples with $n = 2$. Moreover, the π_{ijK} in (5) can be written in factorized form as proposed by Brewer and Donadio (2003). That is,

$$\pi_{ijK} = \pi_i\pi_j(c_i + c_j)/2, \quad (6)$$

and

$$c_i = (n-1)/n\gamma(1 - 2X_i).$$

An implication of approximation (5) is that $\pi_{ijK}/n(n-1)$ does not depend on n. Hence, the corresponding approximation of ρ_z does not depend on n (recall we have assumed that every $X_i < 1/n$).

This nondependence on n would also result had we used the approximation proposed by Hartley and Rao (1962) for randomized PPS sampling:

$$\pi_{ijHR} = n(n-1)X_iX_j$$
$$\{1 + X_i + X_j - \mu_x + 2(X_i^2 + X_j^2 + X_iX_j)$$
$$- 3\mu_x(X_i + X_j - \mu_x - 2\Sigma_{i\in U}X_i^3)\}, \quad (7)$$

where $\mu_x = \Sigma_{i\in U}X_i^2$ (recall $\mu_z = \Sigma_{i\in U}X_iZ_i$). Obviously, $\pi_{ijHR}/n(n-1)$ does not depend on n. At the time Hartley and Rao assumed that $n = O(1)$ as $N \to \infty$. In addition, referring to a private conversation with J.N.K. Rao, Thompson and Wu (2008) state that approximation (7) is valid when $n = o(N)$ as $N \to \infty$. For an example that (5) and (7) can not be used for *any* n and N, see Appendix A.

Since both (5) and (7) lead to approximations for ρ_z in randomized PPS sampling that are $\rho_z\{1+o(1)\}$ as $N \to \infty$ with $n = o(N)$, (5) can be used for calculating ρ_z in practice when $n << N$ and N is large. For ease of the exposition, it is assumed here that there is a positive constant c such that $\rho_z < -c/N$. See also Kott (2005, page 436) who discusses estimating the variance under PPS sampling when $n = O(N^{2/3})$.

Suppose $\gamma = 1 + \mu_x + O(1/N^2)$ and $\mu_x = O(1/N)$ (which follow from the conditions of Theorem 1 below). It is not hard to see that, after dropping $O(1/nN)$ terms, c_i in (6) is identical with $c_{iHR} = (n-1)/\{n(1 + \mu_x - 2X_i)\}$. The latter expression is equation (11) of Brewer and Donadio, which is based on π_{ijHR} in (7).

The approach proposed here is somewhat different from Knottnerus (2003). First, rewrite (5) as

$$\pi_{ijK} = n(n-1)\frac{X_iX_j}{\gamma}\left(\frac{1/2}{1-2X_i} + \frac{1/2}{1-2X_j}\right). \quad (8)$$

Substituting (8) into (4), we obtain a new, simple approximation for ρ_z:

$$\rho_z = \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}} \frac{X_i X_j}{\gamma}\left(\frac{1/2}{1-2X_i}+\frac{1/2}{1-2X_j}\right)\left(\frac{Z_i-Y}{\sigma_z}\right)\left(\frac{Z_j-Y}{\sigma_z}\right)$$

$$= \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}} \frac{X_i X_j}{\gamma}\left(\frac{1}{1-2X_i}\right)\left(\frac{Z_i-Y}{\sigma_z}\right)\left(\frac{Z_j-Y}{\sigma_z}\right)$$

$$= 0 - \sum_{i\in U}\frac{X_i^2}{\gamma(1-2X_i)}\left(\frac{Z_i-Y}{\sigma_z}\right)^2. \qquad (9)$$

In the second line, we used the equality $\sum_{i,j} m_{ij} v_i = \sum_{i,j} m_{ij} v_j$ when $m_{ij} = m_{ji}$. In the last line, we used $\sum_{j\in U} X_j (Z_j - Y) = 0$.

Next, let $\bar{X}$ denote the population mean of $X_1, ..., X_N$ and define σ_x^2 and V_x^2 by

$$\sigma_x^2 = \sum_{i\in U} X_i (X_i - \mu_x)^2,$$

and

$$V_x^2 = \sum_{i\in U} (X_i - \bar{X})^2 / N,$$

respectively. In the following theorem (9) is further simplified.

Theorem 1. Suppose that $(Z_i - Y)/\sigma_z = O(1)$ as $N \to \infty$ and that there are positive constants c and C such that $V_x/\bar{X} < c$, $\sigma_x/\mu_x < c$ and $0 < X_i < C < 1/2$. Then, for large N and $n << N$,

$$\rho_z = -\frac{\sum_{i\in U} X_i^2 (Z_i - Y)^2}{\sum_{i\in U} X_i (Z_i - Y)^2}\left\{1 + O\left(\frac{1}{N}\right)\right\} + O\left(\frac{1}{N^2}\right). \qquad (10)$$

Proof. Because $\bar{X} = 1/N$, it follows from the above assumptions that the weighted mean μ_x $[= \Sigma X_i^2 = N(V_x^2 + \bar{X}^2)]$ is of order $1/N$ and hence, $\sigma_x = O(1/N)$. Because $(1-2X_i)^{-1} = 1 + 2X_i + O(X_i^2)$ for $0 < X_i < C < 1/2$, ρ_z from (9) can be written for $N \to \infty$ as

$$\rho_z = -\sum_{i\in U}\frac{X_i^2}{\gamma}\left(\frac{Z_i - Y}{\sigma_z}\right)^2 + \frac{1}{\gamma} O\left(\sum_{i\in U} X_i^3\right),$$

where $\sum_{i\in U} X_i^3 = \sigma_x^2 + \mu_x^2 = O(N^{-2})$, and

$$\gamma = \frac{1}{2} + \frac{1}{2}\sum_{i\in U} X_i\{1 + 2X_i + O(X_i^2)\}$$

$$= 1 + \mu_x + O\left(\frac{1}{N^2}\right) = 1 + O\left(\frac{1}{N}\right),$$

from which (10) follows. This concludes the proof.

Substituting (10) into (3), we get

$$\text{var}(\hat{Y}_{\text{PPS}}) = \frac{\sigma_z^2}{n} - \frac{n-1}{n}\sum_{i\in U} X_i^2 (Z_i - Y)^2$$

$$= \frac{1}{n}\sum_{i\in U} X_i\{1 - (n-1)X_i\}(Z_i - Y)^2, \qquad (11)$$

which is also given by Hartley and Rao (1962). It is noteworthy that approximation (10) also follows directly from substituting the simple approximation $\pi_{ijAP} = n(n-1) X_i X_j$ into (4). Likewise, use of π_{ijHR} leads to an expression almost similar to (9) and hence to (10). In addition, direct use of π_{ijAP} in (1) or (2) for the SRS case with $X_i = X_j = 1/N$ may lead to errors of more than 100% for populations with $\bar{Y} = V_y^2$; see Knottnerus (2003, pages 274-6). Hence, (1) and (2) are more sensitive to small errors in the π_{ij} than (3) and (4). Furthermore, note that when n is so small that $|n\rho_z| << 1$, we may set $\rho_z = 0$ yielding the with-replacement variance formula of Hansen and Hurwitz (1943).

In order to estimate (3) using ρ_z, denote, as before, a randomly chosen observation from s by z_{s1}. Then we have

$$\sigma_z^2 = \text{var}(z_{s1}) = \text{var}\{E(z_{s1}|s)\} + E\{\text{var}(z_{s1}|s)\}$$

$$= \text{var}(\bar{z}_s) + E\left(\frac{n-1}{n} s_z^2\right),$$

where

$$s_z^2 = \frac{1}{n-1}\sum_{i\in s} (Z_i - \bar{z}_s)^2.$$

Now from (3), it is seen that $s_z^2/(1-\rho_z)$ is an unbiased estimator for σ_z^2. When ρ_z is very small, the term $(1-\rho_z)$ can be neglected. When n is sufficiently large, the ratio ρ_z from (9) can be estimated by

$$\hat{\rho}_{z9} = -\frac{\sum_{i\in s} X_i (Z_i - \bar{z}_s)^2 / \hat{\gamma}(1 - 2X_i)}{\sum_{i\in s} (Z_i - \bar{z}_s)^2},$$

where

$$\hat{\gamma} = \frac{1}{2} + \frac{1}{2n}\sum_{i\in s}\frac{1}{1-2X_i}.$$

Because $\hat{\gamma} \geq 1$ and $X_i \leq 1/n$, we have $\hat{\rho}_{z9} \geq -1/(n-2)$. For the bias of an estimated ratio when n is small, see Cochran (1977, page 160).

In a similar manner ρ_z from (10) can be estimated by

$$\hat{\rho}_{z10} = -\frac{\sum_{i\in s} X_i (Z_i - \bar{z}_s)^2}{\sum_{i\in s} (Z_i - \bar{z}_s)^2} \geq \frac{-1}{n} > \frac{-1}{n-1}.$$

Hence, replacing σ_z^2 and ρ_z in (3) by $s_z^2/(1-\hat{\rho}_{z10})$ and $\hat{\rho}_{z10}$, respectively, leads to a nonnegative variance estimator

with probability 1. This also holds for $\hat{\rho}_{z9}$ when all $X_i \leq 1/(n+1)$. The estimator for $\text{var}(\hat{Y}_{\text{PPS}})$ thus obtained becomes

$$\hat{\text{var}}_{\rho}(\hat{Y}_{\text{PPS}}) = \frac{\{1 + (n-1)\hat{\rho}_{z9}\}s_z^2}{n(1-\hat{\rho}_{z9})}.$$

Moreover, for moderate values of N, estimator $\hat{\rho}_{z9}$ has probably better properties than $\hat{\rho}_{z10}$ because the π_{ijK} underlying (9) satisfy exactly the second-order restrictions irrespective of the values of n and N.

3. Efficiency of $\hat{Y}_{\text{PPS}}$ for large n and N

3.1 Efficiency formulas

Because $X = 1$, the ratio estimator for Y becomes

$$\hat{Y}_R = \frac{\bar{y}_s}{\bar{x}_s} = \frac{\sum_{i\in s} X_i Z_i}{\sum_{i\in s} X_i}.$$

For sufficiently large n the commonly used approximation for its variance is

$$\text{var}(\hat{Y}_R) = \frac{N(N-n)}{n(N-1)} \sum_{i\in U} X_i^2 (Z_i - Y)^2. \qquad (12)$$

From (3) and (12) it can be seen that the efficiency of $\hat{Y}_{\text{PPS}}$ relative to $\hat{Y}_R$ can be written as

$$Eff_{P/R} = \frac{\text{var}(\hat{Y}_R)}{\text{var}(\hat{Y}_{\text{PPS}})} = \frac{(N-n)\sum_{i\in U} X_i^2 (Z_i - Y)^2}{\{1 + (n-1)\rho_z\}\sigma_z^2}, \qquad (13)$$

assuming $N/(N-1) \approx 1$. Combining (10) and (13) gives

$$Eff_{P/R} = \frac{-(N-n)\rho_z}{1 + (n-1)\rho_z}. \qquad (14)$$

Now suppose that the observations Y_i satisfy the model:

$$Y_i = \mu X_i + \varepsilon_i, \qquad (15)$$

with $E(\varepsilon_i) = 0$, $E(\varepsilon_i^2) = \sigma^2 X_i^{\delta}$, and $E(\varepsilon_i \varepsilon_j) = 0$ $(i \neq j)$. Consequently, for the Z_i we have $Z_i = \mu + u_i$ with $E(u_i) = 0$, $E(u_i^2) = \sigma^2 X_i^{\delta-2}$, and $E(u_i u_j) = 0$ $(i \neq j)$. According to Kott (1988), δ often lies between 1 and 2. See also Brewer (1963b). Brewer and Donadio (2003) showed that by assuming a model like (15), (7) and hence (10) and (14) hold when n and N are of the same order as $N \to \infty$. Furthermore, for sufficiently large N we can replace Y as well as the numerator and denominator in (10) by their model expectations. This yields

$$\rho_z = -\frac{\Sigma_{i\in U} X_i^{\delta}}{\Sigma_{i\in U} X_i^{\delta-1}}. \qquad (16)$$

In the next subsections we look more closely at the relationship between δ and the efficiency of $\hat{Y}_{\text{PPS}}$.

3.2 Efficiency of $\hat{Y}_{\text{PPS}}$ when $\delta = 2$

For $\delta = 2$, (16) gives $\rho_z = -\Sigma_{i\in U} X_i^2 = -\mu_x$, which can also be written as

$$\rho_z = -\frac{1}{N}(1 + CV_x^2), \qquad (17)$$

because

$$\frac{1}{N}\sum_{i\in U} X_i^2 = V_x^2 + \bar{X}^2 = \bar{X}^2(1 + CV_x^2),$$

where $\bar{X} = 1/N$ and $CV_x = V_x/\bar{X}$ is the coefficient of variation of the X_i. Substituting (17) into (14) gives

$$Eff_{P/R} = \frac{(N-n)(1+CV_x^2)}{N - (n-1)(1+CV_x^2)}.$$

Hence, for $\delta = 2$, the efficiency of the randomized PPS sample is high when the variability among the X_i is high. When $CV_x = 0$, randomized PPS sampling amounts to SRS sampling and obviously, $Eff_{P/R} = 1$ assuming $(N - n + 1) \approx (N - n)$; note that this assumption holds when N is sufficiently large and $n/N < f_0 < 1$.

Observe that substituting $n = n_{\text{PPS}}(1 + CV_x^2)$ into (12) leads to about the same outcome as (3) and (10) with n_{PPS} instead of n. Hence, when $CV_x = 1.5$, randomized PPS sampling with sample size $n_{\text{PPS}} = 100$ is as efficient as the ratio estimator from an SRS sample of size $n_{\text{SRS}} = 325$. More generally, assuming that $(n-1)/n \approx 1$, it is seen from (3), (10), and (12) that a ratio estimator from an SRS sample of size n_{SRS} is as efficient as a PPS sample of size n_{PPS} when

$$n_{\text{SRS}} = -n_{\text{PPS}}\rho_z N. \qquad (18)$$

3.3 Efficiency of $\hat{Y}_{\text{PPS}}$ for $\delta < 1$ vs $\delta \geq 1$

Another special case is $\delta = 1$. From (16), $\rho_z = -1/N$ when $\delta = 1$. Subsequently, it follows from (14) that under model (15) $Eff_{P/R} = 1 + O(N^{-1})$, provided that $n/N < f_0 < 1$ as $N \to \infty$ irrespective of the value of CV_x. Furthermore, it can be shown that $Eff_{P/R}$ is an increasing function of δ. This is proven below in Lemma 1. Hence, for $\delta < 1$ the randomized PPS estimator is less efficient than the ratio estimator, while for $\delta > 1$ the randomized PPS estimator is more efficient than the ratio estimator.

Lemma 1. Let $Eff_{P/R}$ and ρ_z be defined by (14) and (16), respectively. If $V_x^2 > 0$, then $Eff_{P/R}$ is a monotonically increasing function of δ.

Proof. Write ρ_z from (16) as a weighted mean of the (negative) X_i

$$\rho_z = -u(\delta) = -\sum_{i \in U} w_i X_i,$$

where

$$w_i = \frac{X_i^{\delta-1}}{\sum_{i \in U} X_i^{\delta-1}} \quad \text{[Note that } \mu_x = u(2)\text{]}.$$

Let $X_i > X_j$ $(i \neq j)$, and define $h(\delta)$ as $w_i/w_j = (X_i/X_j)^{\delta-1}$. Since $h(\delta)$ is increasing in δ, the weight of the larger X_i is increasing compared to that of X_j when δ is increasing. Hence, $u(\delta)$ is increasing and ρ_z is decreasing in δ. It suffices therefore to show that $Eff_{P/R}$ is decreasing in ρ_z. Writing (14) as

$$Eff_{P/R} = \frac{-(N-n)}{\rho_z^{-1} + (n-1)},$$

it is seen that $Eff_{P/R}$ is decreasing in ρ_z indeed. This concludes the proof.

3.4 An alternative structure among the disturbances

Finally, suppose the variance of the disturbances in (15) is of the form:

$$\text{var}(\varepsilon_i) = c_1 X_i + c_2 X_i^2 \qquad (0 < c_1, c_2 \leq 1).$$

See Kott (1988). For this case we obtain in analogy with (16)

$$\rho_z = -\sum_{i \in U} \omega_i X_i,$$

where

$$\omega_i = \frac{1 + \phi X_i}{\sum_{i \in U}(1 + \phi X_i)}, \text{ and } \phi = c_2/c_1$$

when $\varphi = 0$, $\rho_z = -1/N$. Hence, when $c_2 = 0$, PPS sampling is only as efficient as the ordinary ratio estimator from SRS sampling. Along the same lines as the proof of Lemma 1, it can be shown that ρ_z is decreasing in φ while $Eff_{P/R}$ is increasing in φ Hence, for this case the randomized PPS estimator is always more efficient than the ratio estimator when c_2 is positive.

4. An application to the Producer Price Index

The Producer Price Index (PPI) in The Netherlands is based on about 2,500 commodity price indexes organized by type of product. The price index for a specific commodity can be written as

$$Y = \sum_{i \in U} X_i Z_i,$$

where Z_i is the price change for that commodity of establishment i relative to the basic period while X_i is the relative sales of that commodity by establishment i in the basic period (recall $\Sigma X_i = 1$).

In the example given here, we examine the price changes of 70 establishments for the commodity *Basic Metal* in December of 2005 relative to December of 2004; see Table 1. We compare the variance of the ratio estimator from an SRS sample with the variance of the HT estimator from a randomized PPS sample when $n = 9$. Applying (12) to these data gives $\text{var}(\hat{Y}_R) = 101$. If the sample had been drawn with replacement the variance would have been 116. Applying (3) and (9) for a randomized PPS sample gives $\text{var}(\hat{Y}_{\text{PPS},\gamma}) = 29.9$. This outcome takes γ into account and lies close to the result $V_{\text{PPS}}^{(sim)} = 29.2$ from a simulation experiment consisting of 80,000 randomized PPS samples of size $n = 9$ from the set of 70 establishments. Hence, $Eff_{P/R} = 3.5$. Because formula (12) for $\text{var}(\hat{Y}_R)$ is only asymptotically unbiased, we also carried out simulations evaluating the mean square error (MSE) and the bias of $\hat{Y}_R$ resulting in $\text{MSE}_R^{(sim)} = 108$ and a relatively small bias of 0.7. This confirms the conjecture that (12) gives an underestimation of the true variance; see Cochran (1977). Hence, for moderate samples the true value of $Eff_{P/R}$ might be somewhat higher than (14) suggests.

Furthermore, it is noteworthy that the simpler formula (10) for ρ_z in combination with (3) gives almost the same result $\text{var}(\hat{Y}_{\text{PPS}}) = 30.7$ even though $N = 70$ is not very large. The with replacement PPS variance would have been 43.8. Hence, the variance reduction for randomized PPS sampling is more than 30% even though the sampling fraction n/N is much smaller. According to (18), formula (12) with $n_{\text{SRS}} = 26$ gives about the same outcome as (3) with $n_{\text{PPS}} = 9$; note: $\rho_z = -0.042$. Hence, the sample sizes differ by a factor 2.9, which is more or less in line with the factor $(1 + CV_x^2) = 3.1$ from subsection 3.2. This should not be surprising because the price changes and their variability hardly depend on the sizes of the company. Fitting a double log regression

$$\ln(Z_i - Y)^2 = \alpha + \beta \ln X_i + v_i \qquad (19)$$

results in the estimate $\hat{\beta} = 0.07$ for the data in Table 1; units with $Z_i = Y$ should be omitted in the regression. The estimate $\hat{\beta} = 0.07$ corresponds with $\hat{\delta} = 2.07$ for the disturbances in (15) which explains the superiority of randomized PPS sampling for this type of data. Also for other commodities $\hat{\delta}$ often was about 2; see Enthoven (2007).

Table 1
Price changes (Z_i) and sizes (X_i) of 70 establishments

i	price change	size	i	price change	size
1	-18.4%	0.0608	36	34.8%	0.0427
2	-16.0%	0.0784	37	13.1%	0.0121
3	3.3%	0.0762	38	31.7%	0.0351
4	12.5%	0.0100	39	-24.8%	0.0074
5	0.0%	0.0029	40	55.3%	0.0009
6	8.3%	0.0006	41	40.5%	0.0066
7	-39.0%	0.0182	42	34.6%	0.0022
8	-25.1%	0.0020	43	1.7%	0.0001
9	1.1%	0.0040	44	0.0%	0.0039
10	4.4%	0.0066	45	3.9%	0.0304
11	-4.9%	0.0039	46	25.4%	0.0209
12	-8.9%	0.0070	47	25.6%	0.0062
13	-7.0%	0.0148	48	0.0%	0.0033
14	-15.0%	0.0108	49	-0.3%	0.0019
15	-10.7%	0.0087	50	66.6%	0.0346
16	-9.0%	0.1079	51	0.0%	0.0039
17	-11.3%	0.0247	52	-2.9%	0.0007
18	10.6%	0.0024	53	15.8%	0.0011
19	-23.2%	0.0001	54	0.0%	0.0026
20	-25.4%	0.0001	55	0.0%	0.0018
21	-80.7%	0.0002	56	11.6%	0.0057
22	13.4%	0.0005	57	0.0%	0.0042
23	-42.5%	0.0010	58	0.0%	0.0236
24	-34.8%	0.0014	59	-1.5%	0.0015
25	-30.0%	0.0126	60	0.0%	0.0003
26	8.0%	0.0530	61	11.7%	0.0067
27	0.0%	0.0208	62	0.0%	0.0012
28	2.1%	0.0119	63	0.8%	0.0040
29	11.3%	0.0208	64	2.0%	0.0009
30	0.7%	0.0322	65	2.3%	0.0018
31	9.5%	0.0447	66	4.7%	0.0026
32	11.5%	0.0018	67	0.9%	0.0064
33	5.8%	0.0174	68	-1.0%	0.0309
34	-6.9%	0.0197	69	-0.5%	0.0005
35	0.0%	0.0124	70	0.0%	0.0006

We conclude this section with a small example showing that randomized PPS is not *always* better than the ratio estimator. Although the data in Table 2 for a population of five units are artificial, a data pattern like this may occur in financial branches where very small financial companies may grow very fast with respect to certain financial variables. This high variability among growth rates of small companies results in a low value for δ. For an SRS sample with $n = 2$ from the five units in Table 2 the variance of the ratio estimator is 211 according to (12); simulations give $\mathrm{MSE}_R^{(sim)} = 323$. This is much less than the variance of 557 found in a simulation consisting of 80,000 randomized PPS samples of size $n = 2$. Formula (3) in combination with (9) gives the same outcome: 557. This would also be the correct variance had sample been drawn according to Brewer (1963a) or Durbin (1967). Formula (11), based on (10), gives a slightly different value, 556.

Regression (19) with the data from Table 2 yields $\hat{\beta} = -3.0$, and hence $\hat{\delta} = -1.0$. In line with the findings of subsection 3.3 this low value $\hat{\delta} = -1.0$ explains why $\hat{Y}_{\mathrm{PPS}}$ is less efficient than $\hat{Y}_R$ in this example. Moreover, the ordinary direct estimator $N\bar{y}_s$ from an SRS sample has a variance of 356, which is even smaller here than the variance in randomized PPS sampling; $\bar{y}_s$ being the sample mean of the Y_i. Hence, for this type of data, the ratio estimator is the best option. Recall that the ratio estimator has a smaller variance than $N\bar{y}_s$ when $b > Y/2X$ where b is the slope of a regression from Y_i on X_i and a constant $(i = 1, ..., N)$; see Knottnerus (2003, page 117). So the data Y_i $(= X_i Z_i)$ in Table 2 certainly do not exhibit a flat trend.

Table 2
Growth rates of assets (Z_i) and sizes (X_i) of 5 establishments

i	growth rate	size
1	200%	0.0455
2	33%	0.1364
3	75%	0.1818
4	33%	0.2727
5	62%	0.3636

5. Summary

This paper compares the variance of the HT estimator $\hat{Y}_{\text{PPS}}$ from a randomized PPS sample with the variance of the classical ratio estimator $\hat{Y}_R$ from an SRS sample of the same size. In this comparison the sampling autocorrelation coefficient ρ_z plays an important role.

When the data pattern of the variables x and z $(= y/x)$ is such that $\rho_z < -1/(N-1)$, it can be shown under mild conditions that $\hat{Y}_{\text{PPS}}$ is more efficient than $\hat{Y}_R$ for sufficiently large n and N, provided that X_i and Z_i are uncorrelated. Under model (15) with $E(\varepsilon_i^2) = \sigma^2 X_i^{\delta}$ it holds that $\rho_z < -1/(N-1)$ when $\delta > 1$. Hence, for this type of data $\hat{Y}_{\text{PPS}}$ is to be preferred. Moreover, it emerges from (14) and (16) that for $\delta = 2$ the relative efficiency of PPS sampling compared to that of the ratio estimator is increasing when CV_x is increasing. In addition, $\hat{Y}_R$ is to be preferred when the data correspond to a model with $\delta < 1$. These findings are confirmed empirically with a simulation study using two different data sets. When model (15) is not applicable, the relative efficiency of $\hat{Y}_{\text{PPS}}$ is given by (14) provided n is large and N is relatively larger. In practice the unknown ρ_z in (14) is replaced by $\hat{\rho}_{z9}$. The fact that $n << N$ does not necessarily mean that the factor $(n-1)\rho_z$ in (3) is always negligible.

Acknowledgements

The views expressed in the article are those of the author and do not necessarily reflect the policy of Statistics Netherlands. The author would like to thank Peter-Paul de Wolf, Sander Scholtus, the Associate Editor and two anonymous referees for their helpful suggestions and corrections.

Appendix A

A counterexample

Equations (5) and (7) cannot always be used for randomized PPS sampling when n and N are of the same order while X_i and Z_i are correlated. To see that, consider a population U consisting of two groups U_1 and U_2 with means $\bar{Y}_1$ and $\bar{Y}_2$, respectively. Both stratum sizes are $N/2$. Let s be a randomized PPS sample of size $n = 3N/4$ from the whole population U. Let the X_i be such that

$$\pi_i = nX_i = \begin{cases} 1 & \text{if} \quad i \in U_1 \\ 0.5 & \text{if} \quad i \in U_2. \end{cases}$$

Obviously, group 1 does not contribute to the variance. The selected units in s from U_2 constitute an ordinary SRS sample of size $N/4$. Hence, for randomized PPS sampling the correct variance formula in this example is

$$\text{var}(\hat{Y}_{\text{PPS}}) = \left(\frac{N}{2}\right)^2 \left(1 - \frac{1}{2}\right) \frac{S_{y2}^2}{N/4} = \frac{NS_{y2}^2}{2},$$

and

$$S_{y2}^2 = \frac{2}{N-2} \sum_{i \in U_2} (Y_{2i} - \bar{Y}_2)^2.$$

However, approximation (11) gives an entirely different, larger outcome unless $\bar{Y}_1 = 2\bar{Y}_2$.

References

Brewer, K.R.W. (1963a). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, 5, 5-13.

Brewer, K.R.W. (1963b). Ratio estimation and finite population: Some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.

Brewer, K.R.W., and Donadio, M.E. (2003). The high entropy variance of the Horvitz-Thompson estimator. *Survey Methodology*, 29, 189-196.

Cochran, W.G. (1977). *Sampling Techniques.* New York: John Wiley & Sons, Inc.

Connor, W.S. (1966). An exact formula for the probability that two specified sampling units will occur in a sample drawn with unequal probabilities and without replacement. *Journal of the American Statistical Association*, 61, 384-390.

Cumberland, W.G., and Royall, R.M. (1981). Prediction models and unequal probability sampling. *Journal of the Royal Statistical Society*, B, 43, 353-367.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25, 193-203.

Durbin, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. *Applied Statistics*, 16, 152-164.

Enthoven, L. (2007). *Cohort calculations* (in Dutch). Report MIC-2007-21, Statistics Netherlands, Voorburg.

Foreman, E.K., and Brewer, K.R.W. (1971). The efficient use of supplementary information in standard sampling procedures. *Journal of the Royal Statistical Society*, B, 33, 391-400.

Gabler, S. (1984). On unequal probability sampling: Sufficient conditions for the superiority of sampling without replacement. *Biometrika*, 71, 171-175.

Hansen, M.H., and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.

Hartley, H.O., and Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.

Hidiroglou, M.A., and Gray, G.B. (1980). Construction of joint probability of selection for systematic P.P.S. sampling. *Applied Statistics*, 29, 107-112.

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

Knottnerus, P. (2003). *Sample Survey Theory*: *Some Pythagorean Perspectives*. New York: Springer-Verlag.

Kott, P.S. (1988). Model-based finite population correction for the Horvitz-Thompson estimator. *Biometrika*, 75, 797-799.

Kott, P.S. (2005). A note on the Hartley-Rao variance estimator. *Journal of Official Statistics*, 21, 433-439.

Qualité, L. (2008). A comparison of conditional Poisson sampling versus unequal probability sampling with replacement. *Journal of Statistical Planning and Inference*, 138, 1428-1432.

Rosén, B. (1997). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, 62, 159-191.

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

Sen, A.R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5, 119-127.

Stehman, S.V., and Overton, W.S. (1994). Comparison of variance estimators of the Horvitz-Thompson estimator for randomized variable probability systematic sampling. *Journal of the American Statistical Association*, 89, 30-43.

Thompson, M.E., and Wu, C. (2008). Simulation-based randomized systematic PPS sampling under substitution of units. *Survey Methodology*, 34, 3-10.

Yates, F., and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, B, 15, 253-261.