Proceedings of the Annual Meeting of the American Statistical Association, August 5-9, 2001

# A Probabilistic Framework for Inference in Finite Population Sampling

José Elías Rodríguez
Facultad de Matemáticas, Universidad de Guanajuato, Apartado Postal 402,
Guanajuato, Gto., 36000 México

**Key Words: Design-Based Inference, Model-Based Inference, Finite Population Sampling**

## 1 Introduction

The objective of this research is to propose a probabilistic framework for inference in finite population sampling. There are a few seminal works, like [9], [7], [3], and [1]. They make a clear separation of two sources of randomness for inference in finite population sampling. The first source is the randomness produced by the sampling design and the second source is the randomness produced by the model of the characteristic of interest. Those seminal works also give an operative form to combine the two sources of randomness. Induced by this inferential approach, a formal method to combine the two sources of randomness will be proposed where all the statistical statements in finite population sampling will be well defined.

To propose a probabilistic framework for inference in finite population sampling, the design approach and the model approach inference will be first presented. Then, a formal method to combine the two approaches will be proposed. This combination will be the result of the construction of a product probability space. The elements of this product probability space will be the probability space induced by the sampling design as well as the probability space induced by the model of the characteristic of interest.

## 2 Preliminaries

First, the notation used throughout this work will be presented. A *finite population* of size $N$ is represented by

$$\mathcal{U} = \{1, \ldots, N\},$$

where $N$ is not necessarily known. The *characteristics of interest* of each element of the population are represented by

$$z_{\mathcal{U}} = \{z_1, \ldots, z_N\}.$$

Here we suppose that $z_k \in \mathbb{R}$; $k = 1, \ldots, N$. Some times the characteristics of interest are modeled by the set of random variables

$$Z_{\mathcal{U}} = \{Z_1, \ldots, Z_N\}.$$

These variables are defined over the probability space $(\Omega, \mathbb{A}, Q)$. Here '$Z_{\mathcal{U}}$ models $z_{\mathcal{U}}$' means that $z_{\mathcal{U}}$ is a possible value of $Z_{\mathcal{U}}$.

The probability space $(\Omega, \mathbb{A}, Q)$ or the join probability distribution function of $Z_{\mathcal{U}}$ is known as a *model* (superpopulation model in [7], [5], and [3]), for the characteristics of interest $z_{\mathcal{U}}$.

It is important to note that the behavior of $z_{\mathcal{U}}$ does not change if we select a sample from $U$, so the model does not have to depend in any sense on the method of sample selection.

A *sample* is a subset, say $s$, of the population $\mathcal{U}$. To generate the samples of $\mathcal{U}$ we utilize a *sampling scheme*. This is a sequence of experiments with the objective to select elements of the population $\mathcal{U}$ into the sample. The resulting probability of selecting these elements will be the sampling design. More details on sampling schemes are in the Section 1.2 of [4]. More accurate, if $S$ is a set of samples of $\mathcal{U}$, then a *sampling design* is a probability function, say $P$, over the $\sigma-$field $2^S$. For practical purposes this probability function $P$ has to be equal to the resulting probability of the sampling scheme above. Also for practical purposes, the natural set of samples $S$ is $\{s \subset \mathcal{U} : P(\{s\}) > 0\}$. If $S$ has that last property, it is called the set of *possible samples* of $\mathcal{U}$ under the sample design $P$. Thus, it can be formed a discrete probability space with the design $P$ and the set of samples $S$, $\left(S, 2^S, P\right)$, where the sampling scheme plays the role of the associated experiment.

If we only use the probability space $\left(S, 2^S, P\right)$ as source of uncertainty for inference in sampling, then it is called the *design-based inference* approach (see [9]). On the other hand, if we only use the model as the source of uncertainty for inference in sampling, then it is called the *model-based inference* approach. This term is used in [9] and [6]. In [1], the term *prediction-based inference* is used instead of model-based inference.

Now, the principal objective of this work is to present a formal method to combine the two sources of uncertainty.

Before, observe that any function from $S$ to $\mathbb{R}$ is a random variable given the structure of the probability space $\left(S, 2^S, P\right)$. A special class of random variables is

$$\lambda_k(s) = 0 \text{ if } k \notin s,$$
$$\lambda_k(s) \neq 0 \text{ if } k \in s,$$

where $s \in S$ and $k \in U$. A particular case of these random variables is

$$I_k(s) = \begin{cases} 0 & \text{si} \quad k \notin s, \\ 1 & \text{si} \quad k \in s, \end{cases}$$

for all $s \in S$ and $k \in U$. The last set of random variables is used in the Horvitz-Thompson type estimators.

Now, suppose that the prediction of the total $T$ given by

$$T = \sum_{k \in \mathcal{U}} Z_k$$

is required. Then, a natural predictor of $T$ is

$$T_\lambda = \sum_{k \in \mathcal{U}} \lambda_k Z_k. \tag{2.1}$$

However, its expected value among other statistical characteristics can not be formally calculated. This is due to the random variables $Z_{\mathcal{U}}$ and the random variables $\lambda_{\mathcal{U}} = \{\lambda_1, \ldots, \lambda_N\}$ are defined over different probability spaces.

In the next section a method to combine the two probability spaces will be shown. Then, a form to redefine the random variables $Z_{\mathcal{U}}$ and $\lambda_{\mathcal{U}}$ over the resulting probability space will be shown too.

# 3 General probability framework for inference in sampling

In this section a formal combination of the two sources of uncertainty will be presented.

First, the product probability space $\left(S \times \Omega, 2^S \times \mathbb{A}, P \times Q\right)$ is formed. Here $S \times \Omega$ is the Cartesian product of $S$ and $\Omega$, $2^S \times \mathbb{A}$ denotes the smallest sigma field generated by the Cartesian products $\left\{A \times B : A \in 2^S \text{ and } B \in \mathbb{A}\right\}$, and $P \times Q$ is a probability function such that $(P \times Q)(A \times B) = P(A) Q(B)$, where $A \in 2^S$ and $B \in \mathbb{A}$. This product probability space will be the source of uncertainty for inference in finite population sampling.

The definition of the product probability $P \times Q$ corresponds to the *non-informative sampling design* concept (see [9] and the Chapter 1 of [2]). If the sampling design is *informative*, we need define the probability function over $2^S \times \mathbb{A}$ in a different manner than above (examples of informative sampling designs can be seen in [10]). A proposal in this direction is given in [8].

Now, the random variables $Z_{\mathcal{U}}$ and $\lambda_{\mathcal{U}}$ are redefined on this new probability space. Let $\lambda_k : S \times \Omega \to \mathbb{R}$ and $Z_k : S \times \Omega \to \mathbb{R}$ be random variables such that

$$\lambda_k\left((s,\omega)\right) = \lambda_k\left(s\right)$$

and

$$Z_k\left((s,\omega)\right) = Z_k\left(\omega\right),$$

for all $(s,\omega) \in S \times \Omega$ and $k \in U$. Note that it is used the same name for the redefined variables and they are numerically equal to the original ones.

In the previous section was pointed out that the model of $Z_{\mathcal{U}}$ does not have to depend in any sense on the method of the sample selection. This condition is reflected in the following proposition as a direct consequence of the above redefinition.

**Proposition 1** *The set of random variables $Z_{\mathcal{U}}$ is stochastically independent of the set of random variables $\lambda_{\mathcal{U}}$.*

***Proof.*** *First, if $A \in 2^S$ and $B \in \mathbb{A}$, then the events $A \times \Omega$ and $S \times B$ are independent, since*

$$\begin{aligned}(P \times Q)\left(A \times \Omega \cap S \times B\right) &= (P \times Q)\left(A \times B\right) \\ &= P\left(A\right) Q\left(B\right) \\ &= P\left(A\right) Q\left(\Omega\right) P\left(S\right) Q\left(B\right) \\ &= (P \times Q)\left(A \times \Omega\right) \\ &\quad \cdot (P \times Q)\left(S \times B\right).\end{aligned}$$

*Secondly, if $B$ and $B'$ are two Borel subsets of $\mathbb{R}$, then for all $j \in U$,*

$$\lambda_j^{-1}\left(B\right) = \left\{s \in S : \ \lambda_j\left(s\right) \in B\right\} \times \Omega$$

*and, for all $k \in U$,*

$$Z_k^{-1}\left(B'\right) = S \times \left\{\omega \in \Omega : \ Z_k\left(\omega\right) \in B'\right\}.$$

*Note that the same name for the original and redefined variables again is used.*

*Finally if $n, m \leq N$ are two integers, $A_1, \ldots, A_n, B_1, \ldots, B_m$ are Borel subsets of $\mathbb{R}$, and $\{i_1, \ldots, i_n\}, \{j_1, \ldots, j_m\} \subseteq U$, then using the two above results*

$$\begin{aligned}(P \times Q)\left(\left[\bigcap_{k=1}^{n} \lambda_{i_k}^{-1}\left(A_k\right)\right] \bigcap \left[\bigcap_{l=1}^{m} Z_{j_l}^{-1}\left(B_l\right)\right]\right) \\ = (P \times Q)\left(\bigcap_{k=1}^{n} \lambda_{i_k}^{-1}\left(A_k\right)\right) \\ \cdot (P \times Q)\left(\left[\bigcap_{l=1}^{m} Z_{j_l}^{-1}\left(B_l\right)\right]\right).\end{aligned}$$

*This equation shows that the set of redefined random variables $Z_{\mathcal{U}}$ is stochastically independent of the set of redefined random variables $\lambda_{\mathcal{U}}$.* ■

Returning to the predictor of $T$ given by the expression (2.1), the bias of this predictor is

$$\sum_{k \in U} E\left(Z_k\right)\left(E\left(\lambda_k\right) - 1\right),$$

where the expected values are calculated with respect to the product probability function $P \times Q$. Therefore, if $E\left(\lambda_k\right) = 1$, for all $k \in U$, then $T_\lambda$ is unbiased.

References [6], [2], [9], and [5] define a *design-unbiased* predictor, say $\widehat{T}$, of $T$. There, this predictor is a design-unbiased predictor of $T$ if, and only if, for a given design $P$, $E_P\left(\widehat{t}\right) = t$ for all $(z_1, \ldots, z_N) \in \mathbb{R}^N$, where $t$ and $\widehat{t}$ are the values of $T$ and $\widehat{T}$, respectively, for each $(z_1, \ldots, z_N)$. There, $E_P$ also means expectation with respect to the sampling design $P$.

Here this property is equivalent to

$$E\left(\widehat{T} - T \middle| Z_1, \ldots, Z_N\right) = 0 \text{ a. e. } [P \times Q].$$

This means that the design-unbiased predictors in [6], [2], [9] and [5] are equivalent to the model-conditional unbiased predictors here.

In the previous references [6], [2], [9], and [5] also define a *model-unbiased* predictor. There, $\widehat{T}$ is a model-unbiased predictor of $T$ if, and only if, for a given model $Q$, $E_Q\left(\widehat{T} - T\right) = 0$ for all $s \in S$. $E_Q$ means expectation with respect to the probability function $Q$. Here this property is equivalent to

$$E\left(\widehat{T} - T \middle| \lambda_1, \ldots, \lambda_N\right) = 0 \text{ a. e. } [P \times Q].$$

This means that the model-unbiased predictors in the sampling literature are equivalent to the design-conditional unbiased predictors here.

Moreover, [2] and [9] define a *QP-unbiased* predictor. There $\widehat{T}$ is a QP-unbiased predictor of $T$ if, and only if, for a given $P$ and $Q$, $E_Q E_P\left(\widehat{T}-T\right)=0$. Here a QP-unbiased predictor is only an unbiased predictor. In the Chapter 4 of [2] is pointed out that a design-unbiased or model-unbiased predictor is a QP-unbiased predictor. Similarly here, it is easy to see that a model-conditional or design-conditional unbiased predictor is an unbiased predictor, since

$$\begin{aligned} E\left(\widehat{T}-T\right) &= E\left[E\left(\widehat{T}-T\right| Z_1,\dots,Z_N\right)\right] \\ &= E\left[E\left(\widehat{T}-T\right| \lambda_1,\dots,\lambda_N\right)\right]. \end{aligned}$$

Another term coined in [5] is the *anticipated variance*. This is defined as

$$E_Q E_P\left[\left(\widehat{T}-T\right)^2\right]-\left[E_Q E_P\left(\widehat{T}-T\right)\right]^2.$$

Here the anticipated variance is equivalent to $Var\left(\widehat{T}-T\right)$ and it can be expressed by

$$Var\left(\widehat{T}-T\right)=MSE\left(\widehat{T}\right)-B^2\left(\widehat{T}\right),$$

where $B\left(\widehat{T}\right)$ is the bias of the predictor $\widehat{T}$. In [5] and [1] they use this quantity to evaluate the efficiency of predictors.

Observe that the anticipated variance can be expressed as

$$\begin{aligned} Var\left(\widehat{T}-T\right) &= Var\left[E\left(\widehat{T}-T\right| Z_1,\dots,Z_N\right)\right] \\ &\quad +E\left[Var\left(\widehat{T}-T\right| Z_1,\dots,Z_N\right)\right] \\ &= Var\left[E\left(\widehat{T}-T\right| \lambda_1,\dots,\lambda_N\right)\right] \\ &\quad +E\left[Var\left(\widehat{T}-T\right| \lambda_1,\dots,\lambda_N\right)\right]. \end{aligned}$$

If $\widehat{T}$ is a model-conditional unbiased predictor, then

$$\begin{aligned} Var\left(\widehat{T}-T\right) &= E\left[Var\left(\widehat{T}-T\right| Z_1,\dots,Z_N\right)\right] \\ &= E\left[Var\left(\widehat{T}\right| Z_1,\dots,Z_N\right)\right] \end{aligned}$$

A similar statement can be found in [5]. Under the same condition, the anticipated variance is equal to

$$\begin{aligned} MSE\left(\widehat{T}\right) &= E\left(\widehat{T}^2\right)+E\left(T^2\right)-2E\left(\widehat{T}T\right) \\ &= E\left(\widehat{T}^2\right)+E\left(T^2\right) \\ &\quad -2E\left[TE\left(\widehat{T}\right| Z_1,\dots,Z_N\right)\right] \\ &= E\left(\widehat{T}^2\right)-E\left(T^2\right). \qquad (3.1) \end{aligned}$$

Furthermore, if $\widehat{T}$ is a design-conditional unbiased predictor, then

$$Var\left(\widehat{T}-T\right)=E\left[Var\left(\widehat{T}-T\right| \lambda_1,\dots,\lambda_N\right)\right].$$

As was the case before, a similar statement can be found in [5]. Also under the same condition, the anticipated variance is the $MSE\left(\widehat{T}\right)$ but it has not the same expression like (3.1).

One particular case of the predictor (2.1) of $T$ is the *Horvitz-Thompson predictor* of $T$. This predictor is defined by

$$T_\pi=\sum_{k\in U}\frac{I_k}{\pi_k}Z_k,$$

provided that $\pi_k>0$ for all $k\in\mathcal{U}$. This is a model-conditional unbiased predictor, since $E\left(I_k\right)=\pi_k$ for all $k\in\mathcal{U}$. However, it is not necessarily a design-conditional unbiased predictor.

The $MSE$ of the predictor $T_\pi$ is equal to:

$$\sum_{k\in U}\frac{1}{\pi_k}E\left(Z_k^2\right)+\sum\sum_{j\neq k\in U}\frac{\pi_{jk}}{\pi_j\pi_k}E\left(Z_jZ_k\right)-E\left(T^2\right).$$

It is not difficult to show that an unbiased estimator of this $MSE$ is given by

$$\begin{aligned} &\sum_{k\in U}\left(\frac{1}{\pi_k}-1\right)\frac{I_k}{\pi_k}Z_k^2 \\ &+\sum\sum_{j\neq k\in U}\left(\frac{\pi_{jk}}{\pi_j\pi_k}-1\right)\frac{I_jI_k}{\pi_{jk}}Z_jZ_k, \end{aligned}$$

provided that $\pi_{jk} = E\left(I_j I_k\right) > 0$ for all $j \neq k \in \mathcal{U}$.

The objective in presenting the above statistical statements is to demonstrate the utility of the product probability space proposed in this section.

## 4 Conclusions

The principal point of the probability framework presented here is as follows. If the two sources of randomness for inference in finite population sampling are required, first the combination of the two approaches has to be formalized. This was made proposing the product probability space $\left(S \times \Omega, 2^S \times \mathbb{A}, P \times Q\right)$ and redefining the random variables $Z_{\mathcal{U}}$ and $\lambda_{\mathcal{U}}$. Then the calculation of any statistical characteristic can be made in the formal way. Examples of these statistical characteristics are the bias and anticipated variance of any predictor of $T$.

## References

[1] Brewer, K. R. W. (1999). Design-based or prediction-based inference? Stratified random vs stratified balanced sampling. *International Statistical Review*, **67**, 35-47.

[2] Cassel, C. M., C. E. Särndal and J. H. Wretman (1977). *Foundations of inference in survey sampling.* John Wiley & Sons.

[3] Hansen, M. H., W. G. Madow and B. J. Tepping (1983). An evaluation of model-dependent and probability-sampling inference in sample surveys. *Journal of the American Statistical Association*, **78**, 776-793.

[4] Hedayat, A. S. and B. K. Sinha (1991). *Design and inference in finite population sampling.* John Wiley & Sons.

[5] Isaki, C. T. and W. A. Fuller (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, **77**, 89-96.

[6] Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, **57**, 377-387.

[7] Royall, R. M. and W. G. Cumberland (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, **76**, 66-77.

[8] Rubin-Bleuer, S. (2000). Some issues in the analysis of complex survey data. *Statistics Canada Series, Methodology Branch, Business Survey Methods Division*, BSMD-20-001 E.

[9] Särndal, C. E. (1978). Design-based and Model-based inference in survey sampling. *Scandinavian Journal of Statistics*, **5**, 27-52.

[10] Zacks, S. (1969). Bayes sequential designs for sampling finite populations. *Journal of the American Statistical Association*, **64**, 1342-1349.