UNIVERSIDAD CARLOS III DE MADRID
working papers

Working Paper 03-52
Statistics and Econometrics Series 11
October 2003

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

OPTIMAL RANDOM SAMPLING DESIGNS IN RANDOM FIELD SAMPLING

José E. Rodríguez and Fernando Ávila*

Abstract

A Horvitz-Thompson predictor is proposed for spatial sampling when the characteristic of interest is modeled as a random field. Optimal sampling designs are deduced under this context. Fixed and variable sample size are considered.

Keywords: Horvitz-Thompson estimator; Spatial sampling; Optimal random sampling designs; Variable size samples.

*Rodríguez, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, C/ Madrid, 126, 28903 Getafe. Madrid, e-mail: jerodrig@est-econ.uc3m.es y School of Mathematics, University of Guanajuato, México, e-mail: elias@cimat.mx; Ávila, Centro de Investigación en Matemáticas, Guanajuato, México. This work was completed while the first author was a visiting professor at the Carlos III of Madrid University. Supported, in part, by CONCYTEG grant 03-16-K118-027 to José E. Rodríguez.

The Horvitz-Thompson Predictor in Random Field Sampling

José E. Rodríguez[1]

School of Mathematics, University of Guanajuato, Guanajuato, México

Fernando Ávila

Centro de Investigación en Matemáticas, Guanajuato, México

Abstract: A Horvitz-Thompson predictor is proposed for spatial sampling when the characteristic of interest is modeled as a random …eld. Optimal sampling designs are deduced under this context. Fixed and variable sample size are considered.

Keywords: Horvitz-Thompson estimator; spatial sampling; optimal random sampling designs; variable size samples.

1 Introduction

The Horvitz-Thompson type estimators are widely used in …nite population estimation when the estimation procedure is based on probability sampling. Cordy (1993) extends this type of estimator to populations distributed over spatial domains where the characteristic of interest is conceptualized as a deterministic continuous function de…ned on these domains. The Horvitz-Thompson estimator proposed in Cordy (1993) has been generally used on environmental sampling, see Stevens (1997). In the present work, this characteristic is mod-

[1] Correspondence to: José Elías Rodríguez, Apartado Postal 402, 36000 Guanajuato, Gto México, e-mail: elias@cimat.mx
This work was completed while the …rst author was a visiting professor at the Carlos III de Madrid University.
Supported, in part, by CONCYTEG grant 03-16-K118-027 to José E. Rodríguez.

eled by a random field and a Horvitz-Thompson predictor is proposed. Optimal sampling designs are deduced under this context.

When the region of interest is discretized to a finite grid of points, optimal sampling designs were established in Aldworth & Cressie (1999) among others. In the present work, the random sampling designs, including the optimal designs, are defined over all the points of the region of interest.

In order to introduce the Horvitz-Thompson predictor and their optimal sampling designs, first the spatial estimation is briefly reviewed. Secondly, The Horvitz-Thompson predictor is proposed and his related optimal sampling designs are deduced. Finally, this predictor is studied under variable size samples.

2 Spatial estimation

In this work, the population is a subset of $\mathbb{R}^d$, $U \subset \mathbb{R}^d$ such that $|U| > 0$ ($|\cdot|$ denotes volume under Lebesgue measure). The characteristic of interest of the population is represented by $z_U = \{z(x) : x \in U\}$, where $z(x) \in \mathbb{R}$, for all $x \in U$.

A finite sample is a set of points $\{x_1, \ldots, x_n\}$, such that each $d-$dimensional $x_k \in U$ and n is the fixed sample size.

A sampling design, of size n, is the joint distribution function G_n of a set of random variables $\{X_1, \ldots, X_n\}$, where each X_k is a $d-$dimensional random variable with support in U. The sample points $\{x_1, \ldots, x_n\}$ are possible realizations of these random variables. If G_n has density function g_n, then g_n is also named the sampling design. Additionally, if the random variables $\{X_1, \dots, X_n\}$ are independent and identically distributed with marginal distribution function G, then either G or its respective density function is called a random sampling design.

The population parameter to be estimated is the total $t := \int_{\mathcal{U}} z(x)\,dx$, provided that this integral exists.

The Horvitz-Thompson estimator of t is given by

$$t_{\pi_n} := \sum_{k=1} \frac{z(X_k)}{\pi_n(X_k)}, \tag{1}$$

where $\pi_n = \sum_{k=1}^{n} g_k$, provided that $\pi_n > 0$ on $\mathcal{U}$, and g_k is the marginal density function of X_k. The estimator (1) was proposed by Cordy (1993).

The estimator (1) has the following properties: a) it is unbiased; b) its variance is

$$\int_{\mathcal{U}} \frac{z^2(x)}{\pi_n(x)}\,dx + \int_{\mathcal{U}}\int_{\mathcal{U}} \frac{z(x)\,z(y)}{\pi_n(x)\,\pi_n(y)}\pi_n(x,y)\,dx\,dy - t^2, \tag{2}$$

where $\pi_n(\cdot,\cdot) = \sum_{j=1}^{n} \sum_{k \neq j} h_{jk}(\cdot,\cdot)$ and h_{jk} is the marginal bivariate density function of (X_j, X_k); c) if $\pi_n(\cdot,\cdot) > 0$ on $\mathcal{U} \times \mathcal{U}$, then an unbiased estimator of the variance in (2) is

$$\sum_{k=1} \frac{z^2(X_k)}{\pi_n(X_k)} + \sum_{j=1} \sum_{k \neq j} \frac{z(X_j)\,z(X_k)}{\pi_n(X_j)\,\pi_n(X_k)} - \sum_{j=1} \sum_{k \neq j} \frac{z(X_j)\,z(X_k)}{\pi_n(X_j, X_k)}.$$

Some restrictions over $z_{\mathcal{U}}$, π_n and $\pi_n(\cdot,\cdot)$ have been established in Cordy

(1993) in order that the estimator (1) has the above properties.

3 The Horvitz-Thompson predictor

Here the characteristic z_U is conceptualized as a realization of the second-order random field $Z_U = \{Z(x) : x \in U\}$, such that $\int_U E\left[Z^2(x)\right] < \infty$. Additionally, this random field is assumed to be continuous in quadratic mean.

Similarly, as in the previous section, the statistic of interest is the total value $T := \int_U Z(x)\,dx$.

In addition, the set of random variables Z_U and the set $\{X_1, \ldots, X_n\}$ are defined over the same probability space. Moreover, we assume that the field Z_U and the set $\{X_1, \ldots, X_n\}$ are stochastic independent, in particular the sampling design G_n will not depend on Z_U.

The above sampling scheme has two sources of randomness. One comes from our uncertainty about the particular values of the quantity of interest on each point of U. The other is generated by the sampling procedure. The first kind of randomness is modeled by the random field and the second one is described by the random sampling design. Furthermore, the random field is used for modeling the dependency among the observations in different sampling points. This random field will be also used for obtaining optimal sampling designs.

Using the form of the total T, a natural predictor is a linear homogeneous predictor, based on the design G_n, as

$$T_{\lambda_n} := \frac{1}{n} \sum_{k=1} \lambda_n (X_k) Z(X_k), \tag{3}$$

where $\lambda_n : U \to \mathbb{R}$ is a function of coefficients. The same predictor can be found in Schoenfelder & Cambanis (1982). The coefficients of this predictor do not require knowledge of the random field model and thus this predictor is nonparametric.

Schoenfelder & Cambanis (1982) obtained the necessary and sufficient conditions in order that the MSE of the predictor (3) tends to zero as $n \to \infty$.

Now, the bias of the predictor (3) is

$$B\left(T_n^{\lambda}\right) := \int_U \lambda_n(x) E(Z(x)) \, d\overline{G}_n(x) - \int_U E(Z(x)) \, dx,$$

where $\overline{G}_n = \frac{1}{n} \sum_{k=1}^{n} G_k$, G_k is the marginal distribution function of X_k. If the mean function $E[Z(x)]$ is known for all $x \in U$, then it is possible to find a function of coefficients λ_n such that $B\left(T_n^{\lambda}\right) = 0$. For example, if $E(Z(x)) = m \neq 0$ for all $x \in U$ and if the sample design and λ_n are such that $\int_U \lambda_n(x) \, d\overline{G}_n(x) = |U|$, then the predictor (3) is unbiased. On the other hand, if the uniform sampling design is used and $\lambda_n = |U|$ a.e., then the predictor (3) is also unbiased.

But if the mean function is not known and a non-uniform sampling design is desired, there is still the possibility of obtaining an unbiased predictor.

Proposition 1 Let G_n be a sampling design and λ_n be a nonnegative function on U, such that $\int_A \lambda_n(x)\, d\overline{G}_n(x) = |A|$, for all Borel subsets A of U. If moreover the mean function of Z_{U} is Lebesgue-integrable on U, then the predictor (3) is unbiased.

Given the assumptions for the function λ_n and the mean function of Z_{U}, the proof of the last proposition is a direct application of the chain rule.

The conditions for the sampling design and the function of coefficients in the last proposition imply uniform unbiasedness. The predictor (3) is unbiased for all Lebesgue-integrable mean functions under the conditions of the last proposition. In the spatial prediction by Kriging, the uniform unbiasedness property is also held by the Kriging predictor (e.g. Cressie, 1993, p. 120).

The utility of Proposition 1 is that under this choice of the sampling design and the function of coefficients, if $\overline{G}_n$ has density $\overline{g}_n$, then

$$|A| = \int_A \lambda_n(x)\, d\overline{G}_n(x) = \int_A \lambda_n(x)\, \overline{g}_n(x)\, dx,$$

for all Borel subsets A of U. This also means that the predictor (3) is unbiased if $\lambda_n \overline{g}_n = 1$ a.e. [Lebesgue] over U.

Using the above random sampling design as well as the function of coefficients, the following proposition can be proved.

Proposition 2 If the sample design associated with $\{X_1, \ldots, X_n\}$ is such that the distribution function $\overline{G}_n$ has a density function $\overline{g}_n$ and $\lambda_n \overline{g}_n = 1$ a.e.

[Lebesgue] over $\mathcal{U}$, then the predictor (3) becomes the unbiased predictor

$$T_{\pi_n} = \sum_{k=1} \frac{Z(X_k)}{\pi_n(X_k)}, \tag{4}$$

where π_n is as before, provided that $\pi_n > 0$ on $\mathcal{U}$.

The unbiased predictor (4) is the Horvitz-Thompson predictor of T when $z_{\mathcal{U}}$ is modeled with the random field $Z_{\mathcal{U}}$. This predictor extends the work of Cordy (1993).

Furthermore, if an unbiased linear homogeneous predictor of T is required, then the two previous propositions mean that the Horvitz-Thompson predictor should be used. If another exists, this is equal to the Horvitz-Thompson predictor a.s.

Now, the MSE of the predictor (4) is

$$\int_{\mathcal{U}} \frac{E\left[Z^2(x)\right]}{\pi_n(x)} dx + \int_{\mathcal{U}} \int_{\mathcal{U}} \frac{E[Z(x) Z(y)]}{\pi_n(x) \pi_n(y)} \pi_n(x, y)\, dx dy - E\left(T^2\right), \tag{5}$$

where $\pi_n(\cdot, \cdot)$ is as before.

If $\pi_n(\cdot, \cdot) > 0$ on $\mathcal{U} \times \mathcal{U}$, then an unbiased estimator of the MSE (5) is

$$\sum_{k=1} \frac{Z^2(X_k)}{\pi_n^2(X_k)} + \sum_{j=1} \sum_{k \neq j} \frac{Z(X_j) Z(X_k)}{\pi_n(X_j) \pi_n(X_k)} - \sum_{j=1} \sum_{k \neq j} \frac{Z(X_j) Z(X_k)}{\pi_n(X_j, X_k)}. \tag{6}$$

3.1 Simple random sampling

Here the simple random sampling design means that the set of random variables $\{X_1, \ldots, X_n\}$ are independent and identically distributed as G. Note that the marginal distribution function G does not change with n. The same definition is given in Schoenfelder & Cambanis (1982).

Under this sampling design the predictor (4) of the total T has the form

$$T_{\pi n,SRS} = \frac{1}{n} \sum_{k=1} \frac{Z(X_k)}{g(X_k)}, \tag{7}$$

where g is the density function of the design G with support in U.

The mean squared error of this predictor is

$$\frac{1}{n} \left\{ \int_U \frac{E\left[Z^2(x)\right]}{g(x)} dx - E\left(T^2\right) \right\}. \tag{8}$$

Note that this MSE tends to zero as $n \to \infty$, which shows the consistency in quadratic mean of the predictor (7).

To minimize the MSE (8) with respect to the design g, it is necessary to minimize the first part of that expression.

Proposition 3 For simple random sampling, the MSE (8) is minimized if and only if the sampling design G has a density of the form

$$g(x) = \frac{\sqrt{E[Z^2(x)]}}{\int_U \sqrt{E[Z^2(y)]}\, dy}. \tag{9}$$

Utilizing the Cauchy-Schwarz inequality, we can derive the optimal random sampling design (9). This Proposition corresponds to Proposition (3.1) of Schoenfelder & Cambanis (1982).

Remark 4 A necessary condition to obtain the last optimal random sampling design is to know the second moment function $E\left[Z^2(x)\right]$ for all $x \in U$. This is not a common situation. However, if any information about the variability of the field Z_U is available (e.g. information about $E\left[Z^2(x)\right]$), then a sampling design must be constructed using this information. A subset of U with high variability should have a high selection probability. Conversely, a subset with low variability should have a low selection probability.

4 Variable size samples

Occasionally, the required random sample is a variable size sample. In this situation, it is necessary to reformulate the sample concept given in Section 2. Now, the variable size samples are the realizations of a spatial (or multidimensional) random point process over U. This is denoted by $\{X_{1n}, \ldots, X_{nn}\}$, where each X_{kn} is a $d - dimensional$ random variable over U, and n is a counting process also over U. Moreover, the random variables Z_U and the spatial random point process $\{X_{1n}, \ldots, X_{nn}\}$ are defined over the same probability space as well as being stochastically independents.

In addition, the support of the random variable $n(U)$ is assumed to be in the positive integers. Furthermore, it is supposed that each value of the random variable $n(U)$ has been assigned a sampling design $G_{n(U)}$, the joint distribution function of $\{X_{1\mathrm{n(U)}}, \ldots, X_{\mathrm{n(U)n(U)}}\}$.

In the work of Cordy (1993), the variable size samples are not considered as the realizations of a spatial random point process. There, a variable size sample is only one element from the set of possible variable size samples.

Under this context, a possible Horvitz-Thompson predictor of the total T is

$$T_\pi = \sum_{k=1}^{\mathrm{n(U)}} \frac{Z\left(X_{k\mathrm{n(U)}}\right)}{\pi\left(X_{k\mathrm{n(U)}}\right)}, \tag{10}$$

where $\pi(x) = E\left[\pi_{\mathrm{n(U)}}(x)\right]$, provided that it exists and $\pi > 0$, $\pi_{\mathrm{n(U)}}(x) = \sum_{k=1}^{\mathrm{n(U)}} g_{k\mathrm{n(U)}}(x)$, and $g_{k\mathrm{n(U)}}$ is the marginal density function of $X_{k\mathrm{n(U)}}$ as before. The predictor (10) extends the estimator given in the Theorem 3 of Cordy (1993).

It is not difficult to show that $E\left(T_\pi | \mathrm{n(U)}\right) = \int_{\mathrm{U}} \frac{E(Z(x))}{\pi(x)} \pi_{\mathrm{n(U)}}(x)\, dx$ a.s. From this expression, it is possible to show that the predictor (10) is unbiased.

If the simple random sampling is applied for each value of $\mathrm{n(U)}$, then the predictor (10) takes the form

$$T_{\pi,SRS} = \frac{1}{E[\mathrm{n(U)}]} \sum_{k=1}^{\mathrm{n(U)}} \frac{Z\left(X_{k\mathrm{n(U)}}\right)}{g\left(X_{k\mathrm{n(U)}}\right)},$$

where g is the density function of the SRS and it is invariant with the values of $\mathrm{n(U)}$. The corresponding MSE of this predictor is

$$\frac{1}{E[\mathrm{n(U)}]} \left\{ \int_{\mathrm{U}} \frac{E\left(Z^2(x)\right)}{g(x)} dx - E\left(T^2\right) \right\} + \frac{Var[\mathrm{n(U)}]}{E^2[\mathrm{n(U)}]} E\left(T^2\right). \tag{11}$$

Observe that for given values of the first two moments of $n(U)$, the optimal sampling design is given by the expression (9).

Another possible Horvitz-Thompson predictor of the total T is

$$T_{\pi_{n(U)}} = \sum_{k=1}^{n(U)} \frac{Z\left(X_{kn(U)}\right)}{\pi_{n(U)}\left(X_{kn(U)}\right)}, \tag{12}$$

where $\pi_{n(U)}$ is given in the description of the formula (10), provided that $\pi_{n(U)} > 0$ for each value of $n(U)$.

Given the sample size $n(U)$, the predictor (12) is conditionally unbiased, that is $E\left[T_{\pi_{n(U)}} - T \mid n(U)\right] = 0$ a.s.

Under simple random sampling, the predictor (12) has the form

$$T_{\pi_{n(U)},SRS} = \frac{1}{n(U)} \sum_{k=1}^{n(U)} \frac{Z\left(X_{kn(U)}\right)}{g\left(X_{kn(U)}\right)}.$$

Its corresponding MSE is

$$E\left[\frac{1}{n(U)}\right]\left\{\int_U \frac{E\left(Z^2(x)\right)}{g(x)}dx - E\left(T^2\right)\right\}. \tag{13}$$

Once more, observe that for a given value of the first moment of $1/n(U)$, the optimal sampling design is also given by expression (9).

Examples of optimal sampling designs are given in Rodríguez (2002) for fixed and variable size samples.

5 Final remarks

The optimal sampling design associated with the Horvitz-Thompson predictor is a function of the second moment function of the random …eld (see Proposition 3). If this function is unknown, it is necessary to evaluate the impact on the MSE of the Horvitz-Thompson predictor from not using the correct second moment function. The problem of using an incorrect second moment function in the context of the Kriging predictor has been analyzed in Stein & Handcock (1989) among others. The methods used in that reference could be used for analyzing the e¤ect of an incorrect second moment function on the optimality of the sampling design.

References

Aldworth, J. and N. A. C. Cressie (1999), Sampling Designs and Prediction Methods for Gaussian Spatial Processes, in: S. Ghosh, ed. Multivariate Analysis, Design of Experiments and Survey Sampling (Marcel Dekker) pp. 1-54.

Cordy, C. B. (1993), An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe, Statistics & Probability Letters 18, 353-362.

Cressie, N. A. C. (1993), Statistics for spatial data (Wiley-Interscience, revised ed.).

Rodríguez, J. E. (2002), Optimal Sampling Designs and the Horvitz-Thompson Predictor in Random Fields Sampling, Doctoral Dissertation (Centro de Investigación en Matemáticas, México).

Schoenfelder, C. and S. Cambanis (1982), Random Designs for Estimating Integrals of Stochastic Processes, The Annals of Statistics 10, 526-538.

Stein, M. L. and M. S. Handcock. (1989). Some Asymptotic Properties of Kriging When the Covariance Function Is Misspeci...ed. Mathematical Geology, 21, 171-190.

Stevens, D. L. (1997), Variable Density Grid-Based Sampling Designs for Continuous Spatial Populations, Environmetrics 8, 167-195.

Proofs of results

• Properties of the estimator (1). a) unbiasedness:

$$
\begin{aligned}
E\left(t_{\pi_n}\right) &= \sum_{k=1} E\left[\frac{z\left(X_k\right)}{\pi_n\left(X_k\right)}\right] \\
&= \sum_{k=1} \int_U \frac{z\left(x\right)}{\pi_n\left(x\right)} g_k\left(x\right) dx.
\end{aligned}
$$

Given that $\pi_n = \sum_{k=1}^{n} g_k$, then

$$
\begin{aligned}
E\left(t_{\pi_n}\right) &= \int_U \frac{z\left(x\right)}{\pi_n\left(x\right)} \pi_n\left(x\right) dx \\
&= \int_U z\left(x\right) dx = t.
\end{aligned}
$$

b) Its variance:

$$
\begin{aligned}
Var\left(t_{\pi_n}\right) &= E\left(t_{\pi_n}^2\right) - t^2 \\
&= \sum_{k=1} E\left[\frac{z^2\left(X_k\right)}{\pi_n^2\left(X_k\right)}\right] + \sum_{j=1} \sum_{k \neq j} E\left[\frac{z\left(X_j\right) z\left(X_k\right)}{\pi_n\left(X_j\right) \pi_n\left(X_k\right)}\right] - t^2.
\end{aligned}
$$

Given that $\pi_n\left(x, y\right) = \sum_{j=1}^{n} \sum_{k \neq j} h_{jk}\left(x, y\right)$, then

$$
\begin{aligned}
Var\left(t_{\pi_n}\right) &= \int_U \frac{z^2\left(x\right)}{\pi_n^2\left(x\right)} \pi_n\left(x\right) dx + \int_U \int_U \frac{z\left(x\right) z\left(y\right)}{\pi_n\left(x\right) \pi_n\left(y\right)} \pi_n\left(x, y\right) dxdy - t^2 \\
&= \int_U \frac{z^2\left(x\right)}{\pi_n\left(x\right)} dx + \int_U \int_U \frac{z\left(x\right) z\left(y\right)}{\pi_n\left(x\right) \pi_n\left(y\right)} \pi_n\left(x, y\right) dxdy - t^2.
\end{aligned}
$$

Observe that

$$
t^2 = \int_U \int_U z\left(x\right) z\left(y\right) dxdy,
$$

then the last variance can be rewritten as:

$$\int_U \frac{z^2(x)}{\pi_n(x)}dx + \int_U\int_U \frac{\pi_n(x,y) - \pi_n(x)\pi_n(y)}{\pi_n(x)\pi_n(y)} z(x)z(y)\,dxdy.$$

This is the expression given in Cordy (1993) for the variance.

c) Unbiasedness of the estimator of its variance: If the steps of part (a) are followed, then it is possible to show that

$$E\left[\sum_{k=1} \frac{z^2(X_k)}{\pi_n(X_k)}\right] = \int_U \frac{z^2(x)}{\pi_n(x)}dx.$$

Now

$$E\left[\sum_{j=1}\sum_{k\neq j} \frac{z(X_j)z(X_k)}{\pi_n(X_j)\pi_n(X_k)}\right] = \sum_{j=1}\sum_{k\neq j}\int_U\int_U \frac{z(x)z(y)}{\pi_n(x)\pi_n(y)} g_{jk}(x,y)\,dxdy$$
$$= \int_U\int_U \frac{z(x)z(y)}{\pi_n(x)\pi_n(y)}\pi_n(x,y)\,dxdy.$$

Similarly, it is possible to show that

$$E\left[\sum_{j=1}\sum_{k\neq j} \frac{z(X_j)z(X_k)}{\pi_n(X_j,X_k)}\right] = \int_U\int_U z(x)z(y)\,dxdy$$
$$= t^2.$$

The last three expressions show that the variance estimator

$$\sum_{k=1} \frac{z^2(X_k)}{\pi_n(X_k)} + \sum_{j=1}\sum_{k\neq j} \frac{z(X_j)z(X_k)}{\pi_n(X_j)\pi_n(X_k)} - \sum_{j=1}\sum_{k\neq j} \frac{z(X_j)z(X_k)}{\pi_n(X_j,X_k)}$$

is unbiased. $\square$

$\bullet$ Proof of Proposition 1. The expected value of the predictor T_{λ_n} is

$$
\begin{aligned}
E\left(T_{\lambda_n}\right) &= \frac{1}{n}\sum_{k=1} EE\left[\lambda_n\left(X_k\right) Z\left(X_k\right) \mid X_1, \ldots, X_n\right] \\
&= \frac{1}{n}\sum_{k=1} E\left\{\lambda_n\left(X_k\right) E\left[Z\left(X_k\right)\right]\right\} \\
&= \frac{1}{n}\sum_{k=1}\int_{\mathrm{U}} \lambda_n\left(x\right) E\left[Z\left(x\right)\right] dG_k\left(x\right).
\end{aligned}
$$

Given that $\overline{G}_n = \frac{1}{n}\sum_{k=1}^{n} G_k$, then

$$
E\left(T_{\lambda_n}\right) = \int_{\mathrm{U}} \lambda_n\left(x\right) E\left[Z\left(x\right)\right] d\overline{G}_n\left(x\right). \tag{14}
$$

On the other hand, given that λ_n is a nonnegative function on U, such that $\int_A \lambda_n\left(x\right) d\overline{G}_n\left(x\right) = \int_A dx$ for all Borel subsets A of U, and the mean function $E\left[Z\left(x\right)\right]$ is Lebesgue-integrable on U, then by the chain rule (see P. Billingsley, 1995, Probability and Measure, Wiley, New York, 3rd ed., p. 214)

$$
\begin{aligned}
\int_{\mathrm{U}} \lambda_n\left(x\right) E\left[Z\left(x\right)\right] d\overline{G}_n\left(x\right) &= \int_{\mathrm{U}} E\left[Z\left(x\right)\right] d\left(x\right) \qquad (15) \\
&= E\left(T\right).
\end{aligned}
$$

Combining the expressions (14) and (15), the unbiasedness property is obtained.

$\square$

$\bullet$ Proof of Proposition 2. Observe that $\overline{g}_n = \frac{1}{n}\sum_{k=1}^{n} g_k$, where g_k is the marginal density function of X_k, which has support in U.

Now, given that $\lambda_n \bar{g}_n = 1$ a.e. [Lebesgue], then $\lambda_n = 1/\bar{g}_n$ a.e. $[\bar{g}_n]$. Thus,

$$
\begin{aligned}
T_{\lambda_n} &= \frac{1}{n} \sum_{k=1} \frac{Z(X_k)}{\bar{g}_n(X_k)} \quad \text{a.s.} \\
&= \sum_{k=1} \frac{Z(X_k)}{\sum_{j=1}^{n} g_j(X_k)}
\end{aligned}
$$

If we define $\pi_n = \sum_{j=1}^{n} g_j$, then $T_{\lambda_n} = T_{\pi_n}$ a.s. $\square$

• MSE (5):

$$
\begin{aligned}
E(T_{\pi_n} - T)^2 &= EE\left[T_{\pi_n}^2 - 2T_{\pi_n}T + T^2 \mid X_1, \ldots, X_n\right] \\
&= EE\left[\sum_{k=1} \frac{Z^2(X_k)}{\pi_n^2(X_k)} + \sum_{j=1}\sum_{k \neq j} \frac{Z(X_j)Z(X_k)}{\pi_n(X_j)\pi_n(X_k)}\right. \\
&\quad - 2\sum_{k=1}\int_{\mathcal{U}} \frac{Z(X_k)Z(y)}{\pi_n(X_k)}dy \\
&\quad \left.+ \int_{\mathcal{U}}\int_{\mathcal{U}} Z(x)Z(y)\,dxdy \,\middle|\, X_1, \ldots, X_n\right] \\
&= E\left[\sum_{k=1} \frac{E\left[Z^2(X_k)\right]}{\pi_n^2(X_k)} + \sum_{j=1}\sum_{k \neq j} \frac{E[Z(X_j)Z(X_k)]}{\pi_n(X_j)\pi_n(X_k)}\right. \\
&\quad - 2\sum_{k=1}\int_{\mathcal{U}} \frac{E[Z(X_k)Z(y)]}{\pi_n(X_k)}dy \\
&\quad \left.+ \int_{\mathcal{U}}\int_{\mathcal{U}} E[Z(x)Z(y)]\,dxdy\right] \\
&= \int_{\mathcal{U}} \frac{E\left[Z^2(x)\right]}{\pi_n(x)}dx + \int_{\mathcal{U}}\int_{\mathcal{U}} \frac{E[Z(x)Z(y)]}{\pi_n(x)\pi_n(y)}\pi_n(x,y)\,dxdy \\
&\quad - 2\int_{\mathcal{U}}\int_{\mathcal{U}} E[Z(x)Z(y)]\,dxdy + \int_{\mathcal{U}}\int_{\mathcal{U}} E[Z(x)Z(y)]\,dxdy \\
&= \int_{\mathcal{U}} \frac{E\left[Z^2(x)\right]}{\pi_n(x)}dx + \int_{\mathcal{U}}\int_{\mathcal{U}} \frac{E[Z(x)Z(y)]}{\pi_n(x)\pi_n(y)}\pi_n(x,y)\,dxdy \\
&\quad - E\left(T^2\right).
\end{aligned}
$$

$\square$

• Unbiasedness of the estimator (6) : First,

$$
\begin{aligned}
E\left[\sum_{k=1} \frac{Z^2(X_k)}{\pi_n^2(X_k)}\right] &= EE\left[\sum_{k=1} \frac{Z^2(X_k)}{\pi_n^2(X_k)} \,\middle|\, X_1,\ldots,X_n\right] \\
&= E\left\{\sum_{k=1} \frac{E\left[Z^2(X_k)\right]}{\pi_n^2(X_k)}\right\} \\
&= \int_U \frac{E\left[Z^2(x)\right]}{\pi_n(x)}\,dx. \qquad (16)
\end{aligned}
$$

Now, if a similar procedure is used as before, then it is possible to obtain that

$$
E\left[\sum_{j=1}\sum_{k\neq j} \frac{Z(X_j)\,Z(X_k)}{\pi_n(X_j)\,\pi_n(X_k)}\right] = \int_U\int_U \frac{E[Z(x)\,Z(y)]}{\pi_n(x)\,\pi_n(y)}\pi_n(x,y)\,dx\,dy \qquad (17)
$$

and

$$
E\left[\sum_{j=1}\sum_{k\neq j} \frac{Z(X_j)\,Z(X_k)}{\pi_n(X_j,X_k)}\right] = E\left(T^2\right). \qquad (18)
$$

Thus, combining in an appropriate way the expressions (16), (17), and (18), the property of unbiasedness of the estimator (6) is obtained

• MSE (8): From the proof of the MSE (5), it is possible to obtain that

$$
\begin{aligned}
E\left(T_{\pi_n,SRS} - T\right)^2 &= E\left[\frac{1}{n^2}\sum_{k=1}\frac{E\left[Z^2\left(X_k\right)\right]}{g^2\left(X_k\right)} + \frac{1}{n^2}\sum_{j=1}\sum_{k\neq j}\frac{E\left[Z\left(X_j\right)Z\left(X_k\right)\right]}{g\left(X_j\right)g\left(X_k\right)}\right. \\
&\quad - 2\frac{1}{n}\sum_{k=1}\int_U \frac{E\left[Z\left(X_k\right)Z\left(x\right)\right]}{g\left(X_k\right)}dx \\
&\quad \left. + \int_U\int_U E\left[Z\left(x\right)Z\left(y\right)\right]dxdy\right] \\
&= \frac{1}{n}\int_U \frac{E\left[Z^2\left(x\right)\right]}{g\left(x\right)}dx + \frac{n\left(n-1\right)}{n^2}\int_U\int_U E\left[Z\left(x\right)Z\left(y\right)\right]dxdy \\
&\quad - 2\int_U\int_U E\left[Z\left(x\right)Z\left(y\right)\right]dxdy + \int_U\int_U E\left[Z\left(x\right)Z\left(y\right)\right]dxdy \\
&= \frac{1}{n}\left\{\int_U \frac{E\left[Z^2\left(x\right)\right]}{g\left(x\right)}dx - E\left(T^2\right)\right\}.
\end{aligned}
$$

$\square$

• Proof of Proposition 3: From the Cauchy-Schwarz inequality,

$$
\begin{aligned}
\left(\int_U \sqrt{E\left[Z^2\left(x\right)\right]}\,dx\right)^2 &= \left[\int_U \frac{\sqrt{E\left[Z^2\left(x\right)\right]}}{\sqrt{g\left(x\right)}}\sqrt{g\left(x\right)}\,dx\right]^2 \\
&\leq \int_U \frac{E\left[Z^2\left(x\right)\right]}{g\left(x\right)}dx \int_U g\left(x\right)\,dx \\
&\leq \int_U \frac{E\left[Z^2\left(x\right)\right]}{g\left(x\right)}dx.
\end{aligned}
$$

The equality is achieved if and only if $g\left(x\right) = K\sqrt{E\left[Z^2\left(x\right)\right]}$, where K is a constant. Given that g is a density function, then $K = 1/\left(\int_U \sqrt{E\left[Z^2\left(y\right)\right]}\,dy\right)$. This shows that the MSE (8) is minimized if and only if the sampling design G

has density of the form

$$g(x) = \frac{\sqrt{E\left[Z^2(x)\right]}}{\int_U \sqrt{E\left[Z^2(y)\right]}\,dy}.$$

¤

• MSE (11) : From the proof of the MSE (8), it is possible to obtain that

$$
\begin{aligned}
E\left(T_{\pi,SRS}-T\right)^2 &= EE\left[\left(T_{\pi,SRS}-T\right)^2 \Big| n(U), X_1,\ldots,X_{n(U)}\right] \\
&= E\left[\frac{1}{E^2[n(U)]}\sum_{k=1}^{n(U)}\frac{E\left[Z^2(X_k)\right]}{g^2(X_k)}+\frac{1}{E^2[n(U)]}\sum_{j=1}^{n(U)}\sum_{k\neq j}\frac{E\left[Z(X_j)Z(X_k)\right]}{g(X_j)g(X_k)}\right. \\
&\quad -2\frac{1}{E[n(U)]}\sum_{k=1}^{n(U)}\int_U\frac{E\left[Z(X_k)Z(x)\right]}{g(X_k)}dx \\
&\quad \left.+\int_U\int_U E\left[Z(x)Z(y)\right]dxdy\right] \\
&= EE\left[\frac{1}{E^2[n(U)]}\sum_{k=1}^{n(U)}\frac{E\left[Z^2(X_k)\right]}{g^2(X_k)}+\frac{1}{E^2[n(U)]}\sum_{j=1}^{n(U)}\sum_{k\neq j}\frac{E\left[Z(X_j)Z(X_k)\right]}{g(X_j)g(X_k)}\right. \\
&\quad -2\frac{1}{E[n(U)]}\sum_{k=1}^{n(U)}\int_U\frac{E\left[Z(X_k)Z(x)\right]}{g(X_k)}dx \\
&\quad \left.+\int_U\int_U E\left[Z(x)Z(y)\right]dxdy \Big| n(U)\right] \\
&= E\left[\frac{n(U)}{E^2[n(U)]}\int_U\frac{E\left[Z^2(x)\right]}{g(x)}dx+\frac{n(U)[n(U)-1]}{E^2[n(U)]}\int_U\int_U E\left[Z(x)Z(y)\right]dxdy\right. \\
&\quad \left.-2\frac{n(U)}{E[n(U)]}\int_U\int_U E\left[Z(x)Z(y)\right]dxdy+\int_U\int_U E\left[Z(x)Z(y)\right]dxdy\right] \\
&= \frac{1}{E[n(U)]}\int_U\frac{E\left[Z^2(x)\right]}{g(x)}dx+\frac{E\left[n^2(U)\right]-E[n(U)]}{E^2[n(U)]}\int_U\int_U E\left[Z(x)Z(y)\right]dxdy \\
&\quad -2\int_U\int_U E\left[Z(x)Z(y)\right]dxdy+\int_U\int_U E\left[Z(x)Z(y)\right]dxdy \\
&= \frac{1}{E[n(U)]}\left\{\int_U\frac{E\left[Z^2(x)\right]}{g(x)}dx-E\left(T^2\right)\right\} \\
&\quad +\frac{E\left[n^2(U)\right]}{E^2[n(U)]}E\left[T^2\right]-E\left[T^2\right] \\
&= \frac{1}{E[n(U)]}\left\{\int_U\frac{E\left[Z^2(x)\right]}{g(x)}dx-E\left(T^2\right)\right\}+\frac{Var[n(U)]}{E^2[n(U)]}E\left[T^2\right].
\end{aligned}
$$

∎

• MSE (13) : Following the proof of the MSE (8) , it is possible to obtain that

$$
\begin{aligned}
E\left[\left(T_{\pi_{n(U)},SRS} - T\right)^2\right] &= E E\left[\left(T_{\pi_{n(U)},SRS} - T\right)^2 \,\middle|\, n(U)\right] \\
&= E\left\{\frac{1}{n(U)}\left[\int_U \frac{E\left[Z^2(x)\right]}{g(x)}dx - E\left[T^2\right]\right]\right\} \\
&= E\left(\frac{1}{n(U)}\right)\left\{\int_U \frac{E\left[Z^2(x)\right]}{g(x)}dx - E\left[T^2\right]\right\}.
\end{aligned}
$$

∎