Stockholm University

Master's Thesis in Statistics

Department of Statistics

Examensarbete i statistik för masterexamen,
Statistiska institutionen

A Without-replacement Fixed-sized Sampling Design

Edgar Bueno

Examensarbete 30 högskolepoäng, ht 2014

Handledare (supervisor): Michael Carlson

A without-replacement fixed-size sampling design

Edgar Bueno*

Abstract

A sampling design with unequal probabilities, called q-sampling, is introduced. It is shown that q-sampling is in fact, a sampling design and its associated inclusion probabilities are presented. The design is proposed to be coupled with the well known Horvitz-Thompson estimator or the proposed, and also unbiased, q-estimator. An approximate expression for the variance and a variance estimator for the strategy (q-sampling–q-estimator) are proposed. Performance of q-sampling is studied using a Monte-Carlo simulation study and is compared with three well known designs: simple random sampling, systematic sampling and Pareto sampling. The main conclusions are that, when coupled with the Horvitz-Thompson estimator, q-sampling behaves similar to simple random sampling. On the other hand, when coupled with the q-estimator, q-sampling lies in between simple random sampling and Pareto sampling.

Keywords: Sampling with unequal probabilities; Pareto sampling; systematic sampling; q-sampling; Monte-Carlo simulation.

*E-mail: embuenoc@hotmail.com. Supervisor: Michael Carlson

Acknowledgements

I would like to thank to my supervisor, Michael Carlson, for his time, patience and suggestions; to my parents, I owe it all to them; to my whole family for always being there for me; and also to Mónica, for her infinite patience and abnegation. I would also like to thank Colfuturo for the financial support during all this time.

Contents

Introduction

In a general setting of survey sampling we are interested in the population total of a variable y of the N elements that constitute the population of interest, i.e. we are interested in

$$t_y = \sum_U y_k$$

where y_k is the value of the variable of interest for the k-th element in the population, U.

In order to obtain an estimate of t_y a sample, s, is selected from U under a given design, $p(\cdot)$. Then, the observations of y for the elements in s are collected and an estimate is calculated by using a given estimator $Q(S)$. The combination of a sampling design and an estimator is called *strategy*.

It is known that when auxiliary variables are available, they can be used in the design, in the estimator or in both. In particular, when an auxiliary variable x that is approximately proportional to y is known for every element in U, it can be used in the design in order to reduce the variance of the estimator.

There are many designs that allow to select a sample with probabilities proportional to the size variable x. Among them, we can find random-size with replacement (e.g. Poisson with replacement), random-size without-replacement (e.g. Poisson sampling), fixed-size with-replacement (e.g. pps sampling) and fixed-size without replacement (e.g. πps sampling) designs. It is expected that the reduction in variance due to the use of the auxiliary variable is greater for the latter, but, on the other hand, some difficulties arise when selecting a sample under this approach.

It is important to note that the proportionality can be relative to different measures, for example, the inclusion probabilities (if we are interested in using the Horvitz-Thompson estimator) or the selection probability in each drawing (if we are interested in using the Hansen-Hurwitz estimator).

The problem of selecting a sample with selection probabilities strictly proportional to x has already been solved. On the other hand, to select a sample with inclusion probabilities strictly proportional to x has not been an easy task, especially if we take into account that a simple selection algorithm is strongly desirable and that we are usually interested in samples of size greater than two. Furthermore, if the Horvitz-Thompson estimator is to be used, it is also desirable that all the second-order inclusion probabilities be greater than zero, easy to calculate and the inclusion covariances should be smaller than zero.

Many available methods fulfill some, but not all, of the conditions above. For example, Brewer's method is defined for samples of size $n = 2$; Sunter's scheme is only "approximately"proportional; and Pareto sampling is assymptotically a πps design, but for small samples the results are only approximations.

Based on the above discussion, the main objective of this thesis is to propose a sampling design that uses an auxiliary variable to select a fixed-size without-replacement sample for any sample size n ($0 < n < N$). The main advantage of this design with respect to other methods is that the r-th ($0 < r \leq n$) order inclusion probabilities are positive and easily calculated. However, the design also has disadvantages that will be pointed out later.

A second objective of the thesis is to couple the proposed design with the Horvitz-

Thompson estimator (π-estimator):

$$\hat{t}_\pi = \sum_s \frac{y_k}{\pi_k}$$

where π_k is the inclusion probability of the k-th element.

It is known that if the π-values were strictly proportional to y, then, the variance of the π-estimator would be zero. Here arises the main disadvantage of the proposed design: it doesn't allow for selecting a (strict) πps for x variables that are severely right-skewed. As a consequence, the third objective of the thesis is to propose an estimator that is unbiased under the proposed design and has no restrictions regarding the distribution of the auxiliary variable x.

In the first section of the thesis some basic concepts will be defined. Also some examples of estimator and sampling designs will be presented. In the second section, the proposed design will be defined and it will be shown that it is, in fact, a sampling design. Also, an expression for the r-th order inclusion probabilities will be shown and an algorithm of selection will be given. In the final part of the section some examples will be developed. In the third section the strategy composed of the proposed design and the π-estimator will be discussed. In the fourth section the strategy composed of the proposed design and the proposed estimator will be discussed. A comparison of the two strategies developed in sections three and four, together with other strategies will be carried out in the fifth section.

1 Theoretical framework

This section is divided into three parts. Basic concepts, like sampling design and estimator, will be presented on the first part. In the second part the Horvitz-Thompson and the Hansen-Hurwitz estimators are defined. The third part will be devoted to presenting different sampling designs that use auxiliary information.

The definitions in the next section are based on those presented in Särndal *et. al.* (1992). Therefore, the reader is referred to that source for a more comprehensive description of the concepts. Also, the notation used here follows that used in Särndal *et. al.* (1992).

1.1 Basic concepts

Let U be a population composed of N elements labeled $\{1, 2, \cdots, N\}$, y_k and x_k the values, associated with the k-th element, of a variable of interest and an auxiliary variable, respectively. The y-values are assumed to be unknown prior to the sampling and only those selected in the sample will be known. On the other hand the x-values are assumed to be known beforehand for every element in the population.

A *sampling frame* that lists the N elements in U is assumed to be at hand. Formally *sampling frame* is defined as (Särndal *et. al.*, 1992)

> any material or device used to obtain observational access to the finite population of interest. It must be possible with the aid of the frame to (1) identify and select a sample in a way that respects a given probability design and (2) establish contact with selected elements.

We are interested in the population total of y, denoted by t_y, i.e. we are interested in

$$t_y = \sum_U y_k$$

It can be seen that t_y is known only through a complete enumeration of the y-values. This procedure is often very expensive or hard to achieve. Therefore, we usually observe only a subset of U and we use it to estimate the total of interest. This subset is called a *sample* and will be denoted by s. A sample is any subset of U (including the empty set and U itself) and it may include the same element more than once. Two main aspects are to be considered regarding a sample: how it is selected and how it will be used to obtain an estimate of the total. The former is related to the concept of *sampling design*; the latter, to the concept of *estimator*.

Let Ω^* be the set of subsets of U. A *sampling design* is any probability distribution over Ω^* and will be denoted by $p(s)$. In other words, a sampling design is the probability assigned to each possible sample. As $p(s)$ is a probability function over Ω^*, it satisfies the following conditions

$$\sum_{\Omega^*} p(s) = 1 \tag{1}$$

$$p(s) \geq 0 \text{ for every } s \in \Omega^* \tag{2}$$

Note that Ω^* is an infinite set. Usually a large number of elements (samples) in Ω^* have probability equal to zero. The *support* of a sampling design is the set of samples with probability strictly positive and will be denoted by Ω. So, conditions (1) and (2) can be rewritten in terms of the support as

$$\sum_{\Omega} p(s) = 1 \tag{3}$$

$$p(s) > 0 \text{ for every } s \in \Omega \tag{4}$$

Once the sampling design is defined, one sample must be selected according to the probabilities defined by the design. A simple method is to enumerate the support and select a sample by generating a random number from a uniform distribution. Unfortunately, although simple to explain, this method is usually impossible to carry out because the cardinality of Ω is very large even for small populations; therefore, methods that avoid the complete enumeration of the samples in the support but respect the probabilities given by the design must be used. These methods are called *sample selection schemes*.

Using a sample selection scheme one sample is selected, the y values for the elements in the sample are collected and an estimate for the total is obtained through an *estimator*. An *estimator* is any real-valued function of the sample. Although any statistic can be considered as an estimator, usually, the difference relies on the fact that an estimator has certain desired properties regarding a given parameter. In our case, the parameter of interest is the population total, t_y.

1.2 Estimators

Two estimators of the total will be described in this subsection: **i.** the Horvitz-Thompson estimator (Horvitz and Thompson, 1952), also called π-estimator that will be denoted by $\hat{t}_\pi$; and **ii.** the Hansen-Hurwitz estimator (Hansen and Hurwitz, 1943) that will be denoted by $\hat{t}_{\text{pwr}}$.

1.2.1 The Horvitz-Thompson estimator

For any k in U, let I_k a function that indicates whether k is included in the sample or not, i.e.

$$I_k = \begin{cases} 1 & k \text{ is included in the sample} \\ 0 & k \text{ is not included in the sample} \end{cases}$$

The expected value of the function I_k under the design $p(\cdot)$ is called the *inclusion probability* of k and will be denoted by π_k,

$$\pi_k = E_p(I_k) = \sum_{\Omega} p(s) I_k$$

In other words, the inclusion probability of k is the probability that the k-th element is actually selected in the sample.

The π-estimator is defined as

$$\hat{t}_\pi = \sum_s \frac{y_k}{\pi_k} \tag{5}$$

It can be shown that $\hat{t}_\pi$ is an unbiased estimator of the total,

$$E_p(\hat{t}_\pi) \equiv \sum_{\Omega} p(s) \hat{t}_\pi = t_y$$

Also, the value of the estimator varies according to the realized sample, so $\hat{t}_\pi$ has a variance. Before defining the variance, we need to define the *second order inclusion probabilities*.

In the same sense that π_k is the probability that the k-th element is selected in the sample, the *second order inclusion probability* of k and l is the probability that both elements, k and l, are selected simultaneously in the sample. This probability is denoted by π_{kl} and is formally defined as

$$\pi_{kl} = E_p(I_k I_l) = \sum_{\Omega} p(s) I_k I_l \tag{6}$$

Note that, by definition, $\pi_{kk} = \pi_k$. Higher order inclusion probabilities can be defined in an analogous way: The *r-th order inclusion probability* of $k_1, k_2, \cdots, k_r$ is the probability that the elements $k_1, k_2, \cdots, k_r$ are selected simultaneously in the sample. This probability is denoted by $\pi_{k_1 k_2 \cdots k_r}$ and is formally defined as

$$\pi_{k_1 k_2 \cdots k_r} = \sum_{\Omega} p(s) I_{k_1} I_{k_2} \cdots I_{k_r} \tag{7}$$

A design that satisfies

$$\pi_k > 0 \text{ and } \pi_{kl} > 0 \quad (k, l = 1, 2, \cdots, N)$$

is called *measurable*.

Now that the second order inclusion probabilities have been defined, the expression for the variance of the π-estimator can be presented:

$$V_p\left(\hat{t}_\pi\right) \equiv \sum_{\Omega} p(s) \left(\hat{t}_\pi - t_y\right)^2 = \sum_U \sum_U \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \tag{8}$$

where Δ_{kl} is the covariance between I_k and I_l,

$$\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$$

It can be seen that in order to calculate the variance of the π-estimator all the y-values are required. But we only know the y-values of those elements included in the sample, so $V_p\left(\hat{t}_\pi\right)$ is also a parameter that must be estimated. If $\pi_{kl} > 0$ for all $k, l \in U$, the statistic

$$\hat{V}_p\left(\hat{t}_\pi\right) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \tag{9}$$

is an unbiased estimator of $V_p\left(\hat{t}_\pi\right)$. An alternative expression for the variance of the π-estimator for fixed-size designs is

$$V_p\left(\hat{t}_\pi\right) \equiv \sum_\Omega p(s)\left(\hat{t}_\pi - t_y\right)^2 = -\frac{1}{2} \sum_U \sum_U \Delta_{kl} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \tag{10}$$

Also, for fixed-size designs, an alternative estimator of the variance is

$$\hat{V}_p\left(\hat{t}_\pi\right) = -\frac{1}{2} \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \tag{11}$$

If $\pi_{kl} > 0$ for all $k, l \in U$, the estimator is unbiased; furthermore, if $\Delta_{kl} < 0$ for all $k, l \in U$, it is nonnegative.

The double sum in the expressions for the variance, and its estimators, leads to hard calculations due to the number of terms involved in it. Other, but biased, estimators of the variance of the π-estimator have been proposed. Some of these do not need the second order inclusion probabilities or the covariances.

It is important to note that if we could make the π-values exactly proportional to the y-values (i.e. $\pi_k = \alpha y_k$ for all $k \in U$), the variance of the π-estimator would be equal to zero. Unfortunately, even when the π-values are usually defined through the design, the y-values remain unknown (only those selected in the sample will be known), so it is impossible to obtain a strict proportionality. Even so, often there is (at least) one variable that is available prior to the design stage and is correlated with the variable of interest. This variable, which is required to be greater than zero and known for every element in U, is often called an *auxiliary variable* and will be denoted by x.

We would be interested in a design that assigns inclusion probabilities proportional to the auxiliary variable, x, so that the variance of the π-estimator is reduced when x and y are well correlated. Such designs are commonly known as πps designs. The discussion of such designs will be relegated to Section 1.3.4.

1.2.2 The Hansen-Hurwitz estimator

Let $p_1, p_2, \cdots, p_N$ be a set of known values associated with the elements in U such that

$$p_k > 0 \text{ for all } k \in U \qquad \text{and} \qquad \sum_U p_k = 1$$

Consider the following sample selection scheme: One out of the N elements in U is selected using the p-values as selection probabilities. Once the element has been

selected it is placed back into the population. This experiment is repeated m times, so that m elements are selected.

The above scheme generates a with-replacement design often denoted as *pps*-sampling. This design will be discussed in Section 1.3.3.

For such a design, the estimator

$$\hat{t}_{\text{pwr}} = \frac{1}{m} \sum_{i=1}^{m} \frac{y_k}{p_k} \tag{12}$$

is unbiased for the total, t_y. This estimator will be called pwr-estimator.

Note that a different notation has been used in the sums for the π-estimator (5) and the pwr-estimator (12). This fact can be explained as follows: two different concepts of sample can be observed under a with-replacement design. One is the set of m selected elements; this set includes each element as many times as it has been selected and is called the *ordered sample*. The pwr-estimator uses this set in order to obtain an estimate. A second concept is the set of different elements selected and is called the *set sample*. The π-estimator uses this set.

The variance of the pwr-estimator can be expressed as

$$V_p\left(\hat{t}_{\text{pwr}}\right) \equiv \sum_{\Omega} p(s) \left(\hat{t}_{\text{pwr}} - t_y\right)^2 = \frac{1}{m} \sum_{U} p_k \left(\frac{y_k}{p_k} - t_y\right)^2 \tag{13}$$

An unbiased estimator of the variance is

$$\hat{V}_p\left(\hat{t}_{\text{pwr}}\right) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \left(\frac{y_k}{p_k} - \hat{t}_{\text{pwr}}\right)^2 \tag{14}$$

Unlike the expressions for the variance of the π-estimator, the expressions for the pwr-estimator are based on a single sum. This fact makes the calculation of the variances (and its estimates) notably simpler for the case of the pwr-estimator in comparison to the π-estimator.

In a sense analogous to that described for the π-estimator, the variance of the pwr-estimator can be strongly reduced through the use of an auxiliary variable that is highly correlated with the variable of interest: if we could get the p-values to be exactly proportional to the y-values, the variance of the pwr-estimator would be zero.

1.3 Sampling designs

Three characteristics can be used as a simple way to categorize a sample design:

1. whether it is with or without replacement;
2. whether it is of fixed or random size; and,
3. whether it uses auxiliary information or not.

It is often observed that without replacement are more efficient than with replacement designs; fixed size are more efficient than random size; and, designs that use auxiliary information are more efficient than those that do not.

These three characteristics generate eight groups that are listed as follows (an example of each case is listed between brackets):

1. No auxiliary info. - Random size - With rep. (Bernoulli with replacement);
2. No auxiliary info. - Random size - Without rep. (Bernoulli without rep.);
3. No auxiliary info. - Fixed size - With rep. (SRS with replacement);
4. No auxiliary info. - Fixed size - Without rep. (SRS without replacement);
5. Auxiliary info. - Random size - With rep. (Poisson with replacement);
6. Auxiliary info. - Random size - Without rep. (Poisson without replacement);
7. Auxiliary info. - Fixed size - With rep. (pps);
8. Auxiliary info. - Fixed size - Without rep. (πps)

For a comprehensive description of these designs, see, for example (Tillé, 2006) or (Särndal *et. al.*, 1992). Those designs that do not use auxiliary information will not be discussed in this document, the only exception being Simple Random Sampling without replacement. This design will be considered for two reasons: the theory behind it is useful to illustrate the concepts already defined in Section 1, and, this design is often considered as a reference when comparisons are to be made. Also, the Poisson with replacement design is not considered, given that it is known to be an inefficient design. On the other hand, Poisson sampling and *pps*-sampling are presented only with illustrative purposes, but they will not be used in the comparisons in Section 5.

1.3.1 Simple Random Sampling without replacement

Let Ω_{srs} be the set composed of the $\binom{N}{n}$ without replacement samples of size n out of N elements. Simple Random Sampling without replacement, denoted SRS, is the design that has Ω_{srs} as its support and assigns the same probability to each sample in it, i.e.

$$p(s) = \frac{1}{\binom{N}{n}}$$

One sample selection scheme to select a sample from a SRS is the following: associate, to each element $k \in U$, a value u_k such that u_k is a random number from a uniform distribution Unif(0,1) and then, select the n elements with the smallest u-values.

Under SRS the first and second order inclusion probabilities are

$$\pi_k = \frac{n}{N} \qquad \pi_{kl} = \frac{n(n-1)}{N(N-1)} \tag{15}$$

Then, the π-estimator takes the form

$$\hat{t}_\pi = \frac{N}{n} \sum_s y_k \tag{16}$$

The variance of $\hat{t}_\pi$ can be rewritten as

$$V\left(\hat{t}_\pi\right) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{y,U} \quad \text{where } S^2_{y,U} = \frac{1}{N-1}\sum_U \left(y_k - \bar{y}_U\right)^2 \text{ and } \bar{y}_U = \frac{1}{N}\sum_U y_k$$

and the estimator for the variance is

$$\hat{V}\left(\hat{t}_\pi\right) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{y,s} \quad \text{where } S^2_{y,s} = \frac{1}{n-1}\sum_s \left(y_k - \bar{y}_s\right)^2 \text{ and } \bar{y}_s = \frac{1}{n}\sum_s y_k \tag{17}$$

1.3.2 Poisson without replacement

Consider the following sample selection scheme. Associate, to each $k \in U$, two values: a first value, π^*_k $(0 < \pi^*_k \leq 1)$, and a second value, u_k such that u_k is a random number from a uniform distribution Unif(0,1) (note that the π^*-values are fixed and known, while the u-values are random). If $u_k < \pi^*_k$, element k is included in the sample.

The design generated by the above scheme is known as Poisson without replacement sampling. Note that the support for this design, Ω_{pois}, consists of the power set of U, i.e. the 2^N subsets of U. Under this scheme, the probability of each sample in the support can be written as

$$p(s) = \prod_s \pi^*_k \prod_{U-s} (1 - \pi^*_k)$$

Under Poisson sampling the first and second order inclusion probabilities are

$$\pi_k = \pi^*_k \qquad \pi_{kl} = \pi^*_k \pi^*_l$$

The expression for the π-estimator cannot be simplified, so

$$\hat{t}_\pi = \sum_s \frac{y_k}{\pi_k}$$

The variance of $\hat{t}_\pi$ can be rewritten as

$$V\left(\hat{t}_\pi\right) = \sum_U (1 - \pi_k)\frac{y_k^2}{\pi_k}$$

and the estimator for the variance is

$$\hat{V}\left(\hat{t}_\pi\right) = \sum_s (1 - \pi_k)\frac{y_k^2}{\pi_k^2}$$

Now, suppose that an auxiliary and positive variable x is available for every $k \in U$. Henceforth, we assume that all the x-values are such that $x_k < \frac{1}{n}$. We can define

$$\pi^*_k = n\frac{x_k}{t_x} \quad \text{where } t_x = \sum_U x_k$$

and n is the expected sample size.

If there is a high correlation between x and the (unknown) variable y, the variance will be reduced. In an extreme situation where $y_k = \alpha x_k$, the variance would be equal to zero.

1.3.3 Proportional-to-size sampling with-replacement

Consider again the sample selection scheme described in Section 1.2.2. The support generated by this scheme, Ω_{pps}, consists of the N^m with-replacement samples of size m out of N elements; and the probability associated with each sample is

$$p(s) = \prod_s p_k$$

This design is often coupled with the pwr-estimator. Expressions for the estimator, its variance and an unbiased estimator for the variance have been already presented in Section 1.2.2.

Suppose that an auxiliary and positive value x is available for every $k \in U$. We can define

$$p_k = \frac{x_k}{t_x} \quad \text{where } t_x = \sum_U x_k$$

As in the case of Poisson sampling, a high correlation between x and y will lead to a great reduction in the variance. A perfect linear relation between x and y would give a variance equal to zero.

Alternatively, the strategy (pps–π-estimator) can be considered. The first and second order inclusion probabilities are

$$\pi_k = 1 - (1 - p_k)^m \qquad \pi_{kl} = 1 - (1 - p_k)^m - (1 - p_l)^m + (1 - p_k - p_l)^m$$

Note that even under a perfect linear correlation between x and y, the variance for this strategy is not equal to zero. This is a good reason why the strategy (pps–pwr-estimator) is often preferred. Even so, sometimes smaller variances are obtained through the former strategy.

1.3.4 Proportional-to-size sampling without-replacement

As mentioned above, usually, fixed size designs are more efficient than random size; without replacement designs, are more efficient than with replacement; and, designs that use auxiliary information are more efficient than those that does not use it. So, we would expect a without-replacement fixed size with auxiliary information design to be the best combination for a design. Many of such designs already exist. Tillé (2006) and Hanif and Brewer (1980), for example, present reviews of available designs.

One design that satisfies the three conditions above is the so called πps design, where the inclusion probabilities are created proportionally to the auxiliary variable, so that the variance of the strategy (πps–π-estimator) is expected to be small. Even when there are many schemes that allow for selecting a sample that follows a πps design, other properties are strongly desired besides those already mentioned. Among these properties we can mention: the scheme should be easily implemented, it should be useful for selecting samples of size greater than two, and the second order inclusion probabilities should be easily computable and greater than zero. If an estimator different than the π-estimator may be considered, the last condition can be ignored.

Unfortunately, it is not easy to find a scheme that simultaneously satisfies all the conditions above. Two schemes that offer interesting properties are: systematic πps with fixed ordered frame and Pareto sampling. These schemes are briefly described as follows.

Systematic πps (There are some variations of this scheme. The one described here is usually called *systematic πps with fixed ordered frame*. More information about the design can be found in Tillé (2006).) Let x_k be the value of the auxiliary variable associated with the k-th element, with $x_k > 0$ for all $k \in U$ and t_x the total of x, i.e.

$$t_x = \sum_U x_k$$

Assume that the sampling frame is ordered regarding a given pattern. For example, ordered according the x-values, according to a different auxiliary variable, according to an identification variable or according to any other "natural"order.

Define the variable z_k as the cumulative sum of the first k values of x in the frame, i.e.

$$z_k = \sum_{i=1}^{k} x_i$$

Let u be a random number from a uniform distribution Unif(0,L), where

$$L = \frac{t_x}{n}$$

Select the n elements such that

$$z_{k-1} \leq u + [(i-1)L] < z_k \quad i = 1, 2, \cdots, n$$

The scheme described above generates a strict πps design, i.e. the inclusion probabilities, π_k, generated by the scheme satisfy the relation

$$\pi_k = n\frac{x_k}{t_x},$$

that is also without-replacement and fixed-size. The selection method is simple to implement in practice and the sample size is not a restriction. On the other hand, the second-order inclusion probabilities generated by it are not easily obtained and, often, many of them are equal to zero; so systematic sampling is not a measurable design and the unbiased estimators for the variance presented in (9) and (11) cannot be calculated. Many alternative estimators are available. Different estimators have been proposed, but all of them seem to perform well under different situations (see, for example (Wolter, 2007)).

Systematic πps is considered as an efficient scheme in the sense that its variance is usually small compared to other designs or even to other πps schemes. However, as there is no unbiased estimator for the variance, it is hard, in practice, to determine how well it is working.

Pareto sampling Let π_k^* be the desired inclusion probability associated with the k-th element. The π^*-values must satisfy

$$\sum_U \pi_k^* = n$$

where n is the desired sample size. Often, the π_k^*-values are defined as

$$\pi_k^* = n\frac{x_k}{t_x}$$

with x an auxiliary variable always greater than zero expected to be highly and positively correlated with y, and

$$t_x = \sum_U x_k$$

Associate, to each $k \in U$, a value u_k, such that u_k is a random number from a uniform distribution Unif(0,1). Define the variable z_k as

$$z_k = \frac{u_k \left(1 - \pi_k^*\right)}{\pi_k^* \left(1 - u_k\right)}$$

The sample consists of the n elements with the smallest z-values.

The scheme described above generates a without-replacement fixed size design, which is also easy to implement in practice and works for every sample size. Even so, Pareto sampling is not a strict πps, it is only approximately πps, i.e.

$$\pi_k \approx \pi_k^*$$

The exact first and second order inclusion probabilities generated by the scheme are not easy to obtain, so the π-estimator and the variance estimators (9) and (11) cannot be calculated. An alternative estimator for the total, that resembles the π-estimator is

$$\hat{t}_{\mathrm{q}\pi} = \sum_s \frac{y_k}{\pi_k^*} \tag{18}$$

An expression that approximates the variance of $\hat{t}_{\mathrm{q}\pi}$ is

$$AV\left(\hat{t}_{\mathrm{q}\pi}\right) = \frac{N}{N-1}\left[t_{y^2(1-\pi^*)/\pi^*} - \frac{t_{y(1-\pi^*)}^2}{t_{\pi^*(1-\pi^*)}}\right] \tag{19}$$

where

$$t_{y^2(1-\pi^*)/\pi^*} = \sum_U y_k^2(1-\pi_k^*)/\pi_k^* \qquad t_{y(1-\pi^*)} = \sum_U y_k(1-\pi_k^*)$$
$$t_{\pi^*(1-\pi^*)} = \sum_U \pi_k^*(1-\pi_k^*)$$

An estimator for the variance is

$$\hat{V}\left(\hat{t}_{\mathrm{q}\pi}\right) = \frac{n}{n-1}\left[\hat{t}_{y^2(1-\pi^*)/\pi^*} - \frac{\hat{t}_{y(1-\pi^*)}^2}{\hat{t}_{\pi^*(1-\pi^*)}}\right] \tag{20}$$

where

$$\hat{t}_{y^2(1-\pi^*)/\pi^*} = \sum_s \frac{y_k^2(1-\pi_k^*)/\pi_k^*}{\pi_k^*} \qquad \hat{t}_{y(1-\pi^*)} = \sum_s \frac{y_k(1-\pi_k^*)}{\pi_k^*}$$
$$\hat{t}_{\pi^*(1-\pi^*)} = \sum_s \frac{\pi_k^*(1-\pi_k^*)}{\pi_k^*}$$

Although is not strictly πps, Pareto sampling is, as systematic sampling, considered a very efficient design. For a more comprehensive presentation of Pareto sampling, see (Rosén, 1997).

2 A without-replacement fixed-size sampling design that uses auxiliary information

In this section a without-replacement fixed-size sampling design which uses auxiliary information is proposed. Also, the inclusion probabilities associated with this design are calculated. An algorithm that allows to selecting a sample from the design is described. In the final part of the section some examples are presented.

Result 1. Let U be the population that consists of N elements labeled $\{1, 2, \cdots, N\}$ and let, for every $k \in U$, q_k be a known value associated with the k-th element. Let, also, Ω_{q} be the set of the $\binom{N}{n}$ without-replacement samples of size n out of the N elements in U. The function that assigns the probability

$$p(s) = \frac{1}{\binom{N-1}{n-1}} \sum_{s} q_k \tag{21}$$

to any set s in Ω_{q}, is a sampling design if

$$t_q = \sum_{U} q_k = 1 \tag{22}$$

$$\sum_{i=1}^{n} q_{(i)} > 0 \tag{23}$$

where $q_{(i)}$ are the order statistics of the variable q.

A proof of the result is shown in the appendix. The support generated by the design in (21) coincides with Ω_{srs}, the support of a SRS.

As this design is based on the q-values, it will be called q-sampling. Note that the q-values do not need to be strictly positive, some of them can be negative. Condition (23) establishes a "limit"for the negative q-values allowed: the sum of the n smallest q-values must be greater than zero.

Now that, by Result 1, we know that q-sampling is, in fact, a sampling design, we can calculate the inclusion probabilities. In the following result, expressions for the first-order, second-order, and, in general, r-th order ($r \leq n$) inclusion probabilities are obtained.

Result 2 (Inclusion probabilities)**.** Under q-sampling as defined in (21), the first order inclusion probability of the k-th element is

$$\pi_k = \frac{1}{N-1} \left[(N-n)\, q_k + (n-1) \right] \tag{24}$$

The second order inclusion probability of the elements k and l ($k \neq l$) is

$$\pi_{kl} = \frac{n-1}{(N-1)(N-2)} \left[(N-n)\, (q_k + q_l) + (n-2) \right] \tag{25}$$

In general, the r-th order inclusion probability of the elements $k_1, k_2, \cdots, k_r$ is

$$\pi_{k_1 k_2 \cdots k_r} = \frac{\prod_{i=1}^{r-1} (n-i)}{\prod_{i=1}^{r} (N-i)} \left[(N-n) \sum_{i=1}^{r} q_{k_i} + (n-r) \right] \tag{26}$$

The result is proved in the appendix. The first and second order inclusion probabilities will be used in the next section, when q-sampling is coupled with the π-estimator. Note that the π_k defined in (24) are valid inclusion probabilities, in the sense that $\sum_U \pi_k = n$, where n is the desired sample size:

$$\sum_U \pi_k = \sum_U \frac{1}{N-1}\left[(N-n)\,q_k + (n-1)\right] = \qquad \text{(by Result (2))}$$

$$\frac{1}{N-1}\left[(N-n)\sum_U q_k + N(n-1)\right] = \qquad \text{(after some algebra)}$$

$$\frac{1}{N-1}\left[(N-n) + N(n-1)\right] = \qquad \text{(by condition (22))}$$

$$\frac{1}{N-1}\left[Nn - n\right] = n$$

The selection algorithm that will be proposed later in this section requires the calculation of conditional inclusion probabilities.

Result 3 (Conditional inclusion probabilities)**.** The conditional inclusion probability of the element k, given that the elements $l_1, l_2, \cdots, l_r$ are also in the sample s, is

$$\pi_{k|l_1 l_2 \cdots l_r} = \frac{\pi_{l_1 l_2 \cdots l_r k}}{\pi_{l_1 l_2 \cdots l_r}} = \frac{n-r}{N-r-1}\frac{\left[(N-n)\left(\sum_{i=1}^r q_{l_i} + q_k\right) + (n-r-1)\right]}{\left[(N-n)\sum_{i=1}^r q_{l_i} + (n-r)\right]}$$

The result is proved in the appendix. A draw-by-draw sample selection scheme for q-sampling is described as follows.

Sample selection scheme

1. For a given set of valid q-values (validity is to be understood under conditions (22) and (23)), calculate the associated π-values; and select one element with probabilities proportional to the π-values, say element l_1 is selected.

2. Calculate the conditional inclusion probabilities for the remaining elements given l_1, $\pi_{k|l_1}$; and select one element with probabilities proportional to these conditional probabilities, say element l_2 is selected.

3. Calculate the conditional inclusion probabilities for the remaining elements given l_1, l_2, $\pi_{k|l_1 l_2}$; and select one element with probabilities proportional to these conditional probabilities, say element l_3 is selected.

4. Continue this process until n elements are selected.

The following example illustrates the procedure.

Example 1 (Ilustrating the sample selection scheme)**.** Let U be a population of size $N = 10$ with q-values equal to

$$q = [-0.0500 \quad -0.0125 \quad 0.0250 \quad 0.0625 \quad 0.1000$$
$$0.1000 \quad 0.1375 \quad 0.1750 \quad 0.2125 \quad 0.2500]$$

The goal is to draw a sample of size $n = 4$. Note that (22) and (23) are satisfied.

1. The inclusion probabilities are calculated using (24), for example, for the first element we have:

$$\pi_1 = \frac{1}{N-1}\left[(N-n)q_1 + (n-1)\right] = \frac{1}{10-1}\left[(10-4)(-0.0500) + (4-1)\right] = 0.3$$

The set of inclusion probabilities is

$$\pi = \begin{bmatrix} 0.300 & 0.325 & 0.350 & 0.375 & 0.400 & 0.400 & 0.425 & 0.450 & 0.475 & 0.500 \end{bmatrix}$$

The cumulative sum of the π-values is

$$z = \begin{bmatrix} 0.300 & 0.625 & 0.975 & 1.350 & 1.750 & 2.150 & 2.575 & 3.025 & 3.500 & 4.000 \end{bmatrix}$$

One value from a uniform distribution Unif(0,4) is realized: $u_1 = 3.799$, so the element $k = 10$ is selected.

2. The conditional inclusion probabilities are calculated using (26). For example, for the first element we have:

$$\pi_{1|10} = \frac{n-r}{N-r-1}\frac{\left[(N-n)\left(q_{10}+q_1\right) + (n-r-1)\right]}{\left[(N-n)q_{10} + (n-r)\right]} = \frac{4-1}{10-1-1}\frac{\left[(10-4)\left(0.2500 + (-0.0500)\right) + (4-1-1)\right]}{\left[(10-4)0.2500 + (4-1)\right]} = 0.2667$$

The nine conditional inclusion probabilities are

$$\pi_{k|10} = [0.2667 \quad 0.2854 \quad 0.3042 \quad 0.3229 \quad 0.3417 \quad 0.3417 \quad 0.3604 \quad 0.3792 \quad 0.3979 \quad \text{NA}]$$

The cumulative sum of the conditional probabilities is

$$z_{k|10} = [0.2667 \quad 0.5521 \quad 0.8563 \quad 1.1792 \quad 1.5208 \quad 1.8625 \quad 2.2229 \quad 2.6021 \quad 3.0000 \quad \text{NA}]$$

One value from a uniform distribution Unif(0,3) is realized: $u_2 = 0.378$, so the element $k = 2$ is selected.

3. The new set of conditional inclusion probabilities is calculated. For example, for the first element we have:

$$\pi_{1|10,2} = \frac{n-r}{N-r-1}\frac{\left[(N-n)\left(q_{10}+q_2+q_1\right) + (n-r-1)\right]}{\left[(N-n)\left(q_{10}+q_2\right) + (n-r)\right]} = \frac{4-2}{10-2-1}\frac{\left[(10-4)\left(0.2500 + (-0.0125) + (-0.0500)\right) + (4-2-1)\right]}{\left[(10-4)\left(0.2500 + (-0.0125)\right) + (4-2)\right]} = 0.1773$$

The eight conditional inclusion probabilities are

$$\pi_{k|10,2} = [0.1773 \quad \text{NA} \quad 0.2148 \quad 0.2336 \quad 0.2523 \quad 0.2523 \quad 0.2711 \quad 0.2889 \quad 0.3087 \quad \text{NA}]$$

The cumulative sum of the conditional probabilities is

$$z_{k|10,2} = [0.1773 \quad \text{NA} \quad 0.3921 \quad 0.6257 \quad 0.8780 \\ 1.1303 \quad 1.4015 \quad 1.6913 \quad 2.0000 \quad \text{NA}]$$

One value from a uniform distribution Unif(0,2) is realized: $u_3 = 0.484$, so the element $k = 4$ is selected.

4. The new set of conditional inclusion probabilities is calculated. For example, for the first element we have:

$$\pi_{1|10,2,4} = \frac{n-r}{N-r-1}\frac{[(N-n)(q_{10}+q_2+q_4+q_1)+(n-r-1)]}{[(N-n)(q_{10}+q_2+q_4)+(n-r)]} = \\ \frac{4-3}{10-3-1}\frac{[(10-4)(0.2500+(-0.0125)+0.0625+(-0.0500))+(4-3-1)]}{[(10-4)(0.2500+(-0.0125)+0.0625)+(4-3)]} = \\ 0.0893$$

The seven conditional inclusion probabilities are

$$\pi_{k|10,2,4} = [0.0893 \quad \text{NA} \quad 0.1161 \quad \text{NA} \quad 0.1429 \\ 0.1429 \quad 0.1563 \quad 0.1696 \quad 0.1830 \quad \text{NA}]$$

The cumulative sum of the conditional probabilities is

$$z_{k|10,2,4} = [0.0893 \quad \text{NA} \quad 0.2054 \quad \text{NA} \quad 0.3482 \\ 0.4911 \quad 0.6473 \quad 0.8170 \quad 1.0000 \quad \text{NA}]$$

One value from a uniform distribution Unif(0,1) is realized: $u_4 = 0.044$, so the first element is selected. And the selected sample is $s = \{1, 2, 4, 10\}$

The process is summarized in Table 1, where *St* stands for *Step*.

		St 1, $u = 3.799$		St 2, $u = 0.378$		St 3, $u = 0.484$		St 4, $u = 0.044$	
k	q	π	z	$\pi_{k\|10}$	$z_{k\|10}$	$\pi_{k\|10,2}$	$z_{k\|10,2}$	$\pi_{k\|10,2,4}$	$z_{k\|10,2,4}$
1	-0.0500	0.300	0.300	0.2667	0.2667	0.1773	0.1773	0.0893	**0.0893**
2	-0.0125	0.325	0.625	0.2854	**0.5521**	NA	NA	NA	NA
3	0.0250	0.350	0.975	0.3042	0.8563	0.2148	0.3921	0.1161	0.2054
4	0.0625	0.375	1.350	0.3229	1.1792	0.2336	**0.6257**	NA	NA
5	0.1000	0.400	1.750	0.3417	1.5208	0.2523	0.8780	0.1429	0.3482
6	0.1000	0.400	2.150	0.3417	1.8625	0.2523	1.1303	0.1429	0.4911
7	0.1375	0.425	2.575	0.3604	2.2229	0.2711	1.4015	0.1563	0.6473
8	0.1750	0.450	3.025	0.3792	2.6021	0.2899	1.6913	0.1696	0.8170
9	0.2125	0.475	3.500	0.3979	3.0000	0.3087	2.0000	0.1830	1.0000
10	0.2500	0.500	**4.000**	NA	NA	NA	NA	NA	NA

Table 1: Ilustrating the sample selection scheme

Note that, in Example 1, the sum of the conditional inclusion probabilities always yield an integer. This is not a coincidence, in fact,

$$\sum_{U^{(r)}} \pi_{k|l_1 l_2 \cdots l_r} = n - r$$

where $U^{(r)}$ is the set of all elements in U except the elements $l_1, l_2, \cdots, l_r$.

Admittedly, the selection scheme proposed above is more complicated than the schemes described for systematic πps and Pareto sampling. However, the scheme is not hard to implement in practice. A program in R is given in the appendix.

Depending on different settings for the q-values, q-sampling takes some interesting forms, as will be shown in the following examples.

Example 2 (Simple Random Sampling)**.** Let $q_k = \frac{1}{N}$ $(k = 1, 2, \cdots, N)$. Under this setting of the q-values, q-sampling becomes a SRS.

In this case, (21) takes the form

$$p(s) = \frac{1}{\binom{N-1}{n-1}} \sum_s q_k = \frac{1}{\binom{N-1}{n-1}} \sum_s \frac{1}{N} = \frac{1}{\binom{N-1}{n-1}} \frac{n}{N} = \frac{1}{\binom{N}{n}}$$

Conditions (22) and (23) are satisfied:

$$t_q = \sum_U q_k = \sum_U \frac{1}{N} = N\frac{1}{N} = 1$$

$$\sum_{i=1}^{n} q_{(i)} = \sum_{i=1}^{n} \frac{1}{N} = n\frac{1}{N} > 0$$

The first order inclusion probabilities are

$$\pi_k = \frac{1}{N-1}\left[(N-n)\, q_k + (n-1)\right] = \frac{1}{N-1}\left[(N-n)\,\frac{1}{N} + (n-1)\right] = \frac{n}{N}$$

The second order inclusion probabilities are

$$\pi_{kl} = \frac{n-1}{(N-1)(N-2)}\left[(N-n)\,(q_k + q_l) + (n-2)\right] =$$

$$\frac{n-1}{(N-1)(N-2)}\left[(N-n)\left(\frac{1}{N} + \frac{1}{N}\right) + (n-2)\right] = \frac{n(n-1)}{N(N-1)}$$

Example 3 (πps)**.** As discussed in Section 1.3.4, suppose that we want to carry out a πps sampling, i.e. we want to assign some (known and defined beforehand) inclusion probabilities $\pi_1^*, \pi_2^*, \cdots, \pi_N^*$ to the N elements in the population U. As mentioned before, usually this design is intended to be coupled with the π-estimator. This strategy will be discussed in Section 3.

Now, for the sampling design in (21), we know that if we fix the q-values, we obtain the inclusion probabilities in (24). But we can work in the opposite direction: we can fix the π-values and then we obtain the associated q-values. So, solving for π_k in equation (24), we have

$$q_k = \frac{1}{N-n}\left[(N-1)\pi_k^* - (n-1)\right] \tag{27}$$

With the q-values defined in this way, the design (21) takes the form

$$p(s) = \frac{1}{\binom{N-1}{n-1}} \sum_s q_k = \frac{1}{\binom{N-1}{n-1}} \sum_s \frac{1}{N-n} \left[(N-1)\pi_k^* - (n-1)\right] = \frac{1}{\binom{N-2}{n-1}} \left[\sum_s \pi_k^* - \frac{n(n-1)}{N-1}\right]$$

Condition (22) is satisfied:

$$t_q = \sum_U q_k = \sum_U \frac{1}{N-n} \left[(N-1)\pi_k^* - (n-1)\right] = \frac{1}{N-n} \left[(N-1)\sum_U \pi_k^* - N(n-1)\right] = \frac{1}{N-n} \left[(N-1)n - N(n-1)\right] = \frac{1}{N-n}(N-n) = 1$$

where we use the fact that $\sum_U \pi_k^* = n$. With respect to condition (23), we have

$$\sum_{i=1}^{n} q_{(i)} = \sum_{i=1}^{n} \frac{1}{N-n} \left[(N-1)\pi_{(i)}^* - (n-1)\right] = \frac{1}{N-n} \left[(N-1)\sum_{i=1}^{n} \pi_{(i)}^* - n(n-1)\right]$$

which is greater than zero only when

$$\sum_{i=1}^{n} \pi_{(i)}^* > \frac{n(n-1)}{N-1} \tag{28}$$

In other words, q-sampling can be used to generate a πps design only when (28) is satisfied. Unfortunately, (28) is a very strong requirement. Note that the right side in (28) is approximately equal (when N and n are large enough) to $\frac{n^2}{N}$; now, under SRS, the sum of the inclusion probabilities of any n elements is also $n \times \frac{n}{N} = \frac{n^2}{N}$ So, for large samples and populations, the πps sampling generated by q-sampling is limited to nearly constant inclusion probabilities; therefore, q-sampling becomes almost "useless"as a scheme for a strict πps sampling. As an example, suppose that we want to select a πps sample of size $n = 50$ from a population of size $N = 100$ using q-sampling, then, by (28), the 50 smallest conditional probabilities must satisfy

$$\sum_{i=1}^{n} \pi_{(i)}^* > \frac{50(50-1)}{100-1} = \frac{2450}{99}$$

therefore, the remaining 50 largest conditional probabilities must satisfy

$$\sum_{i=n+1}^{N} \pi_{(i)}^* < 50 - \frac{50(50-1)}{100-1} = 50 - \frac{2450}{99} = \frac{2500}{99}.$$

A condition that is only satisfied if the π-values are nearly constant.

In order to avoid this strong drawback an alternative estimator that is also unbiased under q-sampling is proposed in Section 4. The strategy that combines q-sampling and the proposed estimator is not affected by the restriction in (28).

Example 4 ($q$ps)**.** Let x be an auxiliary variable known beforehand for every element in U and $x_k > 0$ $(k = 1, 2, \cdots, N)$. Define the q-values as

$$q_k = \frac{x_k}{t_x} \qquad \text{where } t_x = \sum_U x_k$$

It is known that when the auxiliary variable x is well correlated with the study variable y, and the design is coupled with an appropriate estimator, there is a great gain in using a proportional-to-size design. A strategy that combines this design with an "appropriate"estimator will be discussed in the fourth section.

Under this setting, q-sampling takes the form

$$p(s) = \frac{1}{\binom{N-1}{n-1}} \sum_s q_k = \frac{1}{\binom{N-1}{n-1}} \sum_s \frac{x_k}{t_x} = \frac{1}{\binom{N-1}{n-1} t_x} \sum_s x_k$$

Conditions (22) and (23) are satisfied:

$$t_q = \sum_U q_k = \sum_U \frac{x_k}{t_x} = \frac{1}{t_x} \sum_U x_k = \frac{1}{t_x} t_x = 1$$
$$\sum_{i=1}^{n} q_{(i)} = \sum_{i=1}^{n} \frac{x_{(i)}}{t_x} = \frac{1}{t_x} \sum_{i=1}^{n} x_{(i)} > 0$$

The first order inclusion probabilities are

$$\pi_k = \frac{1}{N-1} \left[(N-n)\, q_k + (n-1) \right] = \frac{1}{N-1} \left[(N-n)\, \frac{x_k}{t_x} + (n-1) \right]$$

The second order inclusion probabilities are

$$\pi_{kl} = \frac{n-1}{(N-1)(N-2)} \left[(N-n)\,(q_k + q_l) + (n-2) \right] =$$
$$\frac{n-1}{(N-1)(N-2)} \left[(N-n) \left(\frac{x_k}{t_x} + \frac{x_l}{t_x} \right) + (n-2) \right] =$$
$$\frac{n-1}{(N-1)(N-2)} \left[\frac{(N-n)}{t_x} (x_k + x_l) + (n-2) \right]$$

Note that the expression "proportional-to-size"is to be understood in a different way according to the design: in πps sampling the proportionality is with respect to the inclusion probabilities; in pps sampling, it is with respect to the selection probabilities in each draw; in $q$ps, it is with respect to the q-values. And, in the same way, different estimators suits better to each case: π-estimator for a πps, pwr-estimator for a pps. An estimator that suits well to $q$ps will be proposed in Section 4.

3 q-sampling and the π-estimator

In this section we discuss briefly the strategy that couples q-sampling with the π-estimator. Expressions for the estimator, its variance and an unbiased estimator for the variance are presented.

As described in Section 2, for a given sample and a given set of valid q-values, the associated inclusion probabilities can be easily calculated by (24), so the π-estimator, presented in (5), can be implemented

$$\hat{t}_\pi = \sum_s \frac{y_k}{\pi_k} \quad \text{where } \pi_k = \frac{1}{N-1}\left[(N-n)q_k + (n-1)\right] \tag{29}$$

The second order inclusion probabilities can, also, be easily calculated under q-sampling, and as it is a fixed-size design, either expression for the variance, (8) or (10), can be used

$$V\left(\hat{t}_\pi\right) = \sum_U \sum_U \Delta_{kl} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l} = -\frac{1}{2}\sum_U \sum_U \Delta_{kl}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \tag{30}$$

where $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$, $\pi_{kl} = \frac{n-1}{(N-1)(N-2)}\left[(N-n)\left(q_k+q_l\right) + (n-2)\right]$ and π_k as defined in (29).

Under q-sampling, all the π_{kl} are greater than zero, so the variance estimators (9) and (11) are unbiased

$$\hat{V}\left(\hat{t}_\pi\right) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} \tag{31}$$

$$\hat{V}\left(\hat{t}_\pi\right) = -\frac{1}{2}\sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \tag{32}$$

As mentioned before, a sufficient condition for (32) to take nonnegative values is that $\Delta_{kl} < 0$ for all $k, l \in U$. Under q-sampling, it can be seen that Δ_{kl} can be expressed as

$$\Delta_{kl} = \pi_{kl} - \pi_k\pi_l = \frac{N-n}{(N-2)(N-1)^2}\left[(1-n)(1-q_k-q_l) + (2-N)(N-n)q_kq_l\right]$$

which is smaller than zero, for example, when all the q-values are greater than zero.

(30) is an exact expression for the variance of the π-estimator, however, the double sum leads to cumbersome calculations. For this reason, an alternative expression for the variance is presented. This expression is easy to calculate, in the sense that involves only single sums:

$$AV\left(\hat{t}_\pi\right) = \frac{N}{N-1}\left[t_{y^2(1-\pi)/\pi} - \frac{t^2_{y(1-\pi)}}{t_{\pi(1-\pi)}}\right] \tag{33}$$

where

$$t_{y^2(1-\pi)/\pi} = \sum_U y_k^2(1-\pi_k)/\pi_k \qquad t_{y(1-\pi)} = \sum_U y_k(1-\pi_k) \qquad t_{\pi(1-\pi)} = \sum_U \pi_k(1-\pi_k)$$

The reasoning to propose (33) as an approximate variance for the strategy (q-sampling–π-estimator) is as follows. In Pareto sampling, the actual inclusion probabilities, π_k, are nearly equal to the desired inclusion probabilities, π_k^*, so the estimator in (18) is nearly equal to the π-estimator, i.e.

$$\hat{t}_{q\pi} = \sum_s \frac{y_k}{\pi_k^*} \approx \sum_s \frac{y_k}{\pi_k} = \hat{t}_\pi$$

And its variances should also be similar:

$$V\left(\hat{t}_{q\pi}\right) \approx V\left(\hat{t}_{\pi}\right)$$

But, (19) is an approximation to $V\left(\hat{t}_{q\pi}\right)$, so it is also an approximation to $V\left(\hat{t}_{\pi}\right)$. Under q-sampling the exact π-values can be used, and so, formula (33) is obtained.

In the same sense that (33) mimics (19), a simpler (but biased) estimator for the variance is obtained by mimicking (20):

$$\hat{V}\left(\hat{t}_{\mathrm{q}\pi}\right) = \frac{n}{n-1}\left[\hat{t}_{y^2(1-\pi)/\pi} - \frac{\hat{t}^2_{y(1-\pi)}}{\hat{t}_{\pi(1-\pi)}}\right] \tag{34}$$

where

$$\hat{t}_{y^2(1-\pi)/\pi} = \sum_s \frac{y_k^2(1-\pi_k)/\pi_k}{\pi_k} \qquad \hat{t}_{y(1-\pi)} = \sum_s \frac{y_k(1-\pi_k)}{\pi_k}$$

$$\hat{t}_{\pi(1-\pi)} = \sum_s \frac{\pi_k(1-\pi_k)}{\pi_k}$$

According to the results shown in Section 5, it seems that these simpler counterparts of the variance and its estimator can be considered as good approximations.

Note that a perfect association between the y and the q-values will not yield a zero variance in this strategy.

4 An alternative unbiased estimator under q-sampling

In this section an estimator for the total is proposed. It will be shown that this estimator is unbiased under q-sampling. An expression that approximates the variance of the estimator is developed. Also, an estimator for the variance will be presented. Finally, the examples in Section 2 will be used again to illustrate its use with the proposed estimator.

Result 4. Under q-sampling, the estimator

$$\hat{t}_q = \frac{\sum_s y_k}{\sum_s q_k}, \tag{35}$$

which will be called q-estimator, is unbiased for the total

$$t_y = \sum_U y_k$$

A proof of the result is shown in the appendix. Note that if the q-values are perfectly proportional to the unknown y-values, then the variance of the estimator will be zero.

As $\hat{t}_q$ is a ratio, its variance is not easy to calculate. Nevertheless, an expression that approximates the variance is developed using the Taylor's linearization method as shown in, for example, Casella & Berger (2002). According to the method, the

variance of a ratio of random variables, can be approximated in terms of the expected values, variances and covariance:

$$V\left[\frac{Y}{Q}\right] \approx \frac{E^2(Y)}{E^2(Q)}\left[\frac{V(Y)}{E^2(Y)} - 2\frac{Cov(Y,Q)}{E(Y)E(Q)} + \frac{V(Q)}{E^2(Q)}\right] \tag{36}$$

where $E(Y)$, $V(Y)$, $E(Q)$, $V(Q)$ and $Cov(Y,Q)$ are, respectively, the expected value of Y, the variance of Y, the expected value of Q, the variance of Q and the covariance between Y and Q.

Letting $Y = \sum_s y_k$ and $Q = \sum_s q_k$ in (36), the following result is obtained. (A proof of the result is presented in the Appendix).

Result 5. Under q-sampling, the variance of $\hat{t}_q$ can be approximated by

$$AV\left(\hat{t}_q\right) = \frac{(N-1)}{(N-2)}\frac{1}{B^4}\left[CB^2 - 2EAB + DA^2\right] \tag{37}$$

where

$$\begin{aligned}
A &= t_{qy}(N-n) + t_y(n-1)\\
B &= t_{q^2}(N-n) + (n-1)\\
C &= t_{qy^2}(N-n)(N-2n) + \left(t_{y^2} + 2t_{qy}t_y\right)(N-n)(n-1) + t_y^2(n-1)(n-2)\\
D &= t_{q^3}(N-n)(N-2n) + 3t_{q^2}(N-n)(n-1) + (n-1)(n-2)\\
E &= t_{yq^2}(N-n)(N-2n) + \left(t_y t_{q^2} + 2t_{qy}\right)(N-n)(n-1) + t_y(n-1)(n-2)
\end{aligned}$$

and

$$t_y = \sum_U y_k \qquad t_{y^2} = \sum_U y_k^2 \qquad t_{q^2} = \sum_U q_k^2 \qquad t_{q^3} = \sum_U q_k^3$$
$$t_{qy} = \sum_U q_k y_k \qquad t_{yq^2} = \sum_U y_k q_k^2 \qquad t_{qy^2} = \sum_U q_k y_k^2$$

It is true that, at first glance, (37) seems a cumbersome expression. Even so, it is simpler than the variance for the π-estimator, (30), in the sense that it does not require the matrix of second-order inclusion probabilities; (37) is based on simple operations over totals.

An estimator of the variance is obtained by replacing the totals in (37) by its estimates:

Result 6. An estimator of the variance for the strategy (q-sampling–q-estimator) is

$$\hat{V}\left(\hat{t}_q\right) = \frac{(\hat{N}-1)}{(\hat{N}-2)}\frac{1}{\hat{B}^4}\left[\hat{C}\hat{B}^2 - 2\hat{E}\hat{A}\hat{B} + \hat{D}\hat{A}^2\right] \tag{38}$$

where

$$\begin{aligned}
\hat{A} &= \hat{t}_{qy}(\hat{N}-n) + \hat{t}_y(n-1)\\
\hat{B} &= \hat{t}_{q^2}(\hat{N}-n) + (n-1)\\
\hat{C} &= \hat{t}_{qy^2}(\hat{N}-n)(\hat{N}-2n) + \left(\hat{t}_{y^2} + 2\hat{t}_{qy}\hat{t}_y\right)(\hat{N}-n)(n-1) + \hat{t}_y^2(n-1)(n-2)\\
\hat{D} &= \hat{t}_{q^3}(\hat{N}-n)(\hat{N}-2n) + 3\hat{t}_{q^2}(\hat{N}-n)(n-1) + (n-1)(n-2)\\
\hat{E} &= \hat{t}_{yq^2}(\hat{N}-n)(\hat{N}-2n) + \left(\hat{t}_y\hat{t}_{q^2} + 2\hat{t}_{qy}\right)(\hat{N}-n)(n-1) + \hat{t}_y(n-1)(n-2)
\end{aligned}$$

and

$$\hat{N} = \frac{n}{\sum_s q_k} \qquad \hat{t}_y = \frac{\sum_s y_k}{\sum_s q_k} \qquad \hat{t}_{y^2} = \frac{\sum_s y_k^2}{\sum_s q_k} \qquad \hat{t}_{q^2} = \frac{\sum_s q_k^2}{\sum_s q_k}$$
$$\hat{t}_{q^3} = \frac{\sum_s q_k^3}{\sum_s q_k} \qquad \hat{t}_{qy} = \frac{\sum_s q_k y_k}{\sum_s q_k} \qquad \hat{t}_{yq^2} = \frac{\sum_s y_k q_k^2}{\sum_s q_k} \qquad \hat{t}_{qy^2} = \frac{\sum_s q_k y_k^2}{\sum_s q_k}$$

It is important to remark that (37) is not the actual variance of the strategy, it is only an approximation. Also, (38) is not an unbiased estimator of (37), it is only a "consistent estimator". In the next section they are studied in order to see how well they behave with respect to their counterparts. But, before this comparison, the q-estimator will be developed for the three different settings of the q-values presented in examples 2, 3 and 4.

Example 5 (Continuation of Example 2)**.** Let $q_k = \frac{1}{N}$. The q-estimator becomes

$$\hat{t}_q \equiv \frac{\sum_s y_k}{\sum_s q_k} = \frac{\sum_s y_k}{\sum_s \frac{1}{N}} = \frac{N}{n} \sum_s y_k$$

So, when the q-values are all equal to $\frac{1}{N}$, the strategies (q-sampling–q-estimator) and (SRS–π-estimator) are equivalent.

With respect to the variance, using (37) and noting that

$$t_{qy} = \frac{t_y}{N}, \quad t_{q^2} = \frac{1}{N}, \quad t_{qy^2} = \frac{t_{y^2}}{N}, \quad t_{q^3} = \frac{1}{N^2} \quad \text{and} \quad t_{yq^2} = \frac{t_y}{N^2}$$

we obtain

$$AV\left(\hat{t}_y\right) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S_{y,U}^2$$

which coincides exactly with the "true"variance. So, under the setting $q_k = \frac{1}{N}$, the approximation (37) is no longer an approximation, it is the actual value of the variance. On the other hand, with respect to the variance estimator, using (38) and noting that

$$\hat{t}_{qy} = \bar{y}, \quad \hat{t}_y = N\bar{y}, \quad \hat{t}_q^2 = \frac{1}{N}, \quad \hat{t}_{qy^2} = \bar{y}^2, \quad \hat{t}_{y^2} = N\bar{y^2}, \quad \hat{t}_{q^3} = \frac{1}{N^2}, \quad \hat{t}_{yq^2} = \frac{1}{N}\bar{y}$$

we obtain

$$\hat{V}\left(\hat{t}_q\right) = \frac{(n-1)}{n} \frac{N}{(N-1)} \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S_{y,s}^2 \tag{39}$$

Comparing (39) with the unbiased estimator (17), we can see that even when (39) is biased, it is assymptotically unbiased, i.e. is unbiased for large samples and large populations.

Example 6 (Continuation of Example 3)**.** Let q_k defined as in Example 3, i.e.

$$q_k = \frac{1}{N-n}\left[(N-1)\pi_k^* - (n-1)\right]$$

Assuming that the q-values are valid under condition (28), q-sampling with this setting generates a strict πps design. Under this setting, the q-estimator takes the form

$$\hat{t}_y \equiv \frac{\sum_s y_k}{\sum_s q_k} = \frac{\sum_s y_k}{\sum_s \frac{1}{N-n}\left[(N-1)\pi_k^* - (n-1)\right]} = \frac{(N-n)\sum_s y_k}{(N-1)\sum_s \pi_k^* - n(n-1)}$$

Example 7 (Continuation of Example 4)**.** Let q_k defined as in Example 4, i.e.

$$q_k = \frac{x_k}{t_x} \qquad \text{where } t_x = \sum_U x_k \text{ and } x > 0 \text{ for } k = 1, 2, \cdots, N$$

Under this setting, the q-estimator becomes

$$\hat{t}_y \equiv \frac{\sum_s y_k}{\sum_s q_k} = \frac{\sum_s y_k}{\sum_s \frac{x_k}{t_x}} = t_x \frac{\sum_s y_k}{\sum_s x_k} \tag{40}$$

Note that if we could have $x_k = \alpha y_k$, then (40) would be

$$t_x \frac{\sum_s y_k}{\sum_s x_k} = \alpha t_y \frac{\sum_s y_k}{\alpha \sum_s y_k} = t_y$$

So, the variance of the strategy ($q$ps–q-estimator) would be zero. And in this case, both, the approximation to the variance (37) and its estimator (38) are also equal to zero.

5 Assessing $q$ps-sampling

In sections 2, 3 and 4, $q$ps-sampling was defined and its use with two different estimators (the known π-estimator, and the proposed q-estimator) was discussed. In this section a numerical study designed with the goal to assess the behavior of $q$ps-sampling is presented. In the first part of the section, the simulation study and its objectives are described. In the second part, results regarding how well the approximated expressions for the variance work with respect to the variance are presented. In the third part, results regarding how well the variance estimators work with respect to the approximated variances are presented. The fourth part compares the bias of five different strategies. The fifth part presents a comparison of the efficiency of the five strategies in terms of its variance. A comparison of the coverage of the five strategies is presented in the last subsection.

5.1 The simulation study and its objectives

q-sampling was defined in Section 2. It is a design that uses auxiliary information and whose inclusion probabilities are easily obtained. q-sampling may be seen as a family of designs, depending on how the q-values are defined. In particular three different settings were presented. The first setting (where all the q-values are equal) is equivalent to simple random sampling, so this case becomes "uninteresting", in the sense that no use of auxiliary information is made. The second setting is πps sampling, and it was shown that the restriction in (28) should be satisfied by q-sampling in order to obtain a strict πps; unfortunately, this restriction is so strong that is hardly satisfied in practice, so, this case becomes also uninteresting. The third setting was called $q$ps. In this case, the q-values are defined proportionally to a known and positive auxiliary variable. This is, somehow, a natural way to define the q-values and no restrictions are found in this case. For this reason, $q$ps is the only type of q-sampling that will be evaluated through the simulation study in this section.

In Section 3, the use of the π-estimator together with $q$ps sampling was described. For this case, an exact expression for the variance is known (see equation (30)).

However, this expression is hard to compute due to the double sum involved in it, so a simpler but only approximate expression was presented in (33). In Section 4, the use of $q$ps, together with the proposed q-estimator was descibed. No exact expression for the variance is available in this case, however an approximate expression is presented in (37). To study how well these expressions approximate the variance is the first objective of the simulation study, and will be developed in subsection 5.2.

Two unbiased estimators of the variance for the π-estimator were presented in Section 3. Again, these estimators involve a double sum that makes them difficult to compute in practice, for this reason a simpler, but biased, estimator was also presented in (34). Regarding the q-estimator, no unbiased estimator for the variance is available, but a "consistent"estimator was presented in Section 4. To study how biased are these estimators with respect to the approximated variances is the second objective of the simulation study and will be developed in subsection 5.3.

How good a sampling strategy is, is usually assessed in terms of its bias and its variance. The two strategies ($q$ps–π) (Section 3) and ($q$ps–q) (Section 4) are compared to three well known strategies: (SRS–π) (Section 1.3.1), (Pareto–qπ) and (Systematic–π) (Section 1.3.4). The bias of the five strategies under comparison is discussed in Section 5.4. To compare the variance of the five strategies is the third objective of the simulation study and the results are shown in Section 5.5.

Usually, using the central limit theorem, the distribution of the estimators is approximated by a normal distribution, or being more conservative, by a t-distribution. Therefore, confidence intervals are estimated using this approximation. In this sense, it is expected that confidence intervals built in this way will cover the true parameter with a probability of approximately $100(1-\alpha)\%$ for a given confidence level, α. To study the coverage of the five strategies under comparison is the fourth objective of the simulation study. The results are shown in Section 5.6.

A Monte Carlo simulation study was developed in order to investigate the objectives described above. Four populations of sizes $N = 20, 200, 2000, 20000$ were generated as follows.

The values for the auxiliary variable x were generated as N realizations from a χ^2_1 distribution. The intention with this distribution is to recreate the usually observed case in practice of right-skewed variables. Figure 1 shows the histogram for the case $N = 2000$ with the actual density overplotted.

Once the values of the auxiliary variable were observed, thirteen study variables were generated by assigning different values to the βs in the equation

$$y_k^{(i)} = \beta_0 + \beta_1 x_k^{\beta_2} + \epsilon_k \quad (i = 1, 2, \cdots, 13) \qquad \text{with } \epsilon_k \sim N\left(0, \beta_3 x_k^{\beta_4}\right)$$

The different settings for the β-values are summarized in Table 2. The negative values obtained using the equation above were replaced by its absolute value. Furthermore, having into account that Pearson's correlation has no clear interpretation for non-linear associations between x and y, the high, medium and low correlations in Table 2 refer to the rank-based Spearman's correlation coefficient which measures monotonic relations between x and y. Often, in a situation like the one presented in case 1 in Table 2 (linear association without intercept with high correlation), the variance of a design which uses auxiliary information is small compared to the variance of Simple Random Sampling. As the correlation between x and y gets smaller or the type of association is far from linear without intercept, this efficiency tends to

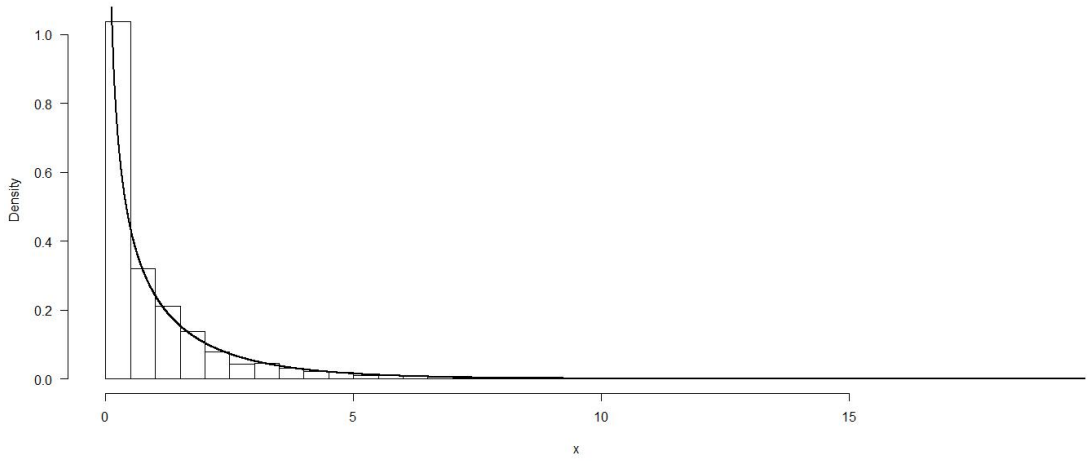

Figure 1: Histogram of x for the case $N = 2000$

disappear, at the point that SRS can be even more efficient than a proportional-to-size design. Cases 2 to 13 allow to study the behavior of the strategies under non-ideal situations.

Case (i)	β_0	β_1	β_2	β_3	β_4	Type	Corr.	ρ
1	0	2.39	1.0	1.00	1.00	Linear without intercept	High	0.95
2	0	0.15	1.0	1.00	0.45	Linear without intercept	Medium	0.65
3	0	0.11	1.0	1.00	0.17	Linear without intercept	Low	0.35
4	10	2.54	1.0	1.00	1.00	Linear with intercept	High	0.95
5	10	0.93	1.0	1.00	0.45	Linear with intercept	Medium	0.65
6	10	0.36	1.0	1.00	0.17	Linear with intercept	Low	0.35
7	0	4.88	0.5	1.00	1.00	Concave	High	0.95
8	0	1.00	0.5	1.23	0.40	Concave	Medium	0.65
9	0	1.00	0.5	2.61	0.15	Concave	Low	0.35
10	0	1.66	2.0	1.00	1.00	Convex	High	0.95
11	0	0.30	2.0	2.64	0.40	Convex	Medium	0.65
12	0	0.33	2.0	3.00	0.10	Convex	Low	0.35
13	0	0.00	1.0	1.00	0.00	Independent	Zero	0.00

Table 2: β-values for the generation of the thirteen study variables y

Figure 2 shows scatter plots for nine selected y-variables for the case $N = 2000$. The effect of the β-values can be more easily interpreted using the figure: β_0 is the intercept, β_1 is a scale factor, β_2 defines the functional form of the association, β_3 is a scale factor for the variance and β_4 is the functional form of the variance. The idea behind β_3 and β_4 is to create heteroscedastic relations between the variables.

Once the populations, U, have been completely defined, different sample sizes are considered:

- For the case $N = 20$, the sample sizes defined were $n = 1, 2, \cdots, 19$;

- For the case $N = 200$, they were $n = 1, 3, 6, 10, 15, 21, 28, 36, 45, 55, 66, 78, 91, 105, 120, 136, 153, 171, 191, 199$;
- For the case $N = 2000$, they were $n = 1, 4, 9, 16, 25, 36, 49, 64, 81, 100, 121, 144, 169, 196, 225, 256, 289, 324, 361, 400$;
- For the case $N = 20000$, they were $n = 1, 50, 125, 250, 500, 1000, 2000$.

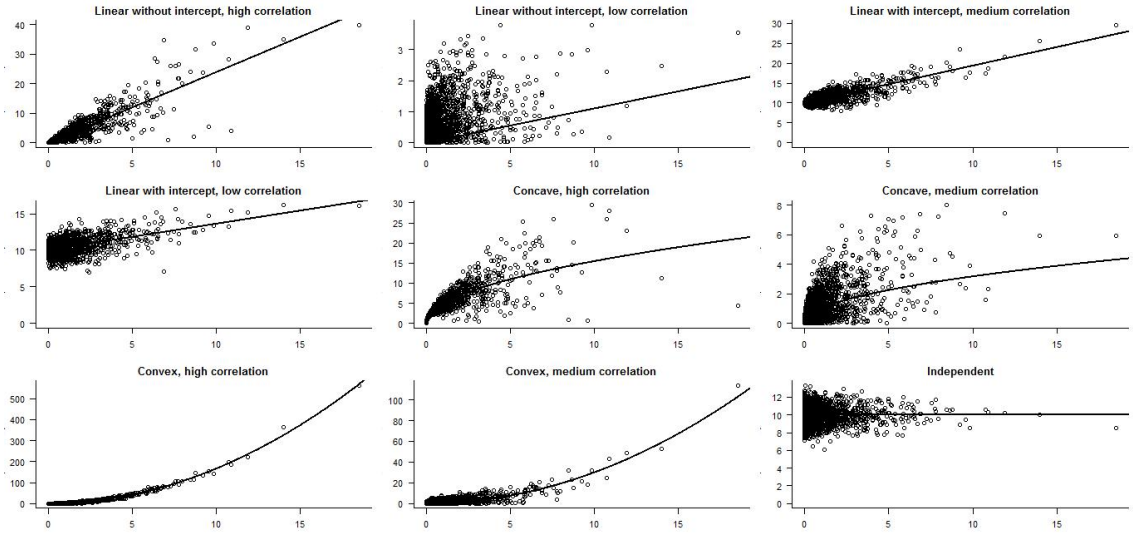

Figure 2: Scatter plots of x and y for nine selected cases in Table 2

For each combination of population size, sample size and study variable, different values of interest (estimates and parameters) are calculated in order to meet the objectives described above. For every population size, the results were similar. Therefore, only the results for the case $N = 2000$ will be presented.

5.2 The approximations to the variance

The first objective of the simulation study is to see how close are the approximate variances of the strategies ($q$ps–π) and ($q$ps–q) to the actual variances, i.e. how close are (33) and (37) to the respective actual variances they intend to approximate. Recall that for the case of the q-estimator, no expression is available to calculate the actual variance; for the case of the π-estimator, even when a compact expression is known for the variance, it requires the calculation of several matrices of dimension $N \times N$. For this reason, the variances were obtained by simulation in the following way.

For each study variable y, $R = 5000$ samples of size n were selected from U using q-sampling, the total of y is estimated using both estimators for each sample. The actual variance is approximated by the variance of the R estimates:

$$\begin{aligned} V_{q\text{ps}}\left(\hat{t}_\pi\right) \approx V_{\text{sim}}\left(\hat{t}_\pi\right) &\equiv \frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{t}_\pi^{(r)} - \bar{\hat{t}}_\pi\right)^2 \quad \text{with } \bar{\hat{t}}_\pi = \frac{1}{R}\sum_{r=1}^{R}\hat{t}_\pi^{(r)} \\ V_{q\text{ps}}\left(\hat{t}_q\right) \approx V_{\text{sim}}\left(\hat{t}_q\right) &\equiv \frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{t}_q^{(r)} - \bar{\hat{t}}_q\right)^2 \quad \text{with } \bar{\hat{t}}_q = \frac{1}{R}\sum_{r=1}^{R}\hat{t}_q^{(r)} \end{aligned} \tag{41}$$

If an approximation to the variance is close to the (simulated) variance, then its ratio will be close to one. Table 3 shows the ratio

$$\frac{AV_{\text{qps}}\left(\hat{t}_{\pi}\right)}{V_{\text{sim}}\left(\hat{t}_{\pi}\right)}$$

It can be seen that, for every variable, the larger the sample size, the better the aproximation approaches the (simulated) variance. With a sampling fraction of 0.05 ($n = 100$), the ratio already lies in the interval (0.95 , 1.05). It does not seem to be a differential behavior among variables, they all seem to approach to one at the same "speed".

n	$y^{(1)}$	$y^{(2)}$	$y^{(3)}$	$y^{(4)}$	$y^{(5)}$	$y^{(6)}$	$y^{(7)}$	$y^{(8)}$	$y^{(9)}$	$y^{(10)}$	$y^{(11)}$	$y^{(12)}$	$y^{(13)}$
1	1.01	1.22	7.51	37.27	1.77	32.59	1.00	1.07	1.43	1.02	0.83	3.77	43.32
4	1.17	1.04	0.98	1.04	1.25	1.28	1.10	1.00	1.02	1.09	1.05	1.01	1.26
9	1.08	1.00	1.02	1.06	1.01	1.12	1.08	1.03	0.99	1.14	1.15	1.00	1.12
16	1.12	1.03	0.98	1.08	1.00	1.04	1.07	1.03	1.02	1.12	1.07	1.03	1.01
25	1.04	1.06	0.99	1.02	1.03	1.02	1.05	1.01	0.97	1.07	0.92	1.05	1.00
36	1.06	1.03	1.02	1.04	1.03	1.02	1.01	1.01	1.02	0.98	1.03	1.07	1.02
49	1.04	0.98	1.00	1.02	1.01	1.03	1.03	1.02	0.96	0.97	0.99	1.03	1.06
64	1.08	1.03	1.01	1.01	1.04	1.00	1.00	1.05	0.97	1.06	1.04	1.05	1.06
81	1.01	0.98	1.00	1.02	1.02	1.03	1.01	1.02	1.01	1.03	0.96	1.01	1.01
100	1.05	1.03	0.98	1.01	1.03	1.01	1.03	1.01	1.00	1.05	0.98	0.99	1.00
121	1.01	1.03	0.99	1.01	1.05	0.98	1.00	1.03	1.01	1.04	1.00	0.98	1.00
144	1.01	1.01	0.98	0.95	1.01	1.02	0.99	0.97	0.98	1.02	1.03	1.00	0.99
169	1.00	1.00	0.98	0.99	1.02	0.99	1.00	1.03	1.01	1.03	1.01	0.96	0.97
196	1.01	1.00	0.98	1.00	0.99	1.02	0.99	1.00	1.03	1.04	0.97	0.98	1.00
225	1.00	1.01	1.01	1.00	1.01	1.03	1.01	1.03	0.98	1.00	0.99	1.03	0.98
256	1.00	0.98	0.95	0.95	1.02	0.98	1.03	0.98	0.97	0.99	1.01	0.99	0.99
289	1.00	0.98	1.00	1.01	0.99	1.01	0.99	0.98	0.96	1.00	0.96	1.00	0.99
324	0.99	1.01	0.97	1.00	1.03	1.02	1.02	0.96	1.02	1.01	1.00	1.00	1.01
361	0.97	1.01	1.01	1.01	1.00	1.04	1.00	1.03	0.98	1.01	1.03	1.00	1.01
400	0.99	1.02	0.97	1.03	1.01	0.98	1.02	0.97	1.04	1.02	1.01	0.99	1.03

Table 3: Approximation to the variance divided by simulated variance for the π-estimator

Table 4 shows the ratio

$$\frac{AV_{\text{qps}}\left(\hat{t}_{q}\right)}{V_{\text{sim}}\left(\hat{t}_{q}\right)}$$

Again, for every variable, the ratio tends to 1 as n increases, although in this case, the approximation is "slower"than for the π-estimator. When the sampling fraction is 0.05 ($n = 100$), the ratio lies in the interval (0.90 , 1.20); a sampling fraction close to 0.15 ($n = 289$) or greater is required for the ratio to lie in the interval (0.95 , 1.05). An interesting observation, that may be pure casuality, is that for the variables with a convex relation, the AV tends to overestimate the variance; for the variables with a concave relation or linear with intercept, the AV tends to underestimate it; whereas for the variables with a linear relation without intercept, the AV lies around the variance. Of course, this last statement cannot be generalized unless a deeper analysis is carried out.

n	$y^{(1)}$	$y^{(2)}$	$y^{(3)}$	$y^{(4)}$	$y^{(5)}$	$y^{(6)}$	$y^{(7)}$	$y^{(8)}$	$y^{(9)}$	$y^{(10)}$	$y^{(11)}$	$y^{(12)}$	$y^{(13)}$
1	1.79	0.27	0.06	0.01	0.00	0.01	0.39	0.16	0.01	7.84	0.92	0.06	0.01
4	1.44	0.62	0.32	0.17	0.22	0.18	0.80	0.50	0.27	3.97	1.85	0.37	0.18
9	1.18	0.82	0.56	0.43	0.49	0.47	0.91	0.70	0.51	2.71	1.91	0.92	0.50
16	1.12	0.90	0.71	0.69	0.64	0.62	0.99	0.80	0.72	1.99	1.49	1.07	0.61
25	1.11	0.91	0.77	0.76	0.73	0.76	1.01	0.88	0.81	1.59	1.26	1.08	0.78
36	1.07	0.92	0.84	0.82	0.84	0.83	0.98	0.93	0.84	1.34	1.29	1.05	0.79
49	1.08	0.95	0.87	0.84	0.84	0.88	0.97	0.99	0.90	1.22	1.21	0.99	0.87
64	1.06	0.96	0.92	0.89	0.92	0.90	1.01	0.98	0.91	1.23	1.13	1.02	0.89
81	1.05	0.98	0.92	0.93	0.92	0.89	1.00	0.97	0.96	1.17	1.07	1.00	0.94
100	1.01	0.99	0.93	0.91	0.91	0.95	0.99	0.98	0.95	1.17	1.04	1.01	0.91
121	1.04	0.97	0.94	0.95	0.96	0.93	1.01	0.96	0.97	1.13	1.08	0.98	0.97
144	1.01	0.96	0.94	0.90	0.95	0.94	0.96	1.02	0.93	1.11	1.08	1.01	0.94
169	1.02	0.97	0.99	0.93	0.98	0.93	0.98	1.00	0.99	1.08	1.05	0.99	0.98
196	1.04	1.00	0.98	0.96	0.98	0.98	0.98	0.96	0.96	1.07	1.02	0.99	0.95
225	1.01	0.99	1.00	0.97	0.98	0.95	0.99	0.98	0.95	1.05	1.03	1.02	0.96
256	1.01	1.03	0.99	0.92	0.97	0.93	1.00	1.01	1.01	1.03	1.03	1.01	0.95
289	1.02	1.01	0.98	0.98	0.95	0.99	1.01	0.97	0.97	1.02	1.01	1.02	0.96
324	1.00	1.02	0.99	0.97	1.00	1.00	1.00	0.97	0.99	1.04	1.04	0.99	0.98
361	0.97	1.02	0.98	0.98	0.97	0.99	0.97	0.95	0.98	1.02	1.00	1.00	1.00
400	1.00	1.03	0.96	0.99	0.98	1.01	0.97	0.96	0.98	1.04	1.03	0.99	0.99

Table 4: Approximation to the variance divided by simulated variance for the q-estimator

5.3 The variance estimators

The second objective of the simulation study is to see how close is the expectation of the biased variance estimators of the two strategies to the approximate variances, i.e. how close are the expectation of (34) and (38) to (33) and (37), respectively. Given that the expectation of (34) and (38) is unknown, it was obtained by simulation in the following way.

Using the $R = 5000$ samples selected as described above for each study variable and each sample size, estimates for the variance were obtained. The (simulated) expectation of the variance estimator is calculated as the average of these R estimates:

$$\begin{aligned} E\left(\hat{V}_{q\text{ps}}\left(\hat{t}_\pi\right)\right) &\approx E_{\text{sim}}\left(\hat{V}_{q\text{ps}}\left(\hat{t}_\pi\right)\right) \equiv \frac{1}{R}\sum_{r=1}^{R}\hat{V}_{q\text{ps}}\left(\hat{t}_\pi^{(r)}\right) \\ E\left(\hat{V}_{q\text{ps}}\left(\hat{t}_q\right)\right) &\approx E_{\text{sim}}\left(\hat{V}_{q\text{ps}}\left(\hat{t}_q\right)\right) \equiv \frac{1}{R}\sum_{r=1}^{R}\hat{V}_{q\text{ps}}\left(\hat{t}_q^{(r)}\right) \end{aligned} \tag{42}$$

If the (simulated) expectation is close to the approximation to the variance, then its ratio will be close to one. Table 5 shows the ratio

$$\frac{E_{\text{sim}}\left(\hat{V}_{q\text{ps}}\left(\hat{t}_\pi\right)\right)}{AV_{q\text{ps}}\left(\hat{t}_\pi\right)}$$

Although biased, it is observed that, for every variable, as long as the sample size increases, the (simulated) expectation tends to be closer to the approximation to the

variance: with a sampling fraction as small as 2% ($n = 36$), the ratio already lies in the interval (0.95 , 1.05). Also, it can be seen that the (simulated) expectation tends to be smaller than the approximation to the variance. The largest bias is associated with the variables $y^{(10)}$, $y^{(11)}$ and $y^{(12)}$, which are those with a convex association with the auxiliary variable. Again, this last result may be due to pure chance.

n	$y^{(1)}$	$y^{(2)}$	$y^{(3)}$	$y^{(4)}$	$y^{(5)}$	$y^{(6)}$	$y^{(7)}$	$y^{(8)}$	$y^{(9)}$	$y^{(10)}$	$y^{(11)}$	$y^{(12)}$	$y^{(13)}$
4	0.81	0.95	0.99	0.98	0.79	0.78	0.88	0.97	1.01	0.89	0.91	0.96	0.78
9	0.92	0.96	0.98	0.92	0.96	0.93	0.91	0.96	0.99	0.89	0.89	0.99	0.91
16	0.91	0.97	0.99	0.93	1.00	0.98	0.94	0.97	0.98	0.90	0.98	0.97	0.96
25	0.94	0.97	0.99	0.96	0.97	1.00	0.96	0.97	1.00	0.96	1.06	0.98	0.98
36	0.97	0.98	0.99	0.96	0.98	1.00	0.96	0.98	1.00	1.00	0.97	0.98	0.99
49	0.99	0.99	1.00	0.98	0.99	1.00	0.97	0.98	1.00	1.01	0.98	0.97	1.00
64	0.97	0.98	1.00	0.99	0.98	1.00	0.98	0.99	1.00	0.96	0.99	0.97	1.00
81	0.98	0.99	1.00	0.99	0.99	0.99	0.99	0.99	1.00	0.98	1.02	1.00	1.00
100	0.98	0.99	1.00	0.99	0.98	1.00	0.99	0.99	1.00	0.96	1.04	0.99	1.00
121	0.98	1.00	0.99	0.99	0.99	1.00	0.99	0.99	1.00	0.97	1.02	1.00	1.00
144	0.99	1.00	1.00	0.99	0.98	1.00	0.99	1.00	1.00	0.97	1.00	0.97	1.00
169	0.99	1.00	1.00	0.99	0.99	1.00	0.99	0.99	1.00	0.98	1.00	1.03	1.00
196	0.99	0.99	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.97	0.99	1.02	1.00
225	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.99	1.01	0.99	1.00
256	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	0.97	1.01	0.99	1.00
289	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.99	1.00	1.00	0.99	0.99	1.00
324	0.99	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.01	1.00
361	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
400	0.99	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.01	0.98	1.00

Table 5: (Simulated) expectation divided by approximation to the variance for the π-estimator

Table 6 shows the ratio

$$\frac{E_{\text{sim}}\left(\hat{V}_{q\text{ps}}\left(\hat{t}_q\right)\right)}{AV_{q\text{ps}}\left(\hat{t}_q\right)}$$

The proposed estimator for the variance of this strategy, (38), was obtained as a consistent estimate of the approximation to the variance (37) and is not an unbiased estimator. Table 6 suggests that the estimator tends to underestimate the approximation to the variance. This behavior is stronger for the variables $y^{(10)}$ and $y^{(11)}$ (convex association with high and medium correlation with x): for the remaining variables, a sampling fraction of 10% ($n = 196$) is enough to obtain a ratio between 0.95 and 1, while, at this point, the ratio is still smaller than 0.9 for the mentioned variables. On the other hand, as expected, as the sample size increases, the bias decreases.

5.4 Comparing the bias of $q$ps with alternative strategies

The performance of a sampling strategy is often assessed in terms of its bias and its variance. Five strategies were compared in the simulation study:

- The strategy ($q$ps–π), described in Section 3.
- The strategy ($q$ps–q), described in Section 4.

n	$y^{(1)}$	$y^{(2)}$	$y^{(3)}$	$y^{(4)}$	$y^{(5)}$	$y^{(6)}$	$y^{(7)}$	$y^{(8)}$	$y^{(9)}$	$y^{(10)}$	$y^{(11)}$	$y^{(12)}$	$y^{(13)}$
4	0.14	0.40	0.79	0.96	0.85	0.90	0.17	0.50	0.87	0.01	0.13	0.50	0.94
9	0.34	0.54	0.78	0.81	0.78	0.76	0.33	0.62	0.80	0.05	0.22	0.55	0.75
16	0.49	0.67	0.83	0.81	0.82	0.85	0.48	0.73	0.84	0.13	0.34	0.63	0.82
25	0.63	0.75	0.89	0.87	0.86	0.86	0.61	0.81	0.89	0.25	0.45	0.69	0.86
36	0.75	0.82	0.91	0.91	0.90	0.90	0.69	0.86	0.91	0.37	0.53	0.76	0.92
49	0.80	0.87	0.95	0.92	0.92	0.93	0.76	0.89	0.93	0.49	0.60	0.82	0.93
64	0.85	0.88	0.95	0.95	0.95	0.94	0.81	0.92	0.95	0.57	0.69	0.85	0.93
81	0.87	0.91	0.96	0.96	0.97	0.96	0.84	0.94	0.95	0.65	0.76	0.88	0.95
100	0.89	0.93	0.97	0.97	0.97	0.97	0.88	0.95	0.97	0.69	0.82	0.89	0.98
121	0.91	0.94	0.97	0.97	0.98	0.98	0.90	0.96	0.97	0.75	0.84	0.91	0.97
144	0.94	0.96	0.98	0.98	0.98	0.98	0.93	0.96	0.99	0.78	0.86	0.93	0.98
169	0.95	0.96	0.98	0.98	0.98	0.99	0.93	0.97	0.98	0.81	0.88	0.96	0.99
196	0.96	0.97	0.99	0.98	0.98	0.99	0.95	0.97	0.99	0.83	0.88	0.96	0.98
225	0.95	0.97	0.99	0.99	0.98	0.99	0.94	0.98	1.00	0.87	0.92	0.96	0.99
256	0.96	0.98	0.99	0.99	0.99	0.99	0.95	0.98	0.99	0.86	0.93	0.96	0.98
289	0.97	0.98	0.98	1.00	0.99	0.99	0.97	0.99	0.99	0.91	0.92	0.96	0.99
324	0.97	0.98	0.99	0.99	0.99	0.99	0.96	0.99	0.99	0.91	0.93	0.98	0.99
361	0.98	0.98	1.00	0.99	0.99	0.99	0.98	0.99	0.99	0.92	0.95	0.97	0.99
400	0.97	0.99	1.00	0.99	1.00	1.00	0.97	0.99	0.99	0.94	0.97	0.97	1.00

Table 6: (Simulated) expectation divided by approximation to the variance for the q-estimator

- The strategy (SRS–π) is often considered as a benchmark. It is a very known strategy that due to its simplicity is widely used. This strategy does not use auxiliary information, which is, at the same time, an advantage and a disadvantage. Is an advantage in the sense that it can be implemented as long as a sampling frame is available, nothing else is required. It is a disadvantage in the sense that if auxiliary information is available, the strategy does not allow for using it.

- Pareto sampling is a measurable design, which is considered to be an efficient strategy when coupled with the $q\pi$-estimator. Its main drawback is that it is not a strict πps.

- The strategy (Systematic–π) has some interesting characteristics that make it attractive: the design is a strict πps and is often considered as a very efficient strategy. On the other hand, as discussed in Särndal *et. al.* (1992), being a non-measurable design, is hard to obtain in practice valid variance estimation and valid confidence intervals.

No tables of results from the simulation study about the bias are presented here. The reason for this is that the first, third and fifth strategies use the π-estimator, which is known to be unbiased. Also, in Section 4, it was shown that the q-estimator is unbiased under q-sampling. Therefore, the strategy (Pareto–$q\pi$) is the only one that uses a biased estimator. However, it is assymptotically unbiased, as mentioned in Rosén (2000). The results of the simulation study show that Pareto sampling is "virtually"unbiased even for small sample sizes. So bias is not a concern for the strategies under comparison.

5.5 Comparing the variance of $q$ps with alternative strategies

Exact expressions for the variance are known only for the strategies (SRS–π) and ($q$ps–π). For the latter, the expression involves a large number of terms which makes it hard to compute in practice. Therefore, the variances for the five strategies were approximated by simulation as described in Section 5.2.

Rosén (1997), introduces the Variance Increase of a given strategy relative to Pareto sampling. In an analogous way, here the Variance Increase —VI— of a given strategy relative to (SRS–π) is defined as

$$\mathrm{VI}\left(p(\cdot),\hat{t}\right) = \frac{V_p\left(\hat{t}\right)}{V_{\mathrm{SRS}}\left(\hat{t}_\pi\right)} \tag{43}$$

The VI is a measure of the efficiency of a given strategy compared to (SRS–π). When the Variance Increase is smaller (greater) than one, the strategy is more (less) efficient than SRS. As it is not possible to calculate the actual value of VI, it was obtained through simulation as

$$\mathrm{VI}_{\mathrm{sim}}\left(p(\cdot),\hat{t}\right) = \frac{V_{\mathrm{sim},p}\left(\hat{t}\right)}{V_{\mathrm{SRS}}\left(\hat{t}_\pi\right)}, \tag{44}$$

The (simulated) VI —SVI— are compared for different sample sizes and the thirteen study variables described in Table 2. Table 7 shows the SVI for $y^{(1)}$, i.e. the case of linear without intercept association between x and y and a high correlation. (The SVI for SRS is presented only for completeness, although it is, of course, known that its actual value should be equal to one.) It can be seen that, in this case, (SRS–π) is the less efficient strategy. ($q$ps–π) is slightly more efficient, but as the sample size increases, this efficiency vanishes. The strategy ($q$ps–q) is visibly more efficient, with a SVI around 0.2. The two strategies (Pareto–$q\pi$) and (Systematic–π) are even more efficient, and they have very similar SVI values. Similar results were obtained for $y^{(10)}$ and $y^{(11)}$ (convex, with high and medium correlation, respectively), therefore the results of these variables are shown only in the appendix.

Table 8 shows the SVI for the case of linear association between x and y without intercept and medium correlation, $y^{(2)}$. Here, (Pareto–$q\pi$) seems to be a little less efficient than ($q$ps–q), which is, in turn, less efficient than (SRS–π). ($q$ps–π) shows the same behavior than in the previous case: it is slightly more efficient than (SRS–π) for small sample sizes, but this efficiency vanishes in medium to large samples. Finally, (Systematic–π) is the most efficient strategy in this case.

The results for the case of linear association with a low correlation between x and y, variable $y^{(3)}$, are shown in Table 9. In this case the situation is completely opposed to that observed in Table 7 for $y^{(1)}$: (SRS–π) is the most efficient strategy. ($q$ps–π) is slightly less efficient. The SVI of ($q$ps–q) is around 3. (Pareto–$q\pi$) is less efficient; and (Systematic–π) is the least efficient strategy in this case. A similar situation was observed for variables $y^{(4)}$, $y^{(5)}$, $y^{(6)}$, $y^{(9)}$ and $y^{(13)}$ (the three cases with a linear with intercept association, concave with low correlation and the independence case). The tables for these variables are shown in the appendix.

The results for $y^{(7)}$, concave with high correlation, are shown in Table 10. In this case, ($q$ps–q) is less efficient than (SRS–π), with values of SVI between 1.10 and 1.15. As before, ($q$ps–π) is a little more efficient than SRS for small samples. Pareto sampling is more efficient than SRS, with SVI around 0.70. In this case, systematic sampling is, again, the most efficient strategy.

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
1	0.06	0.06	0.06	0.07	0.98
4	0.13	0.32	0.06	0.06	1.00
9	0.17	0.55	0.06	0.06	0.98
16	0.18	0.64	0.06	0.05	1.00
25	0.19	0.77	0.06	0.06	0.98
36	0.19	0.81	0.06	0.06	0.99
49	0.19	0.86	0.06	0.07	0.96
64	0.19	0.85	0.06	0.05	0.98
81	0.20	0.92	0.06	0.06	0.98
100	0.20	0.89	0.06	0.06	0.99
121	0.20	0.94	0.05	0.03	1.03
144	0.20	0.95	0.05	0.06	0.96
169	0.20	0.97	0.05	0.05	0.96
196	0.20	0.96	0.05	0.05	1.00
225	0.20	0.97	0.05	0.09	1.01
256	0.20	0.98	0.04	0.03	1.06
289	0.20	0.98	0.04	0.05	1.02
324	0.20	0.99	0.04	0.03	1.00
361	0.21	1.02	0.04	0.04	0.99
400	0.21	0.99	0.03	0.02	1.03

Table 7: (Simulated) VI for five sampling strategies, study variable: $y^{(1)}$

Table 11 shows the result for the concave case with medium correlation, $y^{(8)}$. Here, systematic sampling is sometimes more efficient than SRS, but in general can be said to be less efficient. ($q$ps–q) is, also, less efficient than SRS, with SVI around 1.15. Pareto sampling is even less efficient with SVI around 1.5. Regarding ($q$ps–π), its behavior is, as always, very similar to SRS, being slightly more efficient in this case.

Table 12 shows the result for $y^{(12)}$, this is the convex case with low correlation. Here, Pareto and Systematic sampling are less efficient than SRS. ($q$ps–q) is more efficient than SRS (with SVI around 0.5). As always, the efficiency of ($q$ps–π) is comparable to that of SRS.

A summary of the results observed from the simulation regarding the efficiency of the five strategies is as follows:

- The variance of the strategy ($q$ps–π) is always similar to the variance of (SRS–π), especially for medium to large samples. For small samples the former is sometimes more and sometimes less efficient than the latter. A loose explanation for this is that, as mentioned in Section 3, a perfect correlation between y and the q-values does not result in a zero variance, somehow, this strategy "wastes"some of the information provided by the auxiliary variable, so much, that the effect of x vanishes for large sample sizes, and from there on, the reduction in variance is due solely to the increasing in sample size, as is the case in SRS.

- A common result was that when ($q$ps–q) is more efficient than (SRS–π), then (Pareto–$q\pi$) and (Systematic–π) are even more efficient. On the other hand, when ($q$ps–q) is less efficient than (SRS–π), then (Pareto–$q\pi$) and (Systematic–

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
1	1.05	1.05	1.11	1.46	1.06
4	1.14	0.58	1.20	1.13	0.97
9	1.13	0.75	0.98	0.94	1.02
16	1.15	0.82	1.02	0.98	1.02
25	1.19	0.84	1.28	1.13	0.99
36	1.22	0.89	1.29	1.00	1.00
49	1.20	0.96	1.21	0.98	1.02
64	1.20	0.92	1.53	0.89	1.04
81	1.19	0.98	1.16	0.94	0.99
100	1.18	0.94	1.10	1.11	0.98
121	1.22	0.95	1.21	0.87	1.03
144	1.23	0.97	1.13	1.09	0.98
169	1.22	0.99	1.12	0.70	1.02
196	1.19	0.99	1.27	0.94	1.02
225	1.20	0.97	1.18	0.88	1.03
256	1.16	1.01	2.11	0.59	1.00
289	1.18	1.01	1.14	0.87	1.00
324	1.17	0.98	1.22	0.54	1.02
361	1.17	0.98	1.18	0.73	1.01
400	1.16	0.97	1.18	0.63	0.99

Table 8: (Simulated) VI for five sampling strategies, study variable: $y^{(2)}$

π) are even less efficient. Somehow ($q$ps–q) seems to be more "conservative"in the sense that, not too much is gained, but not too much is lose.

- The ideal situation in practice is represented by $y^{(1)}$, where the study variable is highly correlated with x and there is also a linear association without intercept between the two variables. In this case Pareto and systematic sampling were more efficient than the remaining strategies.

5.6 Comparing the coverage of $q$ps with alternative strategies

Often, the distribution of the estimates obtained by a given strategy is approximated by a normal or a t distribution. Therefore it is expected that estimated confidence intervals of the form

$$\left(\hat{t} - z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}\left(\hat{t}\right)}\ ,\ \hat{t} + z_{1-\frac{\alpha}{2}}\sqrt{\hat{V}\left(\hat{t}\right)}\right) \tag{45}$$

(where $z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of a normal or a t distribution with $n-1$ degrees of freedom) will cover the parameter, approximately, with a probability of $100(1-\alpha)\%$.

The *coverage* of a strategy is defined as

$$E_p\left(C(s)\right) = \sum_{\Omega} p(s)C(s) \tag{46}$$

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
1	4.41	4.41	7.26	3.62	1.00
4	3.37	0.86	9.33	8.61	1.01
9	3.08	0.85	11.88	15.79	1.01
16	2.94	0.92	7.97	17.40	1.01
25	2.98	0.95	18.45	981.37	0.97
36	2.90	0.93	42.67	22.02	0.99
49	2.89	0.97	14.55	11.51	1.03
64	2.81	0.96	12.95	22.91	1.02
81	2.85	0.98	65.83	124.94	1.01
100	2.82	1.01	11.11	16.57	1.00
121	2.84	1.00	9.62	218.99	0.99
144	2.86	1.00	12.03	10.28	0.95
169	2.71	1.01	14.12	5.94	1.00
196	2.76	1.01	14.10	34.72	0.99
225	2.70	0.98	15.94	20.51	1.01
256	2.75	1.05	33.57	11.40	1.03
289	2.78	1.00	15.89	9.78	0.99
324	2.74	1.03	44.57	319.12	0.99
361	2.79	0.98	19.96	17.67	1.03
400	2.84	1.03	31.17	11.74	1.02

Table 9: (Simulated) VI for five sampling strategies, study variable: $y^{(3)}$

where

$$C(s) = \begin{cases} 1, & \text{if the inteval (45) covers the parameter} \\ 0, & \text{otherwise} \end{cases}$$

The coverage is a parameter and no compact expression is known for it. So, it was simulated in the following way. For each of the $R = 5000$ samples for each strategy (except for systematic sampling), the 95% confidence interval (45) using the t distribution, and then $C(s)$, were calculated. The simulated coverage —SC— is defined as the average of the $C(s)$-values:

$$E_p\left(C(s)\right) \approx E_{sim}\left(C(s)\right) = \frac{1}{R}\sum_{i=1}^{R} C(s)$$

Given that no variance estimator has been defined here for systematic sampling, intervals of the form (45) cannot be computed. Therefore, the following alternative interval was computed for this strategy

$$\left(\hat{t} - z_{1-\frac{\alpha}{2}}\sqrt{V_{sim}\left(\hat{t}\right)}\ ,\ \hat{t} + z_{1-\frac{\alpha}{2}}\sqrt{V_{sim}\left(\hat{t}\right)}\right) \tag{47}$$

It is important to have in mind this adaptation for the following comparisons, as the latter interval is considered to be closer to the expected $100(1-\alpha)\%$ than the former, so the comparison among strategies is not completely fair.

Table 13 shows the (simulated) coverage of the five strategies for the variable $y^{(1)}$. In this case, with a sample size as small as $n = 36$ (sampling fraction around

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
1	0.78	0.78	0.57	0.76	1.02
4	0.90	0.35	0.70	0.58	0.99
9	1.00	0.57	0.69	0.42	1.00
16	1.00	0.69	0.72	0.33	0.99
25	1.02	0.78	0.73	0.39	0.99
36	1.08	0.86	0.74	0.63	0.99
49	1.10	0.88	0.79	0.30	1.00
64	1.07	0.92	0.75	0.38	0.96
81	1.08	0.92	0.74	0.27	0.99
100	1.10	0.92	0.75	0.28	1.00
121	1.08	0.96	0.74	0.18	1.00
144	1.14	0.97	0.77	0.33	0.98
169	1.12	0.97	0.73	0.18	1.01
196	1.12	0.98	0.73	0.18	1.01
225	1.12	0.96	0.71	0.21	1.01
256	1.10	0.95	0.69	0.24	0.98
289	1.09	0.99	0.67	0.19	1.01
324	1.10	0.97	0.70	0.15	1.00
361	1.13	0.99	0.65	0.13	0.99
400	1.13	0.97	0.66	0.20	1.02

Table 10: (Simulated) VI for five sampling strategies, study variable: $y^{(7)}$

0.02), the coverage of the five strategies is already greater than 0.9, a lack of coverage that, can be, somehow, neglected. However, systematic sampling shows the largest coverage, being even greater than the expected 95%; Pareto sampling lies very close to the expected value. As before, the behavior of ($q$ps–π) is very similar to (SRS–π), especially for large samples, while for small samples the former performs slightly better than the latter. Both of them lie a little below the expected 95%. ($q$ps–q) has the smallest coverage in this case. The results for variable $y^{(10)}$ are similar in the sense that the coverage of (systematic–π) is larger than (Pareto–$q\pi$), which is larger than ($q$ps–π), which is slightly larger than (SRS–π), which is, in turn, larger than ($q$ps–p). However, the differences are more notorious in that case. Results are shown in the appendix.

The results in Table 14 correspond to the variable $y^{(2)}$. In this case, again, a sampling fraction of 0.02 ($n = 36$) is enough for the five strategies to show a SC larger than 0.9. Specifically, systematic sampling has the largest coverage, being greater than the expected 95%. For small samples, ($q$ps–π) has a SC slightly greater than (SRS–π); a difference that vanishes in medium to large samples. However, both strategies have a SC very close to the expected 95%. ($q$ps–q) is a little "slower"than the former strategies to achieve the expected coverage, even so, it performs very well. In this case, Pareto sampling has the smallest coverage. Similar results were observed for $y^{(3)}$, $y^{(4)}$, $y^{(5)}$, $y^{(6)}$, $y^{(7)}$, $y^{(8)}$, $y^{(9)}$ and $y^{(13)}$, although the differences among strategies are more dramatical for some cases, for example in $y^{(4)}$, $y^{(5)}$ and $y^{(6)}$, where Pareto has a SC below 75%. The results are shown in the appendix.

Table 15 shows the simulated coverage for $y^{(11)}$. Again, systematic sampling lies

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
1	1.37	1.37	1.07	1.07	1.02
4	1.22	0.59	1.31	1.30	0.99
9	1.19	0.73	1.24	2.79	1.03
16	1.19	0.80	1.18	1.38	1.00
25	1.16	0.88	1.37	0.87	1.00
36	1.14	0.91	1.46	0.95	1.02
49	1.09	0.92	1.31	1.08	1.02
64	1.12	0.90	1.43	1.11	0.98
81	1.13	0.94	1.31	1.05	1.02
100	1.13	0.96	1.29	0.77	0.97
121	1.16	0.94	1.56	1.26	1.03
144	1.10	1.00	1.30	1.50	0.99
169	1.12	0.95	1.48	1.32	1.00
196	1.18	0.99	1.52	0.92	0.99
225	1.16	0.96	1.40	1.24	1.02
256	1.12	1.01	1.35	0.93	0.99
289	1.17	1.01	1.80	1.32	1.00
324	1.17	1.03	1.41	1.33	0.96
361	1.20	0.96	1.48	0.88	0.96
400	1.18	1.03	1.28	1.17	1.01

Table 11: (Simulated) VI for five sampling strategies, study variable: $y^{(8)}$

above the expected 95%. ($q$ps–q) follows, with SC greater than 0.9 starting from $n = 36$. Pareto sampling requires, in this case, a sample greater than 64 to overcome the threshold of a 90% SC. Finally, ($q$ps-π) and (SRS-π) are the worst performers in this case.

Table 16 shows the results for $y^{(12)}$. As before, systematic sampling has the largest SC, lying above the expected 95%. ($q$ps–q) overcomes the threshold of the 90% at a sample size around $n = 36$. ($q$ps–π) and (SRS–π) seem to lie around a SC of 90%. Finally, Pareto sampling has the smallest SC, with values around 80%.

A summary of the results observed from the simulation regarding the coverage of the five strategies follows:

- (Systematic–π) showed the best results, with simulated coverage over the expected 95% in almost every case. Here is important to recall that the confidence intervals for this strategy were obtained in a different way, so the comparison is not completely fair.

- Even with small sample sizes the coverage of ($q$ps–q) is already greater than 90%. Its coverage seems to be very close to the expected 95% in every case except the convex with high-correlation case.

- As was the case in the variance, ($q$ps–π) and (SRS–π) have a very similar behavior, with the former being slightly better than the latter for small sample sizes. Their coverage lies between 0.9 and 0.95 in every case, except the convex association case where it is smaller than 0.9.

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
1	5.39	5.39	3.14	1.25	1.18
4	1.29	0.30	5.00	2.88	0.94
9	0.55	0.48	8.79	2.86	0.92
16	0.47	0.59	11.47	2.15	1.01
25	0.47	0.67	3.15	3.50	1.06
36	0.48	0.72	5.52	2.20	0.99
49	0.51	0.80	7.24	24.68	0.99
64	0.50	0.81	6.42	2.68	1.02
81	0.51	0.87	4.73	3.62	1.02
100	0.50	0.91	10.14	4.23	0.97
121	0.52	0.94	4.15	5.14	1.04
144	0.50	0.93	5.35	3.06	1.05
169	0.52	0.98	14.23	3.51	0.95
196	0.51	0.97	4.78	16.90	1.00
225	0.50	0.93	4.36	6.35	0.99
256	0.50	0.96	5.13	3.83	0.97
289	0.50	0.96	5.83	3.14	0.97
324	0.52	0.97	12.89	4.27	0.99
361	0.51	0.97	7.08	122.62	1.01
400	0.52	0.98	10.51	4.23	1.00

Table 12: (Simulated) VI for five sampling strategies, study variable: $y^{(12)}$

- (Pareto–π) seems to be the most affected by the study variable: in the high-correlation case of the linear with intercept and the convex type association, its coverage lies very close to the expected 95%. However, in some cases, its coverage is strongly affected, taking values significantly below 90%.

6 Conclusions and comments

This section is divided into three parts. In the first part a summary of the implementation of q-sampling in practice is presented. In the second part, some comments about the design itself are presented. The third part discusses the performance of the different strategies compared in Section 5.

Implementation of q-sampling in practice In order to implement q-sampling in practice, a sampling frame should be available and a auxiliary variable, q, should be identified. In general, q must satisfy (22) and (23). For $q$ps, the q-values are defined proportional to a variable x which is always greater than zero.

If the sample size is not defined beforehand, a proxy variable for the variable of interest and the approximation to the variance, (37), may be of help. Once the size has been defined, the sample is selected using the method described in Section 2.

The y-values for the selected sample are collected and the total is estimated by (35) (or the π-estimator). The variance is estimated by (38) and an approximately 95% confidence interval is calculated by (45).

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
4	0.63	0.92	0.95	1.00	0.79
9	0.75	0.89	0.95	0.98	0.82
16	0.83	0.89	0.96	0.97	0.86
25	0.87	0.90	0.95	0.96	0.88
36	0.90	0.91	0.95	0.95	0.90
49	0.91	0.92	0.95	0.96	0.91
64	0.92	0.92	0.95	0.97	0.92
81	0.93	0.92	0.95	0.95	0.92
100	0.93	0.93	0.95	0.94	0.92
121	0.94	0.93	0.95	0.97	0.93
144	0.94	0.93	0.95	1.00	0.93
169	0.94	0.93	0.95	0.97	0.93
196	0.94	0.94	0.95	0.95	0.94
225	0.94	0.94	0.95	0.96	0.94
256	0.94	0.94	0.95	0.94	0.93
289	0.94	0.94	0.95	0.97	0.94
324	0.94	0.94	0.95	0.95	0.94
361	0.94	0.94	0.95	0.93	0.94
400	0.94	0.94	0.95	0.96	0.94

Table 13: (Simulated) coverage for five sampling strategies. Study variable: $y^{(1)}$

Comments about the design q-sampling has some interesting theoretical properties, among them: **i.** it belongs to the family of without-replacement fixed-size designs that use auxiliary information; which is a family often considered to be very efficient; **ii.** the inclusion probabilities of any order are easily obtained; **iii.** because of **ii.**, the unbiased π-estimator can be used together with q-sampling; **iv.** the alternative (and apparently more efficient) q-estimator is also unbiased under this design.

On the other hand, it has also some drawbacks that may be considered as future research. These drawbacks are associated with the estimator which is coupled with the design:

- Using the π-estimator, the ideal situation is a πps design. Unfortunately, to obtain a πps from q-sampling is possible only for very restricted (and uninteresting) cases. Furthermore, if the πps condition is relaxed, and we just decide to use the π-estimator with a given set of q-values, the result seems to be a strategy that behaves almost as simple random sampling, so "wasting"most of the information provided by the auxiliary variable.

- Using the q-estimator, no compact expression for the variance is available. An approximated expression that performs well (not excellent) was obtained by the Taylor's linearization method. It would be interesting to try with a second order approximation. Regarding the variance estimator, a consistent estimator was proposed; however, the results show that it tends to underestimate the AV and medium to large sample sizes are required in order for its bias to be negligible. Further research is needed in order to improve this estimator.

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
4	0.68	0.89	0.88	0.99	0.86
9	0.78	0.90	0.88	0.97	0.87
16	0.85	0.91	0.89	0.98	0.90
25	0.88	0.93	0.89	0.98	0.91
36	0.90	0.93	0.90	0.97	0.92
49	0.91	0.93	0.90	0.97	0.93
64	0.92	0.94	0.91	0.97	0.93
81	0.93	0.93	0.92	0.96	0.94
100	0.93	0.94	0.92	0.96	0.94
121	0.93	0.95	0.92	0.96	0.94
144	0.93	0.94	0.93	0.96	0.95
169	0.94	0.94	0.92	0.97	0.94
196	0.94	0.94	0.93	0.97	0.94
225	0.94	0.95	0.93	0.95	0.94
256	0.95	0.94	0.93	0.96	0.95
289	0.95	0.94	0.93	0.95	0.95
324	0.95	0.94	0.93	0.97	0.94
361	0.95	0.95	0.93	0.97	0.94
400	0.95	0.95	0.94	0.96	0.95

Table 14: (Simulated) coverage for five sampling strategies. Study variable: $y^{(2)}$

Conclusions about the performance of the five strategies It is important to recall that the conclusions are based on a simulation study, so they cannot be generalized in a straightforward way. Even so, according to what was observed in the study, some conclusions are as follows:

- The Monte-Carlo simulation study allowed to compare the five strategies under thirteen different variables of study. Variable $y^{(1)}$ represents the ideal situation in practice: linear association between the auxiliary variable, x, and the study variable, y, with a high correlation and without intercept. In this situation, the strategy (Pareto–$q\pi$) shows up as the best option: its bias can be neglected, its coverage lies over the expected 95% and it is significantly more efficient than the SRS and the two strategies that involve $q$ps. Even when its efficiency can be compared to systematic sampling, the balance tilts towards Pareto sampling when we take into account that it is a measurable design, the approximate variance seems to be very close to the actual variance and the variance estimator seems to be almost free of bias, even for small samples.

- Regarding the other variables, and somehow reinforcing a conjecture presented in Rosén (1997), Pareto sampling seems to be, also, the best choice when the association between x and y is markedly convex or linear without intercept with a medium correlation. On the other hand, when the association between x and y is concave, linear with intercept or it has a low correlation, Pareto sampling becomes very inefficient compared to simple random sampling and its coverage seems to be significantly smaller than the expected.

- Systematic sampling coupled with the π-estimator is free of bias, it seems that

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
4	0.68	0.89	0.90	0.99	0.81
9	0.80	0.86	0.89	0.98	0.80
16	0.85	0.87	0.88	0.97	0.82
25	0.87	0.86	0.90	0.97	0.84
36	0.90	0.87	0.89	0.97	0.85
49	0.91	0.87	0.89	0.96	0.86
64	0.91	0.88	0.89	0.98	0.88
81	0.92	0.88	0.91	0.96	0.87
100	0.93	0.89	0.90	0.96	0.87
121	0.93	0.89	0.90	0.96	0.88
144	0.93	0.89	0.90	0.97	0.89
169	0.92	0.89	0.92	0.97	0.88
196	0.92	0.90	0.91	0.98	0.90
225	0.93	0.90	0.91	0.96	0.89
256	0.93	0.91	0.91	0.96	0.90
289	0.93	0.90	0.92	0.96	0.90
324	0.93	0.90	0.92	0.96	0.91
361	0.93	0.91	0.92	0.97	0.90
400	0.93	0.91	0.92	0.96	0.91

Table 15: (Simulated) coverage for five sampling strategies. Study variable: $y^{(11)}$

coverage is not a problem here and its efficiency is comparable to that of Pareto sampling. Even so, being a non-measurable design, variance estimation is quite problematic.

- $q$ps coupled with the π-estimator showed to be almost equivalent to simple random sampling. This result makes this strategy uninteresting, if we take into account the characteristic simplicity of SRS.

- $q$ps is also unbiased when coupled with the q-estimator, its coverage showed to be close to the expected 95%, except for convex-high correlation case. Regarding its efficiency, in general it lies in between simple random sampling and Pareto sampling: when ($q$ps–q) is more efficient than SRS, Pareto is even more efficient; when ($q$ps–q) is less efficient than SRS, Pareto is even less efficient. It can be said to be more conservative than Pareto or systematic sampling.

- However, there were some cases where ($q$ps-q) was more efficient than the remaining strategies. Taking into account that $q$ps is only a particular case of q-sampling, it may be interesting to investigate if it is possible to improve the performance of the strategy, possibly considering a different setting for the q-values.

A necessary question is whether to recommend or not the use of q-sampling. As always, a necessary answer is that further research and improvements may be done about q-sampling. So far, in the rare case of a study interested only in one variable, the answer would be: no. If powerful auxiliary information is available, Pareto sampling

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
4	0.74	0.91	0.81	0.99	0.87
9	0.82	0.90	0.74	0.99	0.84
16	0.86	0.89	0.73	0.98	0.85
25	0.89	0.88	0.75	0.98	0.85
36	0.91	0.89	0.74	0.98	0.85
49	0.92	0.88	0.76	1.00	0.86
64	0.93	0.89	0.78	0.98	0.87
81	0.93	0.88	0.78	0.98	0.88
100	0.93	0.88	0.79	0.98	0.88
121	0.93	0.89	0.79	0.99	0.88
144	0.93	0.89	0.79	0.98	0.89
169	0.94	0.90	0.80	0.98	0.90
196	0.94	0.90	0.80	0.99	0.90
225	0.94	0.90	0.81	0.98	0.90
256	0.94	0.90	0.81	0.98	0.90
289	0.94	0.91	0.81	0.98	0.91
324	0.94	0.91	0.81	0.97	0.91
361	0.94	0.91	0.81	1.00	0.91
400	0.93	0.91	0.82	0.97	0.91

Table 16: (Simulated) coverage for five sampling strategies. Study variable: $y^{(12)}$

seems to performs better; on the other hand, if weak or no auxiliary information at all is available, SRS would be the choice.

Probably, in a multi-purpose study where good auxiliary information is available for the most important variables, q-sampling could be used. It will reduce the variance of the most important variables, without strongly impacting the remaining variables.

References

Casella, G. and Berger, R. (2002). *Statistical inference*. Thomson Learning.

Hanif, M. and Brewer K. R. W. (1980). *Sampling with Unequal Probabilities without Replacement: A Review*. International Statistical Review **48**, 317-335.

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

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.

Rosén, B. (1997). *On sampling with probability proportional to size*. Journal of statistical planning and inference **62**, 159-191.

Rosén, B. (2000). *On inclusion probabilities for order πps sampling*. Journal of statistical planning and inference **90**, 117-143.

Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer.

Tillé, Y. (2006). *Sampling algorithms*. Springer.

Wolter, K. (2007). *Introduction to variance estimation*. Springer.

A Proof of results

Proof of Result 1. In order to prove Result 1, we need to show that conditions (3) and (4) are satisfied, in other words, we need to show that **i.** the sum of $p(s)$ over the support equals one, and, **ii.** $p(s)$ is greater than zero for any without-replacement sample of fixed size n in the support. The proof of the second part is straighforward:

By (23), for any s in Ω_{q} we have

$$p(s) = \frac{1}{\binom{N-1}{n-1}} \sum_{s} q_k \geq \frac{1}{\binom{N-1}{n-1}} \sum_{i=1}^{n} q_{(i)} > 0$$

In words, if the sum of the n smallest q-values is greater than zero, then the sum of any set of size n of q-values, will be greater than zero.

For the first part, we need to show that

$$\sum_{\Omega_{\mathrm{q}}} p(s) = \sum_{\Omega_{\mathrm{q}}} \frac{1}{\binom{N-1}{n-1}} \sum_{s} q_k = 1$$

Let I_k be an indicator that takes the value 1 if the element k belongs to the sample s and 0 otherwise, then, we have

$$\sum_{\Omega_{\mathrm{q}}} \frac{1}{\binom{N-1}{n-1}} \sum_{s} q_k = \sum_{\Omega_{\mathrm{q}}} \frac{1}{\binom{N-1}{n-1}} \sum_{U} q_k I_k = \frac{1}{\binom{N-1}{n-1}} \sum_{\Omega_{\mathrm{q}}} \sum_{U} q_k I_k = \frac{1}{\binom{N-1}{n-1}} \sum_{U} \sum_{\Omega_{\mathrm{q}}} q_k I_k$$

But each element is included in $\binom{N-1}{n-1}$ samples, so

$$\frac{1}{\binom{N-1}{n-1}} \sum_{U} \sum_{\Omega_{\mathrm{q}}} q_k I_k = \frac{1}{\binom{N-1}{n-1}} \sum_{U} \binom{N-1}{n-1} q_k = \binom{N-1}{n-1} \frac{1}{\binom{N-1}{n-1}} \sum_{U} q_k = \sum_{U} q_k$$

But, by (22),

$$\sum_{U} q_k = 1$$

□

Proof of Result 2. The first order inclusion probability of the element k is the sum of the probabilities of the samples that include that element, i.e.

$$\pi_k = \sum_{s \ni k} p(s) = \sum_{s \ni k} \frac{1}{\binom{N-1}{n-1}} \sum_{s} q_i = \frac{1}{\binom{N-1}{n-1}} \sum_{s \ni k} \sum_{s} q_i$$

Using the indicators defined above

$$\pi_k = \frac{1}{\binom{N-1}{n-1}} \sum_{s \ni k} \sum_{s} q_i = \frac{1}{\binom{N-1}{n-1}} \sum_{s \ni k} \sum_{U} q_i I_i = \frac{1}{\binom{N-1}{n-1}} \sum_{s \ni k} \left[q_k + \sum_{U^{(k)}} q_i I_i \right]$$

where $U^{(k)}$ is the set of all elements in U except the element k. The last expression is obtained by noting that the k-th element is included in all terms in the sum. This expression can be rewritten as

$$\pi_k = \frac{1}{\binom{N-1}{n-1}} \sum_{s \ni k} \left[q_k + \sum_{U^{(k)}} q_i I_i \right] = \frac{1}{\binom{N-1}{n-1}} \left[\sum_{s \ni k} q_k + \sum_{s \ni k} \sum_{U^{(k)}} q_i I_i \right]$$

But, there are $\binom{N-1}{n-1}$ samples that include the k-th element. And, of those samples, $\binom{N-2}{n-2}$ include each of the remaining elements, so

$$\pi_k = \frac{1}{\binom{N-1}{n-1}} \left[\sum_{s \ni k} q_k + \sum_{s \ni k} \sum_{U^{(k)}} q_i I_i \right] = \frac{1}{\binom{N-1}{n-1}} \left[\binom{N-1}{n-1} q_k + \binom{N-2}{n-2} \sum_{U^{(k)}} q_i \right] =$$
$$\frac{1}{\binom{N-1}{n-1}} \left[\binom{N-1}{n-1} q_k + \binom{N-2}{n-2} (1 - q_k) \right] =$$
$$\frac{1}{\binom{N-1}{n-1}} \left[\binom{N-1}{n-1} q_k + \binom{N-2}{n-2} - \binom{N-2}{n-2} q_k \right] =$$
$$\frac{1}{\binom{N-1}{n-1}} \left[\left(\binom{N-1}{n-1} - \binom{N-2}{n-2} \right) q_k + \binom{N-2}{n-2} \right] =$$
$$\frac{1}{\binom{N-1}{n-1}} \left[\binom{N-2}{n-1} q_k + \binom{N-2}{n-2} \right] =$$
$$\frac{N-n}{N-1} q_k + \frac{n-1}{N-1} = \frac{1}{N-1} \left[(N-n) q_k + (n-1) \right]$$

The second order inclusion probability of the elements k and l is the sum of the probabilities of the samples that include simultaneously both elements, i.e.

$$\pi_{kl} = \sum_{s \ni k,l} p(s) = \sum_{s \ni k,l} \frac{1}{\binom{N-1}{n-1}} \sum_{s} q_i = \frac{1}{\binom{N-1}{n-1}} \sum_{s \ni k,l} \sum_{s} q_i$$

Again, using the indicators we have

$$\pi_{kl} = \frac{1}{\binom{N-1}{n-1}} \sum_{s \ni k,l} \sum_{s} q_i = \frac{1}{\binom{N-1}{n-1}} \sum_{s \ni k,l} \sum_{U} q_i I_i = \frac{1}{\binom{N-1}{n-1}} \sum_{s \ni k,l} \left[q_k + q_l + \sum_{U^{(k,l)}} q_i I_i \right]$$

where $U^{(k,l)}$ is the set of all elements in U except the elements k and l. The last expression is obtained by noting that the elements k and l are included in all the terms in the sum. This expression can be rewritten as

$$\pi_{kl} = \frac{1}{\binom{N-1}{n-1}} \sum_{s \ni k,l} \left[q_k + q_l + \sum_{U^{(k,l)}} q_i I_i \right] = \frac{1}{\binom{N-1}{n-1}} \left[\sum_{s \ni k,l} (q_k + q_l) + \sum_{s \ni k,l} \sum_{U^{(k,l)}} q_i I_i \right]$$

But, there are $\binom{N-2}{n-2}$ samples that include both elements, k and l. And, of those samples, $\binom{N-3}{n-3}$ include each of the remaining elements, so

$$\pi_{kl} = \frac{1}{\binom{N-1}{n-1}} \left[\sum_{s \ni k,l} (q_k + q_l) + \sum_{s \ni k,l} \sum_{U^{(k,l)}} q_i I_i \right] =$$
$$\frac{1}{\binom{N-1}{n-1}} \left[\binom{N-2}{n-2}(q_k + q_l) + \binom{N-3}{n-3} \sum_{U^{(k,l)}} q_i \right] =$$
$$\frac{1}{\binom{N-1}{n-1}} \left[\binom{N-2}{n-2}(q_k + q_l) + \binom{N-3}{n-3}(1 - (q_k + q_l)) \right] =$$
$$\frac{1}{\binom{N-1}{n-1}} \left[\binom{N-2}{n-2}(q_k + q_l) + \binom{N-3}{n-3} - \binom{N-3}{n-3}(q_k + q_l) \right] =$$
$$\frac{1}{\binom{N-1}{n-1}} \left[\left(\binom{N-2}{n-2} - \binom{N-3}{n-3} \right)(q_k + q_l) + \binom{N-3}{n-3} \right] =$$
$$\frac{1}{\binom{N-1}{n-1}} \left[\binom{N-3}{n-2}(q_k + q_l) + \binom{N-3}{n-3} \right] =$$
$$\frac{(N-n)(n-1)}{(N-1)(N-2)}(q_k + q_l) + \frac{(n-1)(n-2)}{(N-1)(N-2)} =$$
$$\frac{n-1}{(N-1)(N-2)} \left[(N-n)(q_k + q_l) + (n-2) \right]$$

In general, the r-th order inclusion probability of the elements $k_1, k_2, \cdots, k_r$ is the sum of the probabilities of the samples that include simultaneously the r elements, i.e.

$$\pi_{k_1 k_2 \cdots k_r} = \sum_{s \ni k_1, k_2, \cdots, k_r} p(s) = \sum_{s \ni k_1, k_2, \cdots, k_r} \frac{1}{\binom{N-1}{n-1}} \sum_s q_i = \frac{1}{\binom{N-1}{n-1}} \sum_{s \ni k_1, k_2, \cdots, k_r} \sum_s q_i$$

Again, using the indicators we have

$$\pi_{k_1 k_2 \cdots k_r} = \frac{1}{\binom{N-1}{n-1}} \sum_{s \ni k_1, k_2, \cdots, k_r} \sum_s q_i = \frac{1}{\binom{N-1}{n-1}} \sum_{s \ni k_1, k_2, \cdots, k_r} \sum_U q_i I_i =$$
$$\frac{1}{\binom{N-1}{n-1}} \sum_{s \ni k_1, k_2, \cdots, k_r} \left[\sum_{i=1}^{r} q_{k_i} + \sum_{U^{(r)}} q_i I_i \right]$$

where $U^{(r)}$ is the set of all elements in U except the elements $l_1, l_2, \cdots, l_r$. The last expression is obtained by noting that the elements $l_1, l_2, \cdots, l_r$ are included in all the terms in the sum. Note that this notation is slightly different than that used in the previous cases. This expression can be rewritten as

$$\pi_{k_1 k_2 \cdots k_r} = \frac{1}{\binom{N-1}{n-1}} \sum_{s \ni k_1, k_2, \cdots, k_r} \left[\sum_{i=1}^{r} q_{k_i} + \sum_{U^{(r)}} q_i I_i \right] =$$
$$\frac{1}{\binom{N-1}{n-1}} \left[\sum_{s \ni k_1, k_2, \cdots, k_r} \sum_{i=1}^{r} q_{k_i} + \sum_{s \ni k_1, k_2, \cdots, k_r} \sum_{U^{(r)}} q_i I_i \right]$$

But, there are $\binom{N-r}{n-r}$ samples that include simultaneously the elements, $k_1, k_2, \cdots, k_r$. And, of those samples, $\binom{N-r-1}{n-r-1}$ include each of the remaining elements, so

$$\pi_{k_1 k_2 \cdots k_r} = \frac{1}{\binom{N-1}{n-1}} \left[\sum_{s \ni k_1, k_2, \cdots, k_r} \sum_{i=1}^{r} q_{k_i} + \sum_{s \ni k_1, k_2, \cdots, k_r} \sum_{U^{(r)}} q_i I_i \right] =$$

$$\frac{1}{\binom{N-1}{n-1}} \left[\binom{N-r}{n-r} \sum_{i=1}^{r} q_{k_i} + \binom{N-r-1}{n-r-1} \sum_{U^{(r)}} q_i \right] =$$

$$\frac{1}{\binom{N-1}{n-1}} \left[\binom{N-r}{n-r} \sum_{i=1}^{r} q_{k_i} + \binom{N-r-1}{n-r-1} \left(1 - \sum_{i=1}^{r} q_{k_i} \right) \right] =$$

$$\frac{1}{\binom{N-1}{n-1}} \left[\binom{N-r}{n-r} \sum_{i=1}^{r} q_{k_i} + \binom{N-r-1}{n-r-1} - \binom{N-r-1}{n-r-1} \sum_{i=1}^{r} q_{k_i} \right] =$$

$$\frac{1}{\binom{N-1}{n-1}} \left[\left(\binom{N-r}{n-r} - \binom{N-r-1}{n-r-1} \right) \sum_{i=1}^{r} q_{k_i} + \binom{N-r-1}{n-r-1} \right] =$$

$$\frac{1}{\binom{N-1}{n-1}} \left[\binom{N-r-1}{n-r} \sum_{i=1}^{r} q_{k_i} + \binom{N-r-1}{n-r-1} \right] =$$

$$(N-n) \frac{\prod_{i=1}^{r-1} (n-i)}{\prod_{i=1}^{r} (N-i)} \sum_{i=1}^{r} q_{k_i} + \frac{\prod_{i=1}^{r} (n-i)}{\prod_{i=1}^{r} (N-i)} = \frac{\prod_{i=1}^{r-1} (n-i)}{\prod_{i=1}^{r} (N-i)} \left[(N-n) \sum_{i=1}^{r} q_{k_i} + (n-r) \right]$$

□

Proof of Result 3. Let A be the event "element k is included in the sample", and B be the event "elements $l_1, l_2, \cdots, l_r$ are included in the sample ". The conditional probability of A given B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

But $P(B) = \pi_{l_1 l_2 \cdots l_r}$ and $P(A \cap B) = \pi_{l_1 l_2 \cdots l_r k}$, so

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\pi_{l_1 l_2 \cdots l_r k}}{\pi_{l_1 l_2 \cdots l_r}} = \frac{\frac{\prod_{i=1}^{r} (n-i)}{\prod_{i=1}^{r+1} (N-i)} \left[(N-n) \left(\sum_{i=1}^{r} p_{l_i} + p_k \right) + (n-r-1) \right]}{\frac{\prod_{i=1}^{r-1} (n-i)}{\prod_{i=1}^{r} (N-i)} \left[(N-n) \sum_{i=1}^{r} p_{l_i} + (n-r) \right]}$$

$$\frac{n-r}{N-r-1} \frac{\left[(N-n) \left(\sum_{i=1}^{r} p_{l_i} + p_k \right) + (n-r-1) \right]}{\left[(N-n) \sum_{i=1}^{r} p_{l_i} + (n-r) \right]}$$

□

Proof of Result 4. We need to prove that

$$E_q \left(\hat{t}_q \right) = t_y$$

By definition

$$
\begin{aligned}
E_q\left(\hat{t}_q\right) = \sum_{\Omega_q} p(s)\hat{t}_q = & \qquad \text{(by definition of expected value)} \\
\sum_{\Omega_q} \frac{1}{\binom{N-1}{n-1}} \sum_s q_k \frac{\sum_s y_k}{\sum_s q_k} = & \qquad \text{(using the design and estimator)} \\
\frac{1}{\binom{N-1}{n-1}} \sum_{\Omega_q} \sum_s y_k = & \\
\frac{1}{\binom{N-1}{n-1}} \sum_{\Omega_q} \sum_U y_k I_k = & \qquad \text{(using the } I_k \text{ defined in the proof of Result 1)} \\
\frac{1}{\binom{N-1}{n-1}} \sum_U \sum_{\Omega_q} y_k I_k = & \qquad \text{(interchanging summations)} \\
\frac{1}{\binom{N-1}{n-1}} \sum_U \binom{N-1}{n-1} y_k = & \qquad \text{(each element is included in } \binom{N-1}{n-1} \text{ samples)} \\
\sum_U y_k = t_y &
\end{aligned}
$$

□

Proof of Result 5. Let $Y = \sum_s y_k$ and $Q = \sum_s q_k$. First, note that

$$Cov(Y,Q) = E(XY) - E(X)E(Y) \qquad V(Y) = E(Y^2) \quad \text{and} \quad V(Q) = E(Q^2)$$

so (36) can be rewritten as

$$
\begin{aligned}
V\left[\frac{Y}{Q}\right] \approx \frac{E^2(Y)}{E^2(Q)}\left[\frac{V(Y)}{E^2(Y)} - 2\frac{Cov(Y,Q)}{E(Y)E(Q)} + \frac{V(Q)}{E^2(Q)}\right] = & \\
\frac{1}{E^4(Q)}\left[E\left(Y^2\right)E^2(Q) - 2E(YQ)E(Y)E(Q) + E(Q^2)E^2(Y)\right] & \quad (48)
\end{aligned}
$$

The goal is to obtain expressions for $E(Y)$, $E(Q)$, $E\left(Y^2\right)$, $E\left(Q^2\right)$ and $E(Y,Q)$, and then use these expressions into (48).

First, we will calculate $E(Y)$,

$$
\begin{aligned}
E(Y) = E\left(\sum_s y_k\right) = \sum_{\Omega_q} p(s) \sum_s y_k = & \quad \text{(definition of expected value)} \\
\sum_{\Omega_q} \frac{1}{\binom{N-1}{n-1}} \sum_s q_k \sum_s y_k = \frac{1}{\binom{N-1}{n-1}} \sum_{\Omega_q} \sum_s q_k \sum_s y_k = & \quad \text{(using the design)} \\
\frac{1}{\binom{N-1}{n-1}} \sum_{\Omega_q} \sum_s \sum_s q_k y_l = \frac{1}{\binom{N-1}{n-1}} \sum_{\Omega_q} \sum_U \sum_U q_k y_l I_k I_l = & \quad \text{(using } I_k \text{ as above)} \\
\frac{1}{\binom{N-1}{n-1}} \sum_U \sum_U \sum_{\Omega_q} q_k y_l I_k I_l & \quad \text{(interchanging summations)}
\end{aligned}
$$

Noting that each element appears $\binom{N-1}{n-1}$ times in Ω_q, and each pair k, l appears $\binom{N-2}{n-2}$ times, the last term can be rewritten

$$\frac{1}{\binom{N-1}{n-1}} \sum_U \sum_U \sum_{\Omega_q} q_k y_l I_k I_l =$$

$$\frac{1}{\binom{N-1}{n-1}} \left[\binom{N-1}{n-1} \sum_U q_k y_k + \binom{N-2}{n-2} \sum\sum_{k \neq l} q_k y_l \right] = \quad \text{(as mentioned above)}$$

$$\frac{1}{\binom{N-1}{n-1}} \left[\left(\binom{N-1}{n-1} \sum_U q_k y_k - \binom{N-2}{n-2} \sum_U q_k y_k \right) + \right.$$

$$\left. \left(\binom{N-2}{n-2} \sum_U q_k y_k + \binom{N-2}{n-2} \sum\sum_{k \neq l} q_k y_l \right) \right] = \quad \text{(adding and substracting)}$$

$$\frac{1}{\binom{N-1}{n-1}} \left[\binom{N-1}{n-1} \frac{N-n}{N-1} \sum_U q_k y_k + \binom{N-2}{n-2} \sum_U \sum_U q_k y_l \right] =$$

$$\frac{1}{\binom{N-1}{n-1}} \left[\binom{N-1}{n-1} \frac{N-n}{N-1} \sum_U q_k y_k + \binom{N-2}{n-2} \sum_U y_k \right] = \quad \text{(noting that } t_q = 1\text{)}$$

$$\frac{N-n}{N-1} \sum_U q_k y_k + \frac{n-1}{N-1} \sum_U y_k =$$

$$\frac{1}{N-1} \left[(N-n) t_{qy} + (n-1) t_y \right]$$

Now, in order to calculate $E(Q)$ just replace y by q in the procedure above, so

$$E(Q) = E\left(\sum_s q_k \right) = \frac{1}{N-1} \left[(N-n) t_{q^2} + (n-1) \right]$$

For the remaining terms of interest, consider arbitrary X and Z defined as $X = \sum_s x_k$ and $Z = \sum_s z_k$. We need to calculate $E(XZ)$:

$$E(XZ) = E\left(\sum_s x_k \sum_s z_k \right) = \sum_{\Omega_1} p(s) \sum_s x_k \sum_s z_k = \quad \text{(definition of expectation)}$$

$$\sum_{\Omega_q} \frac{1}{\binom{N-1}{n-1}} \sum_s q_k \sum_s x_k \sum_s z_k = \quad \text{(using the design)}$$

$$\frac{1}{\binom{N-1}{n-1}} \sum_{\Omega_q} \sum_s q_k \sum_s x_k \sum_s z_k = \quad \text{(factorizing)}$$

$$\frac{1}{\binom{N-1}{n-1}} \sum_{\Omega_q} \sum_s \sum_s \sum_s q_k x_l z_m =$$

$$\frac{1}{\binom{N-1}{n-1}} \sum_{\Omega_q} \sum_U \sum_U \sum_U q_k x_l z_m I_k I_l I_m = \quad \text{(using the } I_k\text{)}$$

$$\frac{1}{\binom{N-1}{n-1}} \sum_U \sum_U \sum_U \sum_{\Omega_q} q_k x_l z_m I_k I_l I_m \quad \text{(interchanging summations)}$$

Noting that each element appears $\binom{N-1}{n-1}$ times in Ω_q, each pair k,l appears $\binom{N-2}{n-2}$ and each triplet k,l,m appears $\binom{N-3}{n-3}$ times, the last term can be rewritten

$$\frac{1}{\binom{N-1}{n-1}}\sum_U\sum_U\sum_U\sum_{\Omega_q} q_k x_l z_m I_k I_l I_m =$$

$$\frac{1}{\binom{N-1}{n-1}}\left[\binom{N-1}{n-1}\sum_U q_k x_k z_k + \binom{N-2}{n-2}\left(\sum\sum_{k\neq l} q_k x_k z_l + \sum\sum_{k\neq l} q_k x_l z_k + \sum\sum_{k\neq l} q_l x_k z_k\right) + \binom{N-3}{n-3}\sum\sum\sum_{k\neq l\neq m} q_k x_l z_m\right] =$$

$$\frac{1}{\binom{N-1}{n-1}}\left[\binom{N-1}{n-1}\sum_U q_k x_k z_k + \binom{N-2}{n-2}\left(\sum_U\sum_U q_k x_k z_l + \sum_U q_k x_l z_k + \sum_U q_l x_k z_k - 3\sum_U q_k x_k y_k\right) + \binom{N-3}{n-3}\left(\sum_U\sum_U\sum_U q_k x_l z_m - \sum_U\sum_U q_k x_k z_l - \sum_U\sum_U q_k x_l z_k - \sum_U\sum_U q_l x_k z_k + 2\sum_U q_k x_k z_k\right)\right]$$

$$\frac{1}{\binom{N-1}{n-1}}\left[\left(\binom{N-1}{n-1} - 3\binom{N-2}{n-2} + 2\binom{N-3}{n-3}\right)\sum_U q_k x_k z_k + \left(\binom{N-2}{n-2} - \binom{N-3}{n-3}\right)\left(\sum_U\sum_U q_k x_k z_l + \sum_U\sum_U q_k x_l z_k + \sum_U\sum_U q_l x_k z_k\right) + \binom{N-3}{n-3}\sum_U\sum_U\sum_U q_k x_l z_m\right] =$$

$$\frac{1}{\binom{N-1}{n-1}}\left[\binom{N-1}{n-1}\frac{(N-n)(N-2n)}{(N-1)(N-2)}\sum_U q_k x_k z_k + \binom{N-1}{n-1}\frac{(N-n)(n-1)}{(N-1)(N-2)}\left(\sum_U\sum_U q_k x_k z_l + \sum_U\sum_U q_k x_l z_k + \sum_U\sum_U q_l x_k z_k\right) + \binom{N-1}{n-1}\frac{(n-1)(n-2)}{(N-1)(N-2)}\sum_U\sum_U\sum_U q_k x_l z_m\right] =$$

$$\frac{1}{(N-1)(N-2)}\left[\sum_U q_k x_k z_k (N-n)(N-2n) + \left(\sum_U\sum_U q_k x_k z_l + \sum_U\sum_U q_k x_l z_k + \sum_U\sum_U q_l x_k z_k\right)(N-n)(n-1) + \sum_U\sum_U\sum_U q_k x_l z_m (n-1)(n-2)\right] =$$

$$\frac{1}{(N-1)(N-2)}\left[t_{qxz}(N-n)(N-2n) + \left(t_{qx}t_z + t_{qz}t_x + t_{xz}\right)(N-n)(n-1) + t_x t_z (n-1)(n-2)\right]$$

Now, replace X and Z by $Y = \sum_s y_k$ and $Q = \sum_s q_k$, respectively, we have

$$E(YQ) = E\left(\sum_s y_k \sum_s q_k\right) =$$
$$\frac{1}{(N-1)(N-2)}\left[t_{yq^2}(N-n)(N-2n) + \left(2t_{qy} + t_{q^2}t_y\right)(N-n)(n-1) + t_y(n-1)(n-2)\right]$$

In order to calculate $E\left(Y^2\right)$, replace both X and Z by Y in (A), we have

$$E\left(Y^2\right) = E\left(\sum_s y_k \sum_s y_k\right) =$$
$$\frac{1}{(N-1)(N-2)}\left[t_{qy^2}(N-n)(N-2n) + \left(2t_{qy}t_y + t_{y^2}\right)(N-n)(n-1) + t_y^2(n-1)(n-2)\right]$$

And $E\left(Q^2\right)$ is

$$E\left(Q^2\right) = E\left(\sum_s q_k \sum_s q_k\right) =$$
$$= \frac{1}{(N-1)(N-2)}\left[t_{q^3}(N-n)(N-2n) + 3t_{q^2}(N-n)(n-1) + (n-1)(n-2)\right]$$

Finally, using the five expressions obtained into (48), we obtain the expression in (37). □

B Sampling selection method: Program in R

The program has two inputs: `q`: the vector of q-values, of length N; `n`: the sample size. The output is a vector of length n indicating the selected elements.

```
qsample<- function(q,n) {
N<- length(q)
ido<- 1:N
id<- ido
qno<- q
qsi<- numeric(N)
elige<- numeric(n)

for (i in 1:n) {
alea<- runif(1,0,n+1-i)
pinc<-
   (n-i+1)*((N-n)*(sum(qsi)+qno)+(n-i))/((N-i)*((N-n)*sum(qsi)+n-i+1))
psup<- cumsum(pinc)
pinf<- c(0,psup[-length(pinc)])
elegido<- (pinf<=alea&alea<psup)
elige[i]<- id[elegido>0]
qsi<- q[elige]
qno<- q[setdiff(ido,elige)]
id<- ido[setdiff(ido,elige)]
}
return(elige)
}
```

C Simulated Variance Increase

In section 5.5 tables of the SVI for $y^{(1)}$, $y^{(2)}$, $y^{(3)}$, $y^{(7)}$, $y^{(8)}$ and $y^{(12)}$ were presented. Tables 17 to 23 correspond to the remaining variables.

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
1	60.73	60.73	73.18	307.44	0.98
4	22.30	0.12	148.70	56.99	1.01
9	15.83	0.23	194.90	319.25	1.01
16	12.63	0.42	91.11	555.00	0.96
25	12.97	0.59	1018.59	85.13	0.98
36	12.82	0.67	224.47	98.38	0.99
49	13.00	0.76	305.69	108.29	0.98
64	12.69	0.81	115.25	105.75	0.98
81	12.39	0.84	171.30	211.31	1.01
100	12.77	0.87	922.85	1313.54	1.01
121	12.42	0.89	242.76	252.62	1.00
144	13.09	0.96	1400.47	111.00	0.96
169	12.76	0.94	842.74	4049.04	1.01
196	12.42	0.94	953.82	113.31	0.99
225	12.43	0.95	198.51	125.53	0.99
256	13.07	1.00	168.17	181.03	1.00
289	12.36	0.95	351.80	285.74	1.01
324	12.50	0.97	259.93	608.55	1.02
361	12.35	0.96	6478.28	150.27	0.98
400	12.23	0.94	1195.32	187.91	1.03

Table 17: (Simulated) VI for five sampling strategies, study variable: $y^{(4)}$

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
1	8640.97	8640.97	382.34	347.85	0.97
4	115.36	1.51	7548.02	1397.80	0.99
9	92.85	0.32	1953.41	686.19	0.99
16	92.13	0.32	793.49	896.94	1.01
25	90.24	0.44	1285.93	1548.01	0.98
36	84.37	0.55	888.83	11528.47	0.99
49	87.53	0.66	781.61	3970.37	0.98
64	82.86	0.70	532817.20	744.72	1.02
81	84.44	0.76	1208.31	4127.70	0.98
100	86.23	0.79	12674.49	6574.17	1.03
121	82.77	0.81	13571.54	1814.09	1.01
144	83.63	0.86	1569.03	1813.41	1.01
169	82.11	0.87	4906.47	3746.46	0.98
196	82.37	0.91	1798.13	1740.36	0.98
225	82.17	0.91	63562.39	2039.07	1.00
256	83.64	0.91	1672.94	5076.80	1.02
289	85.44	0.95	2622.05	1660.41	0.98
324	81.80	0.92	3209.37	1536.59	1.01
361	83.86	0.95	7056.21	2506.07	0.99
400	83.06	0.95	2544.41	4250.31	1.00

Table 18: (Simulated) VI for five sampling strategies, study variable: $y^{(5)}$

D Simulated Coverage

In section 5.6 tables of the simulated coverage for $y^{(1)}$, $y^{(2)}$, $y^{(11)}$ and $y^{(12)}$ were presented. Tables for the remaining variables are shown in this section.

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
1	1112.37	1112.37	942.43	614.50	0.97
4	332.69	5.06	2307.30	5303.14	1.01
9	232.70	1.39	2142.29	2109.70	1.03
16	223.24	0.80	2235.76	1970.50	1.00
25	205.44	0.70	1319.84	1594.76	1.03
36	200.97	0.72	2421.73	1711.42	0.99
49	196.90	0.75	6173.98	1606.43	0.98
64	198.43	0.81	32074.35	1845.36	1.00
81	204.33	0.82	2462.10	3698.81	1.03
100	195.32	0.86	9540.40	4101.76	1.01
121	200.71	0.91	8731.39	7249.44	0.98
144	199.85	0.89	8677.83	20386.54	1.01
169	204.12	0.93	6594.75	13646.80	0.98
196	193.37	0.91	13310.64	12312.58	1.01
225	201.21	0.91	141152.88	3815.38	0.99
256	205.15	0.97	3152.13	2597.78	1.01
289	194.63	0.94	3348.19	193533.74	1.02
324	193.18	0.94	3544.63	105841.89	1.00
361	193.88	0.93	2717.29	3815.76	1.01
400	191.57	0.99	9225.12	2183.99	0.99

Table 19: (Simulated) VI for five sampling strategies, study variable: $y^{(6)}$

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
1	25.21	25.21	7.37	6.34	0.99
4	3.90	0.81	9.29	7.54	1.03
9	3.26	0.87	12.04	11.15	1.03
16	2.78	0.88	10.16	5.45	1.00
25	2.74	0.96	8.83	36.93	1.01
36	2.78	0.93	11.77	11.63	1.00
49	2.69	1.00	10.99	21.74	0.96
64	2.70	1.00	22.41	9.35	1.01
81	2.62	0.96	11.89	10.92	1.01
100	2.67	0.98	11.93	13.24	0.99
121	2.62	0.97	12.75	293.73	0.97
144	2.76	1.00	16.22	14.39	1.00
169	2.62	0.98	19.23	13.08	1.01
196	2.70	0.96	47.81	16.85	1.01
225	2.75	1.01	23.62	13.03	1.00
256	2.60	1.02	16.95	79.92	1.00
289	2.69	1.04	13.78	13.69	1.00
324	2.64	0.98	14.03	12.49	1.01
361	2.69	1.02	139.47	12.76	0.98
400	2.68	0.96	51.88	43.67	0.99

Table 20: (Simulated) VI for five sampling strategies, study variable: $y^{(9)}$

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
1	0.05	0.05	0.06	0.06	1.34
4	0.17	0.25	0.05	0.03	0.86
9	0.25	0.40	0.05	0.02	1.14
16	0.32	0.53	0.05	0.02	1.02
25	0.38	0.65	0.05	0.01	1.04
36	0.43	0.78	0.05	0.01	1.02
49	0.46	0.84	0.04	0.01	0.98
64	0.45	0.81	0.04	0.00	0.98
81	0.46	0.86	0.04	0.00	0.97
100	0.46	0.85	0.03	0.00	1.01
121	0.47	0.88	0.03	0.00	1.01
144	0.47	0.91	0.02	0.00	1.02
169	0.49	0.91	0.02	0.00	1.00
196	0.49	0.91	0.01	0.00	1.01
225	0.50	0.95	0.01	0.00	0.98
256	0.50	0.96	0.01	0.00	0.97
289	0.51	0.96	0.01	0.00	1.01
324	0.50	0.96	0.01	0.00	1.02
361	0.50	0.96	0.00	0.00	0.96
400	0.50	0.95	0.00	0.00	1.01

Table 21: (Simulated) VI for five sampling strategies, study variable: $y^{(10)}$

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
1	0.45	0.45	0.27	0.30	0.78
4	0.31	0.28	0.43	0.24	0.90
9	0.29	0.41	0.35	0.26	0.87
16	0.34	0.57	0.32	0.25	1.03
25	0.39	0.76	0.35	0.23	1.09
36	0.37	0.75	0.35	0.27	0.91
49	0.38	0.83	0.34	0.23	1.04
64	0.40	0.83	0.34	0.23	1.00
81	0.42	0.92	0.33	0.26	1.01
100	0.43	0.92	0.38	0.22	1.03
121	0.41	0.92	0.36	0.24	1.00
144	0.41	0.90	0.35	0.21	1.00
169	0.42	0.93	0.35	0.21	1.03
196	0.43	0.97	0.36	0.36	0.97
225	0.42	0.96	0.36	0.27	1.00
256	0.42	0.95	0.40	0.18	1.02
289	0.43	1.00	0.36	0.20	0.97
324	0.42	0.97	0.35	0.20	1.01
361	0.43	0.94	0.34	0.25	1.02
400	0.42	0.96	0.36	0.17	0.98

Table 22: (Simulated) VI for five sampling strategies, study variable: $y^{(11)}$

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
1	888.59	888.59	1435.73	24766.57	0.99
4	357.70	7.03	1130.84	3450.19	1.01
9	234.61	2.55	11960.99	1578.96	1.00
16	244.94	1.64	1714.32	1474.41	1.00
25	215.57	1.28	3238.48	14420.78	0.99
36	227.71	1.12	1954.12	9307.28	0.99
49	216.14	1.02	2060.93	5844.05	0.99
64	216.17	0.99	121498.28	2870.67	1.02
81	209.36	1.02	2641.35	3127.60	0.99
100	218.46	1.02	4637.45	3856.83	1.01
121	207.60	1.01	46357.27	2821.42	0.99
144	215.63	1.02	2985.62	2058.12	1.03
169	209.00	1.03	9123.58	4722.49	0.97
196	216.31	1.01	28081.68	4625.51	1.00
225	214.68	1.02	6527.30	16370.75	1.02
256	216.70	1.01	2831.55	5876.95	1.00
289	215.50	1.01	14069.09	349831.62	1.00
324	210.65	0.99	9597.24	4161.57	1.00
361	207.79	0.99	204208.23	4386.30	1.01
400	210.62	0.97	20890.35	77614.71	1.00

Table 23: (Simulated) VI for five sampling strategies, study variable: $y^{(13)}$

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
4	0.69	0.91	0.75	0.99	0.90
9	0.78	0.92	0.75	0.99	0.91
16	0.85	0.92	0.77	0.99	0.93
25	0.88	0.93	0.78	1.00	0.94
36	0.90	0.94	0.80	0.99	0.94
49	0.91	0.95	0.80	0.98	0.94
64	0.92	0.94	0.81	0.99	0.94
81	0.92	0.95	0.82	1.00	0.94
100	0.93	0.94	0.82	0.97	0.94
121	0.93	0.94	0.82	1.00	0.95
144	0.94	0.94	0.84	0.97	0.95
169	0.94	0.95	0.84	0.96	0.94
196	0.94	0.95	0.84	0.99	0.95
225	0.94	0.95	0.84	0.98	0.95
256	0.95	0.94	0.85	0.98	0.94
289	0.94	0.95	0.85	0.97	0.95
324	0.94	0.95	0.84	1.00	0.95
361	0.94	0.95	0.85	0.98	0.95
400	0.94	0.94	0.86	0.97	0.95

Table 24: (Simulated) coverage for five sampling strategies. Study variable: $y^{(3)}$

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
4	0.71	0.89	0.67	0.99	0.80
9	0.79	0.87	0.64	0.99	0.84
16	0.84	0.89	0.66	1.00	0.86
25	0.87	0.90	0.67	0.98	0.87
36	0.90	0.90	0.69	0.98	0.89
49	0.91	0.92	0.68	0.98	0.91
64	0.92	0.92	0.71	0.98	0.91
81	0.93	0.93	0.71	0.98	0.92
100	0.93	0.93	0.72	0.99	0.92
121	0.93	0.93	0.72	0.98	0.93
144	0.93	0.92	0.71	0.97	0.93
169	0.94	0.93	0.73	1.00	0.93
196	0.93	0.94	0.73	0.97	0.94
225	0.93	0.93	0.72	0.97	0.94
256	0.94	0.93	0.73	0.98	0.94
289	0.94	0.94	0.73	0.98	0.94
324	0.94	0.94	0.73	0.99	0.93
361	0.94	0.94	0.73	0.97	0.94
400	0.95	0.94	0.72	0.97	0.94

Table 25: (Simulated) coverage for five sampling strategies. Study variable: $y^{(4)}$

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
4	0.71	0.94	0.66	1.00	0.85
9	0.80	0.94	0.64	0.99	0.86
16	0.83	0.95	0.67	0.99	0.89
25	0.87	0.93	0.68	0.99	0.91
36	0.89	0.93	0.69	1.00	0.92
49	0.91	0.93	0.69	1.00	0.92
64	0.92	0.94	0.71	0.98	0.93
81	0.92	0.93	0.70	0.99	0.93
100	0.93	0.93	0.71	1.00	0.93
121	0.94	0.94	0.72	0.98	0.94
144	0.93	0.94	0.72	0.98	0.94
169	0.94	0.94	0.72	0.99	0.94
196	0.94	0.94	0.74	0.98	0.94
225	0.94	0.94	0.72	0.98	0.94
256	0.94	0.94	0.72	0.99	0.94
289	0.94	0.94	0.74	0.98	0.95
324	0.94	0.95	0.74	0.98	0.94
361	0.94	0.94	0.73	0.98	0.95
400	0.95	0.95	0.74	0.99	0.94

Table 26: (Simulated) coverage for five sampling strategies. Study variable: $y^{(5)}$

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
4	0.70	0.95	0.65	1.00	0.94
9	0.78	0.95	0.65	0.99	0.95
16	0.83	0.95	0.66	0.99	0.95
25	0.87	0.95	0.67	0.98	0.94
36	0.90	0.95	0.68	0.98	0.95
49	0.92	0.95	0.70	0.98	0.94
64	0.92	0.95	0.71	0.98	0.95
81	0.93	0.95	0.70	0.98	0.95
100	0.93	0.95	0.71	0.99	0.95
121	0.93	0.94	0.72	0.99	0.95
144	0.93	0.95	0.72	1.00	0.95
169	0.94	0.95	0.72	0.99	0.95
196	0.95	0.95	0.72	0.99	0.95
225	0.94	0.95	0.74	0.98	0.95
256	0.93	0.94	0.72	0.98	0.95
289	0.94	0.95	0.73	1.00	0.94
324	0.95	0.95	0.73	1.00	0.95
361	0.94	0.95	0.73	0.98	0.95
400	0.95	0.94	0.74	0.97	0.95

Table 27: (Simulated) coverage for five sampling strategies. Study variable: $y^{(6)}$

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
4	0.69	0.95	0.92	0.99	0.90
9	0.78	0.94	0.91	0.98	0.90
16	0.83	0.93	0.91	0.97	0.92
25	0.87	0.93	0.91	0.98	0.92
36	0.89	0.94	0.92	0.99	0.93
49	0.90	0.94	0.92	0.93	0.93
64	0.91	0.94	0.92	0.98	0.94
81	0.92	0.94	0.93	0.97	0.94
100	0.92	0.94	0.93	0.96	0.94
121	0.93	0.94	0.93	0.94	0.94
144	0.93	0.94	0.92	0.95	0.95
169	0.93	0.94	0.92	0.94	0.94
196	0.93	0.95	0.93	0.98	0.95
225	0.93	0.95	0.93	0.94	0.94
256	0.93	0.95	0.92	0.97	0.95
289	0.94	0.95	0.93	0.96	0.94
324	0.94	0.95	0.93	0.98	0.95
361	0.93	0.95	0.93	0.96	0.95
400	0.94	0.95	0.93	0.98	0.95

Table 28: (Simulated) coverage for five sampling strategies. Study variable: $y^{(7)}$

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
4	0.69	0.90	0.87	0.99	0.86
9	0.79	0.91	0.88	0.99	0.89
16	0.84	0.92	0.88	0.98	0.91
25	0.88	0.92	0.89	0.96	0.93
36	0.90	0.93	0.90	0.97	0.92
49	0.92	0.93	0.91	0.96	0.93
64	0.92	0.94	0.91	0.96	0.94
81	0.93	0.94	0.92	0.97	0.94
100	0.93	0.94	0.92	0.97	0.95
121	0.93	0.94	0.92	0.98	0.94
144	0.94	0.94	0.92	0.99	0.95
169	0.94	0.94	0.92	0.98	0.95
196	0.94	0.94	0.92	0.97	0.95
225	0.94	0.95	0.93	0.99	0.94
256	0.94	0.94	0.92	0.97	0.95
289	0.94	0.94	0.93	0.97	0.95
324	0.95	0.94	0.92	0.97	0.96
361	0.94	0.95	0.93	0.97	0.95
400	0.94	0.94	0.93	0.96	0.94

Table 29: (Simulated) coverage for five sampling strategies. Study variable: $y^{(8)}$

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
4	0.71	0.91	0.75	0.99	0.90
9	0.78	0.93	0.75	0.99	0.91
16	0.84	0.93	0.76	0.97	0.93
25	0.88	0.93	0.76	0.99	0.93
36	0.89	0.94	0.79	0.98	0.94
49	0.91	0.94	0.79	0.99	0.95
64	0.92	0.94	0.80	0.97	0.95
81	0.93	0.95	0.81	0.98	0.94
100	0.93	0.94	0.81	0.97	0.95
121	0.93	0.95	0.83	1.00	0.95
144	0.93	0.95	0.82	0.97	0.94
169	0.94	0.95	0.83	0.98	0.95
196	0.94	0.95	0.84	0.98	0.95
225	0.94	0.95	0.82	0.98	0.95
256	0.94	0.94	0.84	1.00	0.95
289	0.94	0.94	0.83	0.97	0.94
324	0.94	0.95	0.84	0.98	0.94
361	0.94	0.95	0.84	0.98	0.95
400	0.94	0.96	0.84	0.99	0.95

Table 30: (Simulated) coverage for five sampling strategies. Study variable: $y^{(9)}$

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
4	0.44	0.74	0.89	0.99	0.55
9	0.54	0.73	0.91	0.96	0.63
16	0.62	0.75	0.91	1.00	0.69
25	0.70	0.77	0.92	0.99	0.74
36	0.75	0.79	0.94	0.95	0.76
49	0.78	0.81	0.94	1.00	0.79
64	0.79	0.83	0.94	0.98	0.81
81	0.81	0.84	0.95	1.00	0.82
100	0.82	0.85	0.95	0.98	0.84
121	0.82	0.85	0.95	0.94	0.85
144	0.83	0.86	0.95	0.97	0.85
169	0.84	0.87	0.95	0.97	0.86
196	0.84	0.87	0.95	0.97	0.86
225	0.85	0.87	0.95	0.97	0.88
256	0.85	0.87	0.95	0.98	0.88
289	0.86	0.89	0.94	0.95	0.88
324	0.86	0.89	0.95	0.95	0.88
361	0.86	0.89	0.95	0.96	0.89
400	0.87	0.89	0.95	0.99	0.89

Table 31: (Simulated) coverage for five sampling strategies. Study variable: $y^{(10)}$

n	($q$ps - q)	($q$ps - π)	(Pareto - $q\pi$)	(Syst. - π)	(SRS - π)
4	0.72	0.96	0.65	1.00	0.95
9	0.78	0.95	0.65	0.99	0.95
16	0.83	0.95	0.65	0.98	0.95
25	0.87	0.95	0.69	1.00	0.95
36	0.90	0.95	0.69	0.99	0.95
49	0.91	0.95	0.70	0.99	0.95
64	0.92	0.96	0.70	0.98	0.95
81	0.93	0.95	0.71	0.99	0.95
100	0.93	0.95	0.71	0.98	0.95
121	0.93	0.95	0.70	0.97	0.95
144	0.94	0.95	0.72	0.97	0.95
169	0.94	0.95	0.71	0.98	0.96
196	0.93	0.95	0.72	0.98	0.95
225	0.94	0.95	0.72	0.99	0.95
256	0.94	0.95	0.72	0.98	0.95
289	0.94	0.95	0.73	1.00	0.95
324	0.94	0.95	0.72	0.98	0.95
361	0.95	0.95	0.73	0.98	0.95
400	0.94	0.95	0.74	1.00	0.95

Table 32: (Simulated) coverage for five sampling strategies. Study variable: $y^{(13)}$