On the efficiency of randomized PPS sampling with an application to the Producer Price Index

Paul Knottnerus

Discussion paper (09014)

Statistics Netherlands

The Hague/Heerlen, 2009

Explanation of symbols

.	= data not available
*	= provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2005-2006	= 2005 to 2006 inclusive
2005/2006	= average of 2005 up to and including 2006
2005/’06	= crop year, financial year, school year etc. beginning in 2005 and ending in 2006
2003/’04–2005/’06	= crop year, financial year, etc. 2003/’04 to 2005/’06 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands - Facility Services

Cover
TelDesign, Rotterdam

Information
Telephone .. +31 88 570 70 70
Telefax .. +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax .. +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

6008309014 X-10

On the efficiency of randomized PPS sampling with an application to the Producer Price Index

Paul Knottnerus

Summary: This paper examines the efficiency of the Horvitz-Thompson estimator from a systematic probability proportional to size sample drawn from a randomly ordered list. Moreover, the efficiency is compared with that of an ordinary ratio estimator. The results are demonstrated by means of a simulation study with Dutch data on the Producer Price Index. The discussion on the efficiency includes a comparison with rejected Poisson sampling.

Keywords: efficiency, Horvitz-Thompson estimator, optimal allocation, Producer Price Index, probability proportional to size, ratio estimator, variance, rejective Poisson sampling.

1. Introduction

When the study variable y is more or less proportional to a size variable x, one may use the ratio estimator from a simple random sample without replacement (SRS). Another widely used 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 randomized PPS sample.

In recent years several authors paid attention to variance estimation procedures for randomized PPS samples. See, among others, Brewer and Donadio (2003), 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 they only consider cases where the sample size n is fixed while the population size N is increasing. Moreover, these comparisons don't give insight into the variance change due to randomized PPS sampling. 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. However, a formula for the efficiency is missing.

The main aim of this paper is to derive formulas for the efficiency of the randomized PPS estimator relative to the ratio estimator. These formulas take into account the finite population corrections for an arbitrary sample size n. Besides, to illustrate the difference between both estimators, we present a simple formula for the sample size change due to PPS estimator compared to a ratio estimator with the same variance. 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. Furthermore, it is shown that the corresponding variance estimator for randomized PPS sampling is 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 describes an example with data on the Price Producer Index in The Netherlands in order to illustrate the substantial efficiency gain that might be obtained in practice. A counterexample is included to show that PPS is not always advantageous. Section 5 discusses the relationship between rejective Poisson sampling and randomized PPS sampling, including the corresponding variance approximations and their (in)validity for $n \to \infty$. 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 $\hat{Y}_{HT} = \Sigma_{i \in s} Y_i / \pi_i$. Assuming that the π_i are proportional to the sizes X_i and that $X = \Sigma_{i \in U} X_i = 1$, it holds that $\pi_i = nX_i$; it is also assumed that $X_i \leq 1/n$. Defining $Z_i = Y_i / X_i$, we can write Y as a weighted mean of the Z_i, i.e., $Y = \mu_z = \Sigma_{i \in U} X_i Z_i$. Likewise, we can write the HT or randomized PPS estimator for Y as $\hat{Y}_{HT} = \hat{Y}_{PPS} = \bar{z}_s$ where $\bar{z}_s$ stands for the sample mean of the Z_i.

The variance of the randomized PPS estimator equals

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

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

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). Because (1) and (2) are somewhat inconvenient for a further analysis, we prefer the following alternative expression for the variance

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

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

$$\rho_z = \sum_{i \in U} \sum_{\substack{j \in U \\ j \neq i}} \frac{\pi_{ij}}{n(n-1)} \left(\frac{z_i - \mu_z}{\sigma_z} \right) \left(\frac{z_j - \mu_z}{\sigma_z} \right) \qquad (4)$$

For a proof of (3), see Knottnerus (2003, page 103). Recall 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 et al. (1992, page 79). Furthermore, 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. For instance, for [SRS] sampling with[out] replacement $\rho_z = 0$ $[\rho_z = -1/(N-1)]$.

Although for randomized PPS sampling exact expressions for the π_{ij} are available, these calculations might 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)

$$\begin{aligned} \pi_{ijK} &= n(n-1) \frac{X_i X_j (1 - X_i - X_j)}{\gamma (1 - 2X_i)(1 - 2X_j)} \\ \gamma &= \frac{1}{2} + \frac{1}{2} \sum_{i \in U} \frac{X_i}{1 - 2X_i} \end{aligned} \qquad (5)$$

According to the author these π_{ijK} satisfy the second-order restrictions for the π_{ij}

$$\begin{aligned} &\sum_{i,j \in U (j \neq i)} \pi_{ij} = n(n-1) \\ &\sum_{j \in U (j \neq i)} \pi_{ij} = (n-1)\pi_i \end{aligned}$$

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

$$\begin{aligned} \pi_{ijK} &= \pi_i \pi_j (c_i + c_j)/2 \\ c_i &= (n-1)/n\gamma(1 - 2X_i) \end{aligned}$$

An implication of approximation (5) is that $\pi_{ijK} / n(n-1)$ does not depend on n. Hence, the corresponding approximation of ρ_z doesn't depend on n, provided that $n < 1/X_i$ (i=1,...,N). This would also occur when we had used the approximation proposed by Hartley and Rao (1962) for randomized PPS sampling

$$\pi_{ijHR} = n(n-1)X_i X_j\{1 + X_i + X_j - \mu_x + 2(X_i^2 + X_j^2 + X_i X_j) \\ -3\mu_x(X_i + X_j - \mu_x - 2\Sigma_{i\in U} X_i^3)\} \tag{6}$$

Obviously, $\pi_{ijHR} / n(n-1)$ doesn't depend on *n*. At the time Hartley and Rao assumed that *n*=*O*(1) for $N \to \infty$. Following Thompson and Wu (2008), it is now believed that approximation (6) is valid when *n/ N*=*o*(1) for $N \to \infty$. In section 5 we address this issue in more detail. For the meaning of O-, o-, O_p- and o_p-symbols, see Knottnerus (2003, pages 140-1).

The approach proposed here is somewhat different from Knottnerus (2003). In order to get convenient expressions, rewrite (5) as

$$\pi_{ijK} = n(n-1)\frac{X_i X_j}{\gamma}\left(\frac{1/2}{1-2X_i} + \frac{1/2}{1-2X_j}\right) \tag{7}$$

Substituting (7) into (4), we obtain a new, simple approximation for ρ_z

$$\begin{aligned}\rho_z &= \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}} \frac{X_i X_j}{\gamma}\left(\frac{1/2}{1-2X_i} + \frac{1/2}{1-2X_j}\right)\left(\frac{Z_i - Y}{\sigma_z}\right)\left(\frac{Z_j - Y}{\sigma_z}\right) \\ &= \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}} \frac{X_i X_j}{\gamma}\left(\frac{1}{1-2X_i}\right)\left(\frac{Z_i - Y}{\sigma_z}\right)\left(\frac{Z_j - Y}{\sigma_z}\right) \\ &= 0 - \sum_{i\in U} \frac{X_i^2}{\gamma(1-2X_i)}\left(\frac{Z_i - Y}{\sigma_z}\right)^2 \end{aligned} \tag{8}$$

In the second line use is made of $\sum_{i,j} m_{ij} v_i = \sum_{i,j} m_{ij} v_j$ when $m_{ij} = m_{ji}$ and in the last line that $\sum_{j\in U} X_j (Z_j - Y) = 0.$ In the following theorem it is shown that under some mild regularity conditions (8) can be further simplified. The symbols μ_x and σ_x^2 in the theorem are used in analogy with μ_z and σ_z^2 in (3).

Theorem 1. Suppose that there is a constant *c* such that $V_x / \overline{X} < c$ and $\sigma_x / \mu_x < c$, where $\overline{X}$ and V_x^2 stand for the population mean and variance of *x*, respectively. Furthermore, suppose that $(Z_i - Y)/\sigma_z = O(1)$ as $N \to \infty$. Then ρ_z from (8) can be written for $N \to \infty$ as

$$\rho_z = -\frac{\sum_{i\in U} X_i^2 (Z_i - Y)^2}{\sum_{i\in U} X_i\,(Z_i - Y)^2}[1 + O(\frac{1}{N})] + O(\frac{1}{N^2}) \tag{9}$$

Proof. Because $\overline{X} = 1/N,$ it follows from the above assumptions that the weighted mean μ_x $[= \Sigma X_i^2 = N(V_x^2 + \overline{X}^2)]$ is of order 1/*N* and hence, $\sigma_x = O(1/N).$ Also,

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

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

Using these results in combination with a Taylor series expansion of $(1-2X_i)^{-1}$ in (8), we obtain (9). This concludes the proof.

Substituting (9) into (3), we get an alternative expression for the variance

$$\begin{aligned} \text{var}(\hat{Y}_{PPS}) &= \frac{\sigma_z^2}{n} - \frac{n-1}{n}\sum_{i\in U} X_i^2 (Z_i - Y)^2 \\ &= \frac{1}{n}\sum_{i\in U} X_i\{1-(n-1)X_i\}(Z_i - Y)^2 \end{aligned} \qquad (10)$$

We conclude this section with a couple of remarks.

1. Approximation (9) also follows directly from substituting the very simple approximation $\pi_{ijAP} = n(n-1)X_i X_j$ into (4). However, direct use of π_{ijAP} in (1) or (2) for the SRS case with $X_i = X_j = 1/N$ may lead surprisingly to errors of more than 100% for populations with $\overline{Y} = V_y^2$; see Knottnerus (2003, pages 274-6). Hence, (1) and (2) might be more sensitive to errors in the π_{ij} than (3) and (4).

2. In order to estimate (3), denote the sample variance of the Z_i by s_z^2. Noting that

$$\begin{aligned} \sigma_z^2 &= \text{var}(z_{s1}) = \text{var}\{E(z_{s1}|s)\} + E\{\text{var}(z_{s1}|s)\} \\ &= \text{var}(\overline{z}_s) + E\left(\frac{n-1}{n}s_z^2\right) \end{aligned}$$

and using (3), it can be shown that $s_z^2/(1-\rho_z)$ is an unbiased estimator for σ_z^2. When ρ_z is small, the term $(1-\rho_z)$ can be neglected.

3. The parameter ρ_z from (8) can be estimated in practice by

$$\hat{\rho}_{z8} = -\frac{\sum_{i\in s} X_i (Z_i - \overline{z}_s)^2 / \hat{\gamma}(1-2X_i)}{\sum_{i\in s}(Z_i - \overline{z}_s)^2}$$

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

Because $\hat{\gamma} \geq 1$ and $X_i \leq 1/n$, we have $\hat{\rho}_{z8} \geq -1/(n-2)$. Likewise, ρ_z from (9) can be estimated by

$$\hat{\rho}_{z9} = -\frac{\sum_{i\in s} X_i (Z_i - \overline{z}_s)^2}{\sum_{i\in s}(Z_i - \overline{z}_s)^2} \geq \frac{-1}{n} > \frac{-1}{n-1}$$

Hence, substituting s_z^2 and $\hat{\rho}_{z9}$ into (3) leads to a nonnegative variance estimator for $\hat{Y}_{HT}$ with probability 1. This also holds for $\hat{\rho}_{z8}$ when all $X_i \le 1/(n+1)$.

5. Formula (3) for the variance is under a number of assumptions a convenient point of departure for deriving allocation formulas when a sample is to be drawn from a stratified population with PPS-samples within each stratum. For further details, see Appendix A.

3. Efficiency of the randomized PPS estimator

3.1 Efficicency formulas

Because X=1, the ratio estimator from an SRS sample for the population total Y becomes

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

The commonly used approximation for its variance is

$$\text{var}(\hat{Y}_R) = \frac{N(N-n)}{n(N-1)} \sum_{i \in U} X_i^2 (Z_i - Y)^2 \tag{11}$$

See Cochran (1977). From (3) and (11) it can be seen that the efficiency of $\hat{Y}_{PPS}$ compared to that of $\hat{Y}_R$ can be written as

$$Eff_{P/R} = \frac{\text{var}(\hat{Y}_R)}{\text{var}(\hat{Y}_{PPS})} = \frac{(N-n)\sum_{i \in U} X_i^2 (Z_i - Y)^2}{\{1+(n-1)\rho_z\}\sigma_z^2} \tag{12}$$

where we assumed that $N/(N-1) \approx 1$. Combining (9) and (12) gives

$$Eff_{P/R} = \frac{-(N-n)\rho_z}{1+(n-1)\rho_z} \tag{13}$$

Since $Eff_{P/R} = 1$ for $\rho_z = -1/(N-1)$, this means that PPS sampling is to be preferred when $\rho_z < -1/(N-1)$.

To get more insight into the magnitude of ρ_z suppose that the data pattern of the Y_i can be described by

$$Y_i = \mu X_i + \varepsilon_i \qquad (i=1,\ldots,N). \tag{14}$$

with $E(\varepsilon_i | X_i) = 0$ and $E(\varepsilon_i^2 | X_i) = \sigma^2 X_i^{\delta}$. Consequently, for the Z_i we have $Z_i = \mu + u_i$ with $E(u_i | X_i) = 0$ and $E(u_i^2 | X_i) = \sigma^2 X_i^{\delta - 2}$. According to Kott

(1988), δ often lies between 1 and 2. However, unlike Kott we don't assume that the disturbances are uncorrelated. See also Brewer (1963b). In fact, the only point of interest in (9) is the pattern of the terms $(Z_i - Y)^2$ irrespective of the underlying autocorrelation structure of the data.

Assuming that N is sufficiently large, we can replace Y as well as the numerator and denominator in (9) by their model expectations. This yields

$$\rho_z = -\frac{\Sigma_{i\in U} X_i^{\delta}}{\Sigma_{i\in U} X_i^{\delta-1}} \qquad (15)$$

In the next subsections we look at different situations.

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

For $\delta = 2$ (15) gives $\rho_z = -\Sigma_{i\in U} X_i^2$ which can also be written as

$$\rho_z = -\frac{1}{N}(1 + CV_x^2) \qquad (16)$$

because

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

where $\overline{X} = 1/N$ and CV_x stands for the coefficient of variation of the X_i. Substituting (16) into (13) 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, $Eff_{P/R} = 1$ where we ignored the factor $(N+1)/N$.

In order to demonstrate the efficiency gain of randomized PPS sampling for $\delta = 2$ in a somewhat different way, it is useful to notice that substituting $n = n_{PPS}(1 + CV_x^2)$ into (11) leads to the same outcome as (3) and (9) with n_{PPS} instead of n. Hence, when $CV_x = 1.5$, randomized PPS sampling with sample size n_{PPS} =100 is as efficient as the ratio estimator from a SRS sample of size $n_{SRS} = 325$. More generally, it follows from (13) that a ratio estimator from an SRS sample of size n_{SRS} is as efficient as a PPS sample of size n_{PPS} if

$$n_{SRS} = n_{PPS} - 1 + N + \rho_z^{-1}$$

In section 5 it is shown this relationship is also applicable when $n \neq o(N)$ as $N \to \infty$ provided that according to (14) the Z_i and X_i are uncorrelated.

3.3 Efficiency of $\hat{Y}_{PPS}$ when δ=1

Another special case is $\delta = 1$. From (15) it follows that $\rho_z = -1/N$ when $\delta = 1$. Subsequently, it follows from (13) that $Eff_{P/R} = 1 + O(N^{-1})$ 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. For a formal proof, see Lemma 1. Hence, for $\delta < 1$ the randomized PPS estimator is less efficient than the ratio estimator and 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 (13) and (15), respectively. Then $Eff_{P/R}$ is a monotonically increasing function of δ.

Proof. Write ρ_z from (15) as a weighted mean of the (negative) X_i

$$\rho_z = -\mu_x(\delta) = -\sum_{i \in U} w_i X_i$$
$$w_i = \frac{X_i^{\delta-1}}{\sum_{i \in U} X_i^{\delta-1}} \qquad [\mu_x = \mu_x(2)]$$

Assuming that $X_i > X_j \quad (i \neq j)$, it holds that $h(\delta) = w_i / w_j = (X_i / X_j)^{\delta-1}$ is increasing in δ. Hence, the weight of the larger X_i is increasing compared to that of X_j when δ is increasing. This means that $\mu_x(\delta)$ is increasing and ρ_z is decreasing in δ. Since $Eff_{P/R}$ is a decreasing function of ρ_z as can be seen from (13), $Eff_{P/R}$ is increasing in δ. This concludes the proof.

3.4 An alternative structure among the disturbances

A third and last data pattern we look at in this section is the case where the variance of the disturbances in (14) is of the form

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

See Kott (1988). For this case we obtain in analogy with (15)

$$\rho_z = -\sum_{i \in U} w_i X_i$$
$$w_i = \frac{1 + \varphi X_i}{\sum_{i \in U} (1 + \varphi X_i)} \qquad (\varphi = c_2 / c_1)$$

For $\varphi = 0$ we obtain simply $\rho_z = -1/N$. Hence, for $\varphi = 0$ PPS sampling is as efficient as the ordinary ratio estimator from SRS sampling. Along the same lines as in 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.

4. An application to the Producer Price Index

The Producer Price Index (PPI) in The Netherlands is based on about 2500 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 stands for the (relative) turnover of that commodity of establishment *i* in the basic period $(\Sigma X_i = 1)$.

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. For these data we compare the variance of the ratio estimator from an SRS sample with the variance of the HT estimator from a randomized PPS sample. For both samples *n*=9. Applying (11) to these data gives $\mathrm{var}(\hat{Y}_R) = 101$. If the sample had been drawn with replacement the variance would have been 116. Applying (3) and (8) for a randomized PPS sample gives $\mathrm{var}(\hat{Y}_{PPS,\gamma}) = 29.9$. This outcome takes γ into account and lies close to the result $V_{PPS}^{(sim)} = 29.2$ 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 (11) for $\mathrm{var}(\hat{Y}_R)$ is only asymptotically valid, we also carried out simulations for evaluating the mean square error (MSE) of $\hat{Y}_R$ resulting in $MSE_R^{(sim)} = 108$. This confirms the conjecture that (11) 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 (13) suggests. In addition, the bias of 0.7 found in the simulations was in this case rather small compared to the variance.

Furthermore, it is noteworthy that the simpler formula (9) for ρ_z in combination with (3) gives $\mathrm{var}(\hat{Y}_{PPS}) = 30.7$. This is almost the same result as that from (8) although *N*=70 is not very large. With replacement the PPS variance would have been 43.8, almost 50% more. For $n_{EAZT} = 24$ formula (11) gives about the same outcome as (3) with $n_{PPS} = 9$. Hence, the sample sizes differ a factor 2.7 which is more or less in line with the factor $(1 + CV_x^2) = 3.1$ as we have seen in section 3

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

i	price change	turnover	*i*	price change	turnover
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

This should not be surprising since the price changes and their variability hardly depend on the sizes of the company. For instance, a double log regression

$$\ln(Z_i - Y)^2 = \alpha + \beta \ln X_i + v_i \qquad (17)$$

results in the estimate $\hat{\beta} = 0.07$ for the data in Table 1. This corresponds with $\hat{\delta} = 2.07$ for the disturbances in (14) which explains the superiority of randomized PPS sampling for this type of data pattern. Also for other commodities $\hat{\delta}$ often was about 2; see Enthoven (2007).

We conclude this section with a small example in order to show 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 (11); simulations give $MSE_R^{(sim)} = 323$. 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 (8) gives the same outcome of 557. This would also be the correct variance if the randomized PPS sample had been drawn according to Brewer (1963a) or Durbin (1967). Combination of (3) and (9) gives a slightly different value 556. Regression (17) with the data from Table 2 yields $\hat{\beta} = -3.0$ and hence, $\hat{\delta} = -1.0$. In addition, the ordinary direct estimator $N\bar{y}_s$ from an SRS sample has a variance of 356 which is even better here than randomized PPS sampling. Hence, for this type of data pattern 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). So the data Y_i $(= X_i Z_i)$ in Table 2 certainly don't 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. Relationship with rejective Poisson sampling

A related sampling design is a Poisson sample with inclusion probabilities $\pi_i = nX_i$ given the condition that the sample size is *n*. Often one calls this design rejective Poisson sampling. For this design Hájek (1964, page 1520) has shown that the variance of the corresponding estimator $\hat{Y}_{PO|n}$ for *Y* can be approximated by

$$\begin{aligned}
&\mathrm{var}(\hat{Y}_{PO|n}) = \frac{1}{n}\sum_{i\in U} X_i(1-nX_i)(Z_i - Y^*)^2 \qquad (18)\\
&Y^* = \sum\nolimits_{i\in U} \alpha_i Z_i\\
&\alpha_i = X_i(1-\pi_i)/d\\
&d = \sum\nolimits_{i\in U} X_i(1-\pi_i) = 1 - n\mu_x
\end{aligned}$$

provided that $nd \to \infty$. Hájek used for the derivation of (18) the following approximation for the π_{ij}

$$\begin{aligned}
\pi_{ijH} &= n^2 X_i X_j \left\{1 - \frac{(1-nX_i)(1-nX_j)}{nd}\right\} \qquad (19)\\
&\approx n(n-1)X_i X_j (1 - nX_i X_j / d)
\end{aligned}$$

For the sake of convenience, we have dropped some asymptotically irrelevant terms in the last line in order to derive a simple formula for the corresponding ρ_{zH} for rejective sampling; see Theorem 3. The main difference between (18) and (10) is that *Y* is replaced by Y^*. Consequently, when the Z_i and X_i were generated independently, (10) and (18) are asymptotically equivalent. More generally, the following theorem states that under some mild regularity conditions (10) and (18) are asymptotically equivalent for $n, N \to \infty$ irrespective of the data pattern of the Z_i provided that $n/N=o(1)$.

Theorem 2. Let $\breve{\Delta}$ denote the relative difference between (18) and (10). Then

$$\breve{\Delta} = \frac{\mathrm{var}(\hat{Y}_{PO|n}) - \mathrm{var}(\hat{Y}_{PPS})}{\mathrm{var}(\hat{Y}_{PPS})} = \frac{n^2 (r_{xz}\sigma_x)^2}{d\{1+(n-1)\rho_{z9}\}} - \frac{\rho_{z9}}{1+(n-1)\rho_{z9}} \qquad (20)$$

where ρ_{z9} is given by (9) and r_{xz} is defined by

$$r_{xz} = \sum_{i\in U} X_i \left(\frac{X_i - \mu_x}{\sigma_x}\right)\left(\frac{Z_i - \mu_z}{\sigma_z}\right)$$

Furthermore, suppose that $V_x / \overline{X} = O(1)$, $\sigma_x / \mu_x = O(1)$, and there is a constant *c* such that $\rho_{z9} < -c/N < 0$ and $d\{1+(n-1)\rho_{z9}\} > c > 0$ as $N, n \to \infty$. Also assume that there exists an $\alpha > 0$ such that $n/N = O(1/N^{\alpha})$ as $N \to \infty$. Then for $N, n \to \infty$

$$\breve{\Delta} = O(N^{-2\alpha}) + O(n^{-1})$$
$$\frac{\breve{\Delta}}{(n-1)\rho_{z9}} = O(N^{-\alpha}) + O(n^{-1}) \qquad (21)$$

Comment. The meaning of (21) is that under the given assumptions the approximation error in (10) for rejective Poisson sampling is much smaller than the variance reduction due the non-replacement feature as $N, n \to \infty$ and $n/N = O(1/N^{\alpha})$.

Proof. Since $Y^* = \Sigma_{i \in U} \alpha_i Z_i$, we have

$$\sum_{i \in U} \alpha_i (Z_i - Y)^2 - \sum_{i \in U} \alpha_i (Z_i - Y^*)^2 = (Y - Y^*)^2 \qquad (22)$$

Furthermore, (10) can be written as

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

Hence, the difference between (10) and (18), denoted by Δ, can be written as

$$\begin{aligned} \Delta &= \frac{d(Y - Y^*)^2}{n} + \frac{1}{n} \sum_{i \in U} X_i^2 (Z_i - Y)^2 \\ &= \frac{d(Y - Y^*)^2}{n} - \frac{\rho_{z9} \sigma_z^2}{n} \end{aligned} \qquad (23)$$

Moreover,

$$\begin{aligned} Y - Y^* &= \sum\nolimits_{i \in U} (X_i - \alpha_i)(Z_i - \mu_z) \\ &= \sum\nolimits_{i \in U} X_i \{1 - (1 - nX_i)/d\}(Z_i - \mu_z) \\ &= \frac{n}{d} \sum\nolimits_{i \in U} X_i (X_i - \mu_x)(Z_i - \mu_z) \\ &= \frac{n r_{xz} \sigma_x \sigma_z}{d} \end{aligned} \qquad (24)$$

Substituting (24) into (23) and dividing the result by (10) gives (20). Next, in analogy with the proof of Theorem 1 it follows that $\sigma_x = O(1/N)$. Consequently, $n^2 \sigma_x^2 = O(n^2 / N^2) = O(N^{-2\alpha})$ so that $\breve{\Delta} = O(N^{-2\alpha}) + O(n^{-1})$ as $N, n \to \infty$; note that $\rho_{z9} = O(n^{-1})$ because ρ_{z9} can be seen as a weighted mean of the (negative) X_i

and $X_i \le 1/n$. In addition, (21) follows from (20) because according to the above assumption $-(n-1)\rho_{z9} > nc/2N$. This concludes the proof.

Apart from the estimation by means of (10), the variance in (18) can be estimated directly by

$$\hat{\mathrm{var}}(\hat{Y}_{PO|n}) = \frac{1}{n(n-1)}\sum_{i\in s}(1-nX_i)(Z_i-\hat{Y}^*)^2$$
$$\hat{Y}^* = \frac{\sum_{i\in s}(1-\pi_i)Z_i}{\sum_{i\in s}(1-\pi_i)}$$

See Hájek (1964, page 1520) and Berger (2004). The following theorem shows how the variance of $\hat{Y}_{PO|n}$ can be written in the form of (3).

Theorem 3. For $n, N \to \infty,\ \ d > c > 0$, and under the assumptions of Theorem 1, (18) is asymptotically equivalent with

$$\mathrm{var}(\hat{Y}_{PO|n}) = \{1+(n-1)\rho_{zH}\}\frac{\sigma_z^2}{n}$$
$$\rho_{zH} = \rho_{z9} - \frac{nr_{xz}^2\sigma_x^2}{d} + O(\frac{1}{N^2}) \tag{25}$$

Proof. Denote $(Z_i - \mu_z)$ by z_i and $(X_i - \mu_x)$ by x_i. Substituting (19) into (4) and ignoring the asymptotically irrelevant terms in π_{ijH} we get in analogy with (24)

$$\begin{aligned}\rho_{zH} &= \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}} X_i X_j (1-\frac{nX_iX_j}{d}) z_i z_j / \sigma_z^2 \\ &= \rho_{z9} - \frac{n}{d}\sum_{i\in U} X_i^2 z_i (\sum_{j\in U} X_j x_j z_j - X_i^2 z_i)/\sigma_z^2 \\ &= \rho_{z9} - \frac{n}{d}\sum_{i\in U} X_i^2 z_i (r_{xz}\sigma_x\sigma_z - X_i^2 z_i)/\sigma_z^2 \\ &= \rho_{z9} - \frac{nr_{xz}^2\sigma_x^2}{d} + O(\frac{1}{N^2})\end{aligned}$$

Note that in the last line we used $z_i/\sigma_z = O(1)$ and $nX_i = O(1)$. This concludes the proof.

Following Brewer and Donadio (2003, page 190), randomized PPS sampling can be seen as a high entropy design; see also Rosén (1997) and Berger (1998). This means that approximation (18) can be used for randomized PPS sampling as well provided that $nd \to \infty$. Moreover, Theorem 2 reconfirms that for randomized PPS sampling

(10) can be used for large n provided that $n/N = O(N^{-\alpha})$ $(\alpha > 0)$. In addition, if $n/N \neq o(1)$, (10) can still be used when the Z_i and X_i have a zero covariance as can be seen from (25). Furthermore, when $n/N \neq o(1)$ and $r_{xz} \neq 0$, it follows from Theorem 2 that formula (13) for the efficiency of the randomized PPS estimator relative to that of the ratio estimator should be modified as follows

$$Eff^*_{P/R} = \frac{-(N-n)\rho_{z9}}{(1+n\rho_{z9})(1-\breve{\Delta})}$$

as $n \to \infty$ and $d > c > 0$. According to Theorem 2 this modification of (13) is necessary as compensation for its underestimation by $100\breve{\Delta}\%$ when $n/N \neq o(1)$ and $r_{xz} \neq 0$.

Furthermore, it should be noted that for $\pi_i = n/N$ (i=1,…,N) the actual inclusion probabilities $\pi_{iPO|n}$ for rejective sampling are equal to the original π_i. In contrast, for $\pi_i \neq n/N$ the actual $\pi_{iPO|n}$ need not be equal to $\pi_i = nX_i$ when n is small. Therefore, (18) is not to be recommended for a PPS variance approximation when n is small. For instance, for the data in Table 2 (18) gives as variance 494 with $n=2$. This is an underestimation of about 10%. However, for $nd \to \infty$ it is pointed out by Hájek (1964, page 1517) that $\pi_i / \pi_{iPO|n}$ tends to unity uniformly in i.

An approximation for the π_{ij} that can be used for small and large n is the following combination of (5) and (19)

$$\pi_{ijK\,\mathrm{mod}} = n(n-1)X_i X_j \left\{ \frac{1 - X_i - X_j}{\gamma(1-2X_i)(1-2X_j)} - \frac{nX_i X_j}{d} \right\} \tag{26}$$

The corresponding expression for ρ_z becomes in analogy with the proof of Theorem 3

$$\rho_{zK\,\mathrm{mod}} = \rho_{z8} - \frac{n}{d\sigma_z^2}(r_{xz}^2 \sigma_x^2 \sigma_z^2 - \sum_{i \in U} X_i^4 z_i^2) \tag{27}$$

Use of (26) in the two examples given in section 4 leads to more or less the same variances, i.e., 30.2 and 564, respectively.

In order to give some more insight into the difference between (10) and (18), we conclude this section with a counterexample that (10), including approximations (5) and (6) for the π_{ij}, need not be valid when $n/N \neq o(1)$. Consider a population U consisting of two strata U_1 and U_2 with means $\overline{Y}_1$ and $\overline{Y}_2$, 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, stratum 1 doesn't contribute to the variance. The selected elements in *s* from U_2 constitute an ordinary SRS sample of size $N/4$. Hence, in this case the correct variance formula for $\hat{Y}_{PPS}$ is

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

However, approximation (10) gives now an entirely different outcome unless $\bar{Y}_2 = 2\bar{Y}/3$; note that $Y^* = 3N\bar{Y}_2/2$. In contrast, (18) gives the correct outcome apart from a factor $(N-2)/N$. Also, (26) and (27) lead to an asymptotically correct answer. For instance, assuming that $Y_{2i} = \bar{Y}_2$ for all *i*, $n\rho_{zK\,\text{mod}}$ converges to the correct value -1 as $N \to \infty$ which value corresponds with a zero variance. Note that in this case with $Z_{2i} = Y^*$ use of (22) and (24) yields

$$n(\rho_{xz}\sigma_x\sigma_z)^2/\delta = \delta(Y - Y^*)^2/n = \delta\sum_{i\in U} \alpha_i (Z_i - Y)^2/n$$

6. Summary

This paper compares the efficiency of the HT estimator $\hat{Y}_{PPS}$ from a PPS sample with the efficiency of the classical ratio estimator $\hat{Y}_R$ from an SRS sample. It is assumed that for all elements of the population the size variable *x* is known. When the data patterns of the variables *x* and $z\ (= y/x)$ are such that the parameter $\rho_z < -1/(N-1)$, it can be shown that $\hat{Y}_{PPS}$ is more efficient than $\hat{Y}_R$ as $N \to \infty$. Under model (14) with $E(\varepsilon_i^2 | X_i) = \sigma^2 X_i^\delta$ it holds that $\rho_z < -1/(N-1)$ when $\delta > 1$. According to Kott (1988) δ often lies between 1 and 2. Hence, for this type of data pattern $\hat{Y}_{PPS}$ is to be preferred. Moreover, it emerges 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 for data patterns with $\delta < 1$. These findings are demonstrated by means of a simulation study with two different data sets.

The above results hold when $N \to \infty$ and $n/N = o(1)$ or when $N, n \to \infty$ and X_i and Z_i are uncorrelated. In case X_i and Z_i are correlated, the relative efficiency of PPS sampling is increasing when their squared correlation r_{xz}^2 is increasing provided $N, n \to \infty$ and $n/N \neq o(1)$.

Appendix A. Optimal stratum allocation in randomized PPS sampling

Suppose that Y is a sum of H stratum totals

$$Y = \sum_{h=1}^{H} Y_h$$

When randomized PPS sampling is used in all strata, the variance of $\hat{Y}_{PPS,ST}$ is equal to

$$\begin{aligned}
\operatorname{var}(\hat{Y}_{PPS,ST}) &= \sum_{h=1}^{H} \operatorname{var}(\hat{Y}_{h,PPS}) \\
&= \sum_{h=1}^{H} \{1 + (n_h - 1)\rho_h\} \frac{\sigma_h^2}{n_h} \\
&\approx \sum_{h=1}^{H} \frac{\sigma_h^2}{n_h} + \sum_{h=1}^{H} \rho_h \sigma_h^2
\end{aligned}$$

In the last line it is assumed that $(n_h - 1)/n_h \approx 1$. Under the assumption that the Z_i and X_i are uncorrelated, the approximations of ρ_h proposed so far are (asymptotically) independent of n_h. Hence, the allocation problem reduces to

$$\min \sum_{h=1}^{H} \frac{\sigma_h^2}{n_h} \quad \text{subject to} \quad \sum_{h=1}^{H} n_h = n$$

Assuming that the optimal n_h obey $n_h < 1/X_{hi}$ (i=1, …, N_h) the optimal allocation for randomized PPS sampling is in analogy with the Neyman allocation equal to

$$n_h = \frac{\sigma_h}{\sum_{h=1}^{H} \sigma_h} n$$

An unbiased estimator for σ_h^2 is

$$\hat{\sigma}_h^2 = \frac{s_h^2}{1 - \rho_h}$$

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i \in S_h} (Z_{hi} - \bar{z}_h)^2$$

When the differences between the ρ_h are small or when their absolute values are small, they can be ignored.

In addition, note that a price index Y of H commodities can be written as

$$Y = \sum_{h=1}^{H} W_h Y_h = \sum_{h=1}^{H} W_h \left(\sum_{i \in U_h} X_{hi} Z_{hi} \right)$$

with $\Sigma_h W_h = 1$. Along the same lines it can be shown that the optimal allocation becomes

$$n_h = \frac{W_h \sigma_h}{\sum_{h=1}^{H} W_h \sigma_h} n.$$

References

Berger, Y. (1998). Rate of convergence for asymptotic variance for the Horvitz-Thompson estimator, *Journal of Statistical Planning and Inference*, 74, 149-168.

Berger, Y. (2004). A simple variance estimator for unequal probability sampling without replacement, *Journal of Applied Statistics*, 31, 305-315.

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*, John Wiley & Sons, New York.

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.

Deville, J.C. (1999). Variance estimation for complex statisics 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), Statistics Netherlands, No. 2007-21-MIC, 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.

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

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

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*, Springer-Verlag, New York.

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). Rate of convergence for asymptotic variance for the Horvitz-Thompson estimator, *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*, Springer-Verlag, New York.

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.