
Una nota sobre algunas estrategias representativas en muestreo y su validez práctica

A Note on Some Representative Strategies in survey Sampling and its Practical Validity

Andrés Gutiérrez^a
hugogutierrez@usantotomas.edu.co

Resumen

Una estrategia de muestreo es una dupla compuesta de un diseño de muestreo y un estimador. En este artículo se tratará el problema de escoger una estrategia de muestreo representativa para las variables auxiliares con el fin de aumentar la precisión de las estimaciones del total de una variable de interés en una población finita. Aunque existen diseños de muestreo y estimadores que inducen estrategias representativas, se concluye, por medio de una simulación de Monte Carlo, que en términos de eficiencia no siempre es mejor utilizar un estimador de calibración bajo un diseño de muestreo balanceado, como uno podría suponer.

Palabras clave: calibración, estimador de Horvitz-Thompson, estrategia representativa, muestreo balanceado.

Abstract

A sampling strategy is the combination of a sampling design and an estimator. In this paper is presented the problem of choosing a representative strategy with respect to some auxiliary variables, in order to increase the precision of the resulting estimates for the population total. There exist several sampling designs and estimators that induce a representative strategy, but it is concluded, by means of some empirical simulations, that it is not always better the choice of an strategy given by the use of a calibration estimator under a balanced sampling design, as one may suppose.

Key words: Balanced Sampling, Calibration, Horvitz-Thompson Estimator, Representative Sampling Strategy.

^aDocente investigador. Facultad de Estadística. Universidad Santo Tomás.

1. Introducción

Suponga que se tiene una población finita \mathbf{U} , es decir, un conjunto de N unidades $\{u_1, u_2, \dots, u_N\}$ que pueden ser identificadas sin ambigüedad por un rótulo. Sea

$$U = \{1, 2, \dots, N\}$$

el conjunto de estos rótulos. Nótese que el tamaño de la población no es necesariamente conocido. Para todo $k = 1, \dots, N$, la k -ésima unidad de la población tiene asociado un valor de la característica de interés y_k , y un vector de p características de información auxiliar $\mathbf{x}_k = (x_{k1}, \dots, x_{kp})'$. Los valores de $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ se asumen conocidos. El total poblacional de la característica y y del vector de variables auxiliares \mathbf{x} son

$$t_y = \sum_{k=1}^N y_k \quad (1)$$

y

$$\mathbf{t}_x = \sum_{k=1}^N \mathbf{x}_k, \quad (2)$$

respectivamente. Por otro lado, si \mathbf{S}^1 representa una muestra aleatoria que toma el valor \mathbf{s} con probabilidad $p(\mathbf{S} = \mathbf{s}) = p(\mathbf{s})$, entonces $p(\mathbf{s})$, llamado diseño de muestreo, define una distribución de probabilidad multivariante sobre el soporte Q , que representa el conjunto de todas las posibles muestras (Tillé 2006). Luego, $\sum_{\mathbf{s} \in Q} p(\mathbf{s}) = 1$.

Definición 1.1. Siendo $\hat{t}_{\mathbf{s}}(y)$ un estimador² de t_y y $p(\mathbf{s})$ un diseño de muestreo definido sobre Q . A la dupla $(p(\cdot), \hat{t}_{\mathbf{s}}(\cdot))$ se le llama estrategia de muestreo.

La anterior definición parece ser estándar en la literatura actual (Särndal et al. 2003, Gutiérrez 2009) y en algunos textos como el de Hájek & Dupac (1981). Sin embargo, en algunos textos clásicos como Hansen et al. (1953), Cochran (1977) y Kish (1995), el término «diseño muestral» incluye tanto el plan de muestreo como el método de estimación. Es claro que en este artículo se prefiere usar el término «estrategia» como la combinación de los dos, diseño de muestreo y estimador. Siguiendo las directrices de Hájek & Dupac (1981) y de Deville & Tillé (2004), se tiene la siguiente definición.

¹Esta representación vectorial de una muestra aleatoria difiere de la representación clásica de muestra aleatoria como puede ser encontrada en Särndal et al. (2003), en donde la muestra aleatoria se nota como S y representa un conjunto aleatorio y la realización de este conjunto aleatorio es la muestra seleccionada s . Sin embargo, aunque diferentes, indican lo mismo y por tanto en este artículo se utilizará indiferentemente esta notación.

²En términos de notación, dado que \mathbf{S} es una muestra aleatoria, escribimos $\hat{t}_{\mathbf{s}}(y)$ como una función de \mathbf{S} y por lo tanto resulta ser efectivamente una variable aleatoria que pretende estimar a t_y . Cuando se selecciona una muestra \mathbf{s} , entonces $\hat{t}_{\mathbf{s}}(y)$ indica una realización del estimador.

Definición 1.2. Una estrategia de muestreo se dice representativa con respecto a \mathbf{x} , si y sólo si

$$\hat{t}_{\mathbf{s}}(\mathbf{x}) = t_{\mathbf{x}}. \quad (3)$$

Es decir, si el estimador aplicado a las variables auxiliares reproduce exactamente el total poblacional de las mismas.

Siendo \mathcal{P} un conjunto de diseños de muestreo y \mathcal{T} un conjunto de estimadores, en este artículo se aborda el problema de escoger un diseño de muestreo $p(\mathbf{s}) \in \mathcal{P}$ y un estimador del total poblacional $\hat{t}_{\mathbf{S}}(y) \in \mathcal{T}$, tal que la estrategia resultante sea una estrategia representativa que mantenga una alta eficiencia para el estimador.

Es claro que lo que el estadístico desea es disponer de una estrategia representativa para estimar el total poblacional t_y sin sesgo y con varianza nula. Por supuesto, esto es imposible de conseguir en la práctica. Sin embargo, es posible extender ese concepto y ligarlo con la definición anterior, puesto que al tener una estrategia de muestreo que induzca representatividad en las características de información auxiliar, y suponiendo que el comportamiento estructural de esas variables auxiliares es similar al comportamiento estructural de la característica de información auxiliar, entonces es posible suponer que se tenga una buena estrategia de muestreo (sin sesgo y con varianza pequeña) para los parámetros poblacionales de interés.

Sin embargo, como se demostrará más adelante, el estadístico no debe dejarse confundir por el anterior razonamiento y debe tener en cuenta que la representatividad en las características de información auxiliar no necesariamente induce una estrategia de muestreo con varianza pequeña. Es decir, como lo afirma Gutiérrez (2010a), es necesario verificar rigurosamente que exista una buena correlación entre las características de información auxiliar y las características de interés; más aún, se debe verificar que el comportamiento estructural entre la característica de interés y las ponderaciones resultantes de la estrategia de muestreo, ya sean las probabilidades de inclusión inducidas por un diseño de muestreo o los factores de expansión inducidos por un estimador, sea directamente proporcional.

Este artículo constituye una propuesta práctica por reinstaurar el término «representativo», de uso tan común entre los medios que desconocen la temática estadística, y algunas veces, tan poco usado por los mismos estadísticos que trabajan en la selección de muestras. En efecto, se lee y escucha mucho en los medios sobre «muestra representativa», en un intento por afirmar que el carácter estructural de un estudio tiene una base probabilística. Con el fin de relacionar las conclusiones prácticas de Gutiérrez (2010a), en este artículo se intenta acuñar el término para una visión más amplia y con un énfasis empírico. Se llega a interesantes conclusiones que servirán de evidencia para dar un mejor soporte al momento de escoger una estrategia de muestreo.

Con la lectura y entendimiento de los procesos que actúan detrás de una estrategia representativa, el estadístico tendrá mucha más precaución al momento de elegir una estrategia de muestreo, pues, como se nota en los resultados empíricos,

no siempre utilizar una estrategia representativa produce mejores resultados, en términos de eficiencia estadística, que utilizar una estrategia de muestreo no representativa. De este modo, cuando se mencione alguna afirmación con respecto a una «muestra representativa», el estadístico sabrá que sólo cuando se cumple el principio de representatividad, ligado a la regla de oro del muestreo (Gutiérrez 2010a), es válido utilizar una estrategia representativa.

El orden del artículo es como sigue a continuación. En la sección dos se presentan varios estimadores que permiten obtener estrategias representativas y se revisa brevemente el método de calibración. En la sección tres se presenta el método del cubo (Tillé 2006) como una herramienta que permite la selección de muestras balanceadas que junto con el estimador de Horvitz-Thompson induce estrategias representativas. En la sección cuatro se comparan varias estrategias representativas óptimas, tanto en el diseño como en el estimador, para una variable auxiliar, utilizando resultados empíricos producto de una simulación de Monte Carlo. En la última sección se presentan algunos comentarios y conclusiones.

2. Algunos estimadores que inducen estrategias representativas

Sin importar el diseño de muestreo utilizado para la selección de la muestra, si el total poblacional de las variables auxiliares, t_x , es conocido, se puede utilizar esta información para construir un estimador aún más preciso. En este artículo se consideran los estimadores lineales de la forma

$$\hat{t}_S(y) = w_0 + \sum_{k \in S} w_k y_k, \quad (4)$$

en donde los pesos w_k pueden depender del vector de información auxiliar. Nótese que el subíndice k varía dentro de un conjunto aleatorio S . Es claro que no todos los estimadores lineales cumplen la ecuación de representatividad (3). Por ejemplo, siendo $\pi_k = P(k \in S)$, la probabilidad de inclusión de una unidad a la muestra, Horvitz & Thompson (1952) definen el estimador siguiente

$$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k}, \quad (5)$$

con varianza dada por

$$Var(\hat{t}_{y\pi}) = \sum_{k \in U} \sum_{l \in U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

Donde $\Delta_{kl} = Cov(I_k, I_l)$. Este estimador es insesgado pero no utiliza información auxiliar. No es difícil mostrar que, utilizando un diseño de muestreo de tamaño

de muestra fijo, el estimador de $\hat{t}_{y\pi}$ arroja una estrategia representativa sobre el vector de probabilidades de inclusión $\boldsymbol{\pi} = (\pi_1, \dots, \pi_N)'$.

Para el caso en el que se tiene una sola característica de información auxiliar, entonces \mathbf{x}_k se convierte en x_k y por lo tanto, si $\hat{t}_{y\pi}$ y $\hat{t}_{x\pi}$ son los estimadores de Horvitz-Thompson de t_y y t_x respectivamente, entonces es posible construir estimadores que, sin importar el diseño de muestreo, arrojen estrategias representativas sobre el vector de información auxiliar \mathbf{x} . En el caso de una variable auxiliar, se tienen (Chaudhuri & Stenger 2006):

- Estimador de razón: $\hat{t}_R = \hat{t}_{y\pi} \frac{t_x}{\hat{t}_{x\pi}}$
- Estimador de diferencia: $\hat{t}_D = \hat{t}_{y\pi} + t_x - \hat{t}_{x\pi}$
- Estimador de regresión: $\hat{t}_{reg} = \hat{t}_{y\pi} + (t_x - \hat{t}_{x\pi})\hat{b}$. En donde \hat{b} es un estimador del coeficiente de regresión de y sobre x

Los anteriores estimadores satisfacen la ecuación (3) pues estiman el total t_x con varianza nula.

2.1. Estimadores de calibración

En general, es posible construir estimadores que arrojen estrategias representativas; sin embargo, Deville & Särndal (1992) proponen un clase de estimadores lineales

$$\hat{t}_{y,cal} = \sum_{k \in S} w_k y_k, \quad (6)$$

con $w_0 = 0$, y cuyos pesos w_k son tales que

$$\sum_{k \in S} w_k \mathbf{x}_k = \mathbf{t}_x. \quad (7)$$

A los estimadores de la clase (6) se les llama estimadores de calibración. Luego, bajo cualquier diseño de muestreo, los estimadores de calibración arrojan estrategias representativas sobre el vector de variables auxiliares \mathbf{x} . Los pesos w_k se construyen minimizando una distancia, restringida a (7), entre los nuevos pesos y los pesos originales inducidos por el diseño de muestreo, $d_k = \frac{1}{\pi_k}$. De tal manera que los w_k y $\frac{1}{\pi_k}$ sean tan cercanos como sea posible.

El estimador de regresión generalizada propuesto por Cassel & Wretman (1976) es un caso particular de los estimadores de calibración. Éste se encuentra cuando se hallan los pesos de calibración w_k , mediante la minimización de la distancia Ji-cuadrado dada por $\sum_s \frac{(w_k - d_k)^2}{d_k}$ y está dado por

$$\hat{t}_{y,cal} = \sum_{k \in S} w_k y_k = \hat{t}_{y\pi} + (\mathbf{t}_x - \mathbf{t}_{x\pi})' \hat{\mathbf{B}}, \quad (8)$$

donde

$$\hat{\mathbf{B}} = \left(\sum_S \frac{q_k \mathbf{x}_k \mathbf{x}'_k}{\pi_k} \right)^{-1} \left(\sum_S \frac{q_k \mathbf{x}_k y_k}{\pi_k} \right), \quad (9)$$

es un estimador de muestreo para el parámetro de regresión poblacional $\mathbf{B} = (\sum_U q_k \mathbf{x}_k \mathbf{x}'_k)^{-1} (\sum_U q_k \mathbf{x}_k y_k)$ y q_k es un término no correlacionado con π_k , que le da al k -ésimo elemento una ponderación positiva y conocida.

En términos de la precisión del estimador de calibración, Deville & Särndal (1992) comentan que si la relación entre los y_k y \mathbf{x}'_k es perfecta; es decir, $y_k = \mathbf{x}'_k \mathbf{B}$ para todo $k = 1, \dots, N$, entonces la varianza de $\hat{t}_{y,cal}$ es cero. Siendo así, entonces \mathbf{x} permite predecir perfectamente a y . Entonces, es posible hallar un vector \mathbf{B} tal que las diferencias $y_k - \mathbf{x}'_k \mathbf{B}$ sean pequeñas para todo k . Luego, se espera que el método de calibración arroje estimaciones cercanas al total t_y .

Al minimizar otras distancias sujetas a (7), surgen nuevos estimadores de calibración. Acudiendo al resultado 5 de Deville & Särndal (1992), todos aquellos estimadores resultantes son asintóticamente equivalentes al estimador (8); esto hace que la varianza de tales estimadores sea la misma, en términos asintóticos.

3. Algunos diseños de muestreo que inducen estrategias representativas

En esta sección abordamos el problema de la selección de muestras que reproducen automáticamente los totales poblacionales de las variables auxiliares. Utilizando el estimador de Horvitz-Thompson, a un diseño de muestreo $p(\mathbf{s})$ se le llama diseño de muestreo balanceado si éste permite la selección de muestras para las cuales se tiene que

$$\hat{\mathbf{t}}_{x\pi} = \mathbf{t}_x. \quad (10)$$

Nótese que un diseño de muestreo con estas características define una distribución de probabilidad sobre un soporte restringido a (10). Sólo las muestras que satisfagan las ecuaciones de balanceo tienen un probabilidad de selección mayor que cero; es decir, el soporte está dado por

$$Q = \left\{ s \in S : \sum_{k \in U} \frac{\mathbf{x}_k I_k}{\pi_k} = \mathbf{t}_x \right\}, \quad (11)$$

donde

$$I_k = \begin{cases} 1 & \text{si la unidad } k \text{ está en la muestra} \\ 0 & \text{en otro caso} \end{cases}. \quad (12)$$

Así, el estimador de Horvitz-Thompson junto con el diseño un muestreo balanceado definen una estrategia representativa sobre la información auxiliar. En otras palabras³ $Var(\hat{\mathbf{X}}_{HT}) = \mathbf{0}$.

3.1. El método del cubo

Recientemente, Deville & Tillé (2004) desarrollaron el método del cubo, que es un algoritmo de muestreo para obtener muestras balanceadas y se compone de dos fases, llamadas la fase de vuelo y la fase de aterrizaje. En la primera fase, para que las restricciones de representatividad, dadas por (10), sean satisfechas exactamente, se deben redondear a cero o a uno todas las probabilidades de inclusión de primer orden. La fase de aterrizaje consiste en el manejo adecuado del redondeo. En términos del planteamiento del plan de muestreo y la escogencia de probabilidades de inclusión óptimas, el lector puede referirse a Tillé & Anne-Catherine (2005).

3.1.1. Fase de vuelo

Para un muestreo sin reemplazo y de tamaño fijo, el conjunto de todas las posibles muestras de tamaño n se puede ver como un conjunto de vértices de un cubo en el espacio N -dimensional restringido a que la suma de las coordenadas de cada vértice sea n . La fase de vuelo constituye una martingala (una generalización de una caminata aleatoria) que comienza con un vector de probabilidades de inclusión y permanece en la intersección del cubo y el subespacio restringido por las ecuaciones de balanceo. Al final de esta fase, la caminata aleatoria se detiene en un vértice de dicha intersección.

Siguiendo con Tillé (2006), el objetivo de esta fase es escoger aleatoriamente un vértice de

$$K = \{[0, 1]^N \cap \mathbf{Q}\}$$

de tal forma que las ecuaciones de balanceo se reproduzcan a satisfacción, condicionado a las probabilidades de inclusión iniciales, donde $\mathbf{Q} = \boldsymbol{\pi} + \text{kernel}(\mathbf{A})$, la matriz \mathbf{A} está definida por $\mathbf{A} = (\mathbf{x}_1/\pi_1, \dots, \mathbf{x}_N/\pi_1)$ y $\text{kernel}(\mathbf{A}) = \{\mathbf{v} \in \mathbb{R}^N | \mathbf{A}\mathbf{v} = \mathbf{0}\}$. La fase de aterrizaje es sólo necesaria si el vector escogido no es un vértice del cubo y consiste en flexibilizar las restricciones (lo menos posible) para seleccionar una muestra; esto es, un vértice del cubo.

³Nótese que $Var(\hat{\mathbf{X}}_{HT})$ es una matriz de varianzas y covarianzas, y el hecho de que sea nula no significa que las p variables que constituyen la matriz de información auxiliar sean linealmente independientes, sino que los p estimadores de Horvitz-Thompson para las características de información auxiliar son constantes para cualquier posible muestra seleccionada; de esta manera, su varianza es nula y la covarianza entre cualquier par de estimadores es igualmente nula.

3.1.2. La fase de aterrizaje

Al final de la primera fase, el algoritmo no necesariamente se detiene en una muestra. Suponga que en la iteración T , el algoritmo ha alcanzado un vértice de K , el cual no es necesariamente un vértice del cubo $[0, 1]^N$. Este vértice es denotado como $\boldsymbol{\pi}^* = [\pi_k^*] = \boldsymbol{\pi}(T)$. Siendo q el número de componentes no enteras en este vértice, se tiene que si $q = 0$, entonces el algoritmo está completo puesto que todos los componentes del vértice pertenecen al conjunto $\{0, 1\}$. Por otra parte, si $q > 0$, entonces algunas restricciones no pudieron ser satisfechas rigurosamente, y se tiene que algunos componentes del vértice pertenecen al espacio $(0, 1)$.

Por lo tanto, si $U^* = \{k \in U \mid 0 < \pi_k^* < 1\}$, el objetivo es buscar un diseño muestral que arroje una muestra $s^* \subset U^*$ tal que

$$\sum_{k \in s^*} \mathbf{a}_k \approx \sum_{k \in U^*} \mathbf{a}_k \pi_k^*,$$

con $\mathbf{a}_k = \mathbf{x}_k / \pi_k$ y $s^* = s \cap U^*$. Tillé (2006) resuelve este problema mediante programación lineal, aplicando el método simplex y minimizando una función de costo, $C(\mathbf{s})$, sobre todas las posibles muestras compatibles⁴ con $\boldsymbol{\pi}^*$, de tal forma que

- $C(\mathbf{s}) > 0$, para toda \mathbf{s} compatible con $\boldsymbol{\pi}^*$.
- $C(\mathbf{s}) = 0$, para toda \mathbf{s} balanceada.

Esta programación lineal no depende del tamaño poblacional sino sólo del número de variables de balanceo. Si el número de variables auxiliares es muy grande, al final de la fase de vuelo se deben eliminar una o varias variables auxiliares. Por esta razón, Deville & Tillé (2004) argumentan que es importante ordenar las variables de balanceo de acuerdo a su correlación con las variables de interés.

4. Comparación empírica sobre la optimalidad de algunas estrategias representativas

Hasta este punto se ha analizado la representatividad como una cualidad propia de algunas estrategias de muestreo. Esta cualidad puede ir en dos direcciones, la representatividad inducida por el estimador utilizado o la representatividad inducida por el diseño de muestreo al combinarlo con el estimador, por ejemplo el de Horvitz-Thompson. A simple vista se podría pensar que la mejor estrategia, en términos de eficiencia, sería utilizar un diseño de muestreo balanceado junto con un estimador de calibración (ya que ambos inducen estrategias representativas), y

⁴Una muestra \mathbf{s} se dice compatible con un vector $\boldsymbol{\pi}^*$ si $\pi_k^* = I_k$ para todo k tal que π_k^* es un entero.

además una de las conclusiones de Deville & Tillé (2004), basados en los resultados de la simulación con la población MU284 de Särndal et al. (2003), es que la estrategia más precisa es el uso de un estimador de calibración después de haber seleccionado una muestra con un diseño balanceado.

Sin embargo, la motivación de este artículo es precisamente contradecir esta afirmación para casos específicos, y con esto lograr que el lector entienda que la cualidad de representatividad en una estrategia no tiene sentido cuando no se cumple la regla de oro del muestreo; es decir, cuando el comportamiento estructural de la característica de interés no es proporcional al comportamiento estructural de las probabilidades de inclusión π_k , en el caso del diseño de muestreo, o a las ponderaciones w_k , en el caso del estimador. Por tanto, en esta sección, utilizando los resultados de una simulación de Monte Carlo, se muestra que el balanceo y la calibración no siempre constituyen la mejor elección en términos de eficiencia estadística.

Con base en el anterior razonamiento, se realiza un estudio limitado por medio de una simulación con el fin de tener un acercamiento a la optimalidad de los diseños, estimadores y estrategias anteriores. Se simuló una población de tamaño $N = 1000$ de un modelo de superpoblación. Para este propósito, suponemos que la relación entre Y_k y \mathbf{x}_k puede ser descrita mediante un modelo ξ , tal que $Y_k = x_k\beta + \varepsilon_k$ y para todo $k = 1, \dots, N$. Además, se supone que $\varepsilon_1, \dots, \varepsilon_N$ constituyen un conjunto de variables aleatorias con distribución normal tal que:

$$\begin{aligned} E_{\xi}(\varepsilon_k) &= 0 \\ Var_{\xi}(\varepsilon_k) &= \sigma_k^2 = \sigma^2 \times x_k^2. \end{aligned} \tag{13}$$

Dado que Wu (2003) muestra que existen ciertas condiciones para que un diseño de muestreo sea regular⁵, distribuciones de colas pesadas como la distribución log-normal y la distribución gamma con parámetros de escala muy grandes, no podrán ser usadas para generar la información auxiliar. Por tanto, se generan los valores de x de una distribución gama con parámetro de forma uno y parámetro de escala dos. Esta variable toma valores no negativos y está sesgada a la derecha, lo cual es muy común en aplicaciones reales de encuestas por muestreo (Wu 2003).

El valor del parámetro β se fijó igual a treinta. Además, se asumió que los ε_k son independientes, con estructura heteroscedástica. En cada corrida de la simulación se tomaron muestras aleatorias simples sin balancear de tamaño $n = 100$, $n = 300$ y $n = 500$ y muestras balanceadas de igual tamaño. Para seleccionar las muestras aleatorias simples (*MAS*) se utilizó la función `sample` del ambiente base de R (R Development Core Team 2010), y para seleccionar las muestras balanceadas (*BAL*) se utilizó la función `samplecube` del paquete `sampling` (Tillé & Matei 2011), partiendo de probabilidades de inclusión iguales a $\pi_k = n/N$. Para la estimación del parámetro β se utilizó un enfoque basado en el diseño de muestreo,

⁵Un diseño de muestreo es regular si satisface: (i) $\max_{k \in s} \{n \times d_k\} = O(1)$ y (ii) si la distribución asintótica del estimador de Horvitz-Thompson es normal.

siguiendo el enfoque expuesto en Gutiérrez (2009, Sec. 8.4.3). Así, se definieron las siguientes estrategias de muestreo para la estimación del total poblacional, t_y :

- $(MAS, \hat{t}_{y,\pi})$: el estimador clásico con un diseño de muestreo aleatorio simple.
- $(BAL, \hat{t}_{y,\pi})$: el estimador clásico con un diseño de muestreo balanceado.
- $(MAS, \hat{t}_{y,cal})$: el estimador de calibración bajo muestreo aleatorio simple.
- $(BAL, \hat{t}_{y,cal})$: el estimador de calibración con un diseño de muestreo balanceado.

El proceso se repitió $B = 1000$ veces y en cada simulación, el desempeño de un estimador \hat{t}_y fue evaluado usando su sesgo relativo, y su eficiencia relativa⁶, definidas por Wu & Luan (2003), respectivamente, como:

$$SR(\hat{t}_y) = B^{-1} \sum_{b=1}^B \frac{\hat{t}_{y,b} - t_y}{t_y} \quad (14)$$

$$ER(\hat{t}_y) = \frac{ECM(\hat{t}_y)}{ECM_{MAS}(\hat{t}_{y,\pi})} \quad , \quad (15)$$

donde

$$ECM(\hat{t}_y) = B^{-1} \sum_{b=1}^B (\hat{t}_{y,b} - t_y)^2 \quad (16)$$

y el subscrito $ECM_{MAS}(\hat{t}_{y,\pi})$ corresponde al error cuadrado medio del estimador de Horvitz-Thompson bajo un diseño de muestreo aleatorio simple. Por lo tanto, teniendo en cuenta que $\hat{t}_{y,b}$ se calculó en la b -ésima muestra simulada, se nota que grandes valores para ER (> 1) representan una baja eficiencia de la estrategia $(p(\mathbf{s}), \hat{t}_y)$ en comparación a la estrategia $(MAS, \hat{t}_{y,\pi})$. Al contrario, valores cercanos menores a uno o cercanos a cero para ER (< 1) representan una alta eficiencia de la estrategia.

Teóricamente se conoce la varianza del estimador de Horvitz-Thompson y es posible computar la varianza de un estimador de calibración, cuando la estrategia involucra un diseño de muestreo aleatorio simple. Cuando se trata de muestreo balanceado, aunque teóricamente posible, no es factible realizar este cálculo puesto que demandaría mucho esfuerzo computacional. Nótese que existen expresiones cerradas para las varianzas de las estrategias $(MAS, \hat{t}_{y,\pi})$ y $(MAS, \hat{t}_{y,cal})$, pero no existen expresiones asequibles computacionalmente para el cálculo de las varianzas de las estrategias $(BAL, \hat{t}_{y,\pi})$ y $(BAL, \hat{t}_{y,cal})$, puesto que esas varianzas están dadas en términos de dobles sumatorias definidas para todo par de individuos en la

⁶Nótese que aunque ambos diseños de muestreo, el aleatorio simple y el balanceado, comparten las mismas probabilidades de inclusión, la probabilidad de selección de muestras es totalmente diferente para cada diseño, entre otras razones porque el soporte no es el mismo. Por lo tanto, no se espera que ni las estimaciones ni sus varianzas coincidan.

población. Por lo tanto, tratar de realizar una comparación empírica de la eficiencia de las estrategias en términos de varianzas teóricas o coeficientes de variación no es computacionalmente fácil de llevar a cabo. Ahora, aunque existen aproximaciones a estas varianzas (Deville & Tillé 2005), la comparación debe ser exacta. Esta es la razón de la utilización de simulaciones de Monte Carlo para hacer la comparación.

Por otro lado, se consideraron varios escenarios de correlación entre la característica de información auxiliar y la característica de interés. Es así como, cambiando los valores de σ^2 , se crearon tres escenarios: el primero, en donde la correlación es casi perfecta e igual a 0.99, el segundo en donde la correlación es de 0.8, y el último en donde la correlación es de 0.2. Las figuras 1, 2 y 3 muestran, en un diagrama de puntos y en sendos histogramas, el comportamiento estructural de la característica de interés y de la característica de información auxiliar en estos tres escenarios, respectivamente.

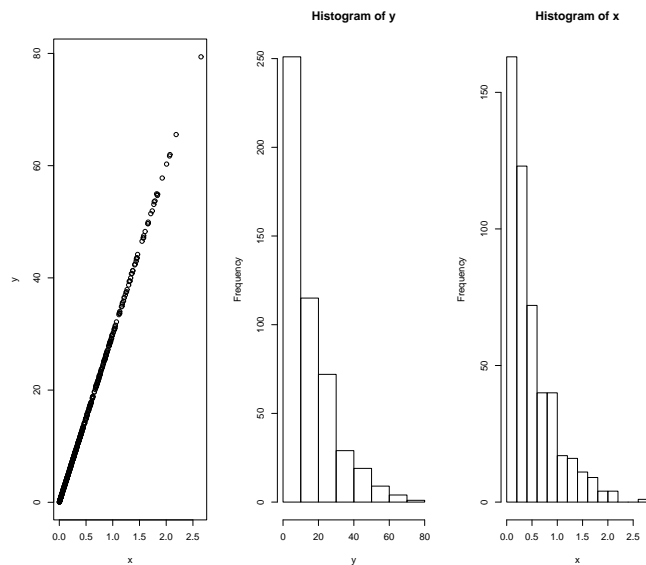


Figura 1: *Diagrama de puntos de y contra x, histograma de y e histograma de x para $\text{corr}(x, y) = 0.99$*

Con respecto al sesgo relativo, todos los resultados estuvieron muy cercanos a cero y, al considerarse despreciables, esas cifras no se muestran en este documento. Lo anterior es natural pues el estimador de Horvitz-Thompson es insesgado bajo cualquier diseño sin reemplazo y el estimador de calibración es asintóticamente insesgado bajo cualquier diseño de muestreo sin reemplazo.

En las tablas 1, 2 y 3, se puede contemplar la optimalidad de los estimadores bajo los diseños de muestreo considerados en la simulación. En particular, se nota que se gana en eficiencia cuando el tamaño de muestra aumenta, bajo todos los escenarios.

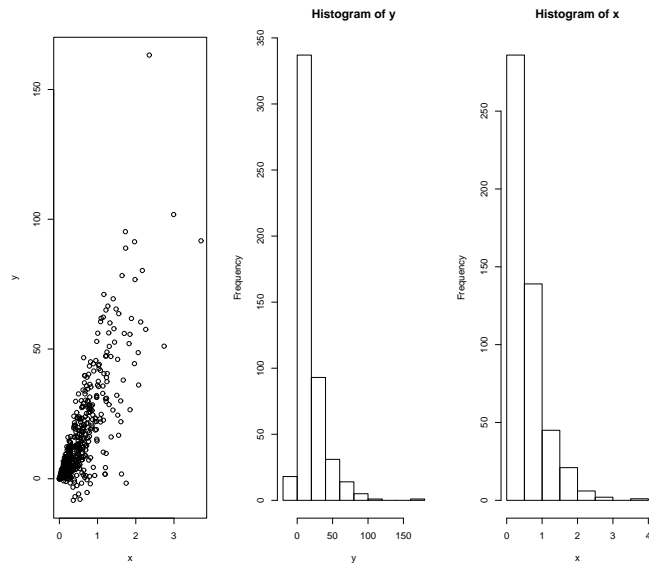


Figura 2: Diagrama de puntos de y contra x , histograma de y e histograma de x para $\text{corr}(x, y) = 0.80$

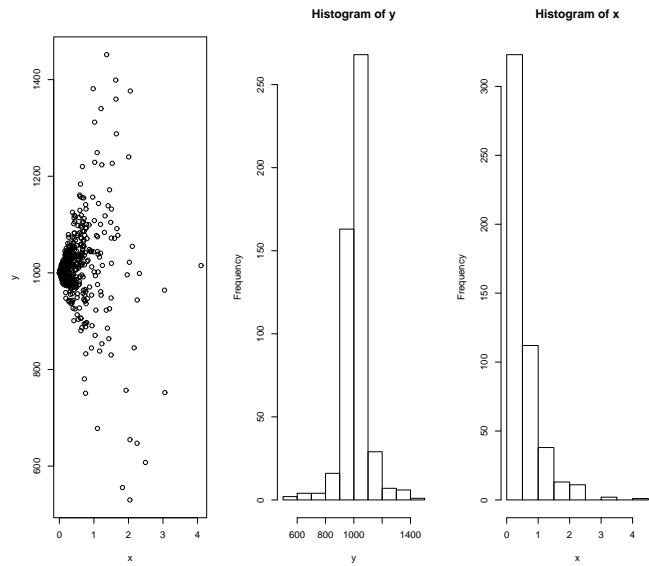


Figura 3: Diagrama de puntos de y contra x , histograma de y e histograma de x para $\text{corr}(x, y) = 0.20$

rios. Como se podría pensar de antemano, en todos los escenarios de correlación, bajo el diseño de muestreo balanceado el estimador de calibración tiene una menor varianza que el estimador de Horvitz-Thompson. La razón de lo anterior se debe a que al utilizar un estimador de calibración con una muestra balanceada, se esta presentando dos veces un proceso de representatividad sobre la misma característica de información auxiliar. Luego, se podría decir, abusando un poco del lenguaje técnico, que se presenta una doble calibración.

n	Muestreo aleatorio simple		Muestreo balanceado	
	$\hat{t}_{y,\pi}$	\hat{t}_{cal}	$\hat{t}_{y,\pi}$	$\hat{t}_{y,cal}$
10	1.000	0.335	0.495	0.345
50	1.000	0.344	0.365	0.336
200	1.000	0.388	0.401	0.380

Tabla 1: Eficiencia relativa de las estrategias de muestreo para $corr(x, y) = 0.99$

n	Muestreo aleatorio simple		Muestreo balanceado	
	$\hat{t}_{y,\pi}$	\hat{t}_{cal}	$\hat{t}_{y,\pi}$	$\hat{t}_{y,cal}$
10	1.000	0.871	0.965	0.908
50	1.000	0.923	0.892	0.889
200	1.000	0.835	0.802	0.799

Tabla 2: Eficiencia relativa de las estrategias de muestreo para $corr(x, y) = 0.80$

n	Muestreo aleatorio simple		Muestreo balanceado	
	$\hat{t}_{y,\pi}$	\hat{t}_{cal}	$\hat{t}_{y,\pi}$	$\hat{t}_{y,cal}$
10	1.000	1.075	1.115	1.093
50	1.000	0.998	1.027	1.020
200	1.000	1.008	1.063	1.065

Tabla 3: Eficiencia relativa de las estrategias de muestreo para $corr(x, y) = 0.20$

En los escenarios con alta correlación, se nota que bajo muestreo aleatorio simple, el estimador de calibración es siempre más eficiente que el estimador de Horvitz-Thompson. Asimismo, bajo muestreo balanceado, el estimador de calibración es siempre más eficiente que el estimador de Horvitz-Thompson. En estos dos escenarios, cuando la correlación es casi perfecta, la ganancia en eficiencia de las estrategias $(BAL, \hat{t}_{y,\pi})$, $(MAS, \hat{t}_{y,cal})$ y $(BAL, \hat{t}_{y,cal})$ es muy alta, con respecto a $(MAS, \hat{t}_{y,\pi})$. Sin embargo, cuando la estructura de correlación baja, la ganancia en eficiencia no es tan grande.

En el escenario en donde no se tiene una buena correlación, el comportamiento es diferente. Bajo muestreo aleatorio simple, al utilizar un estimador de calibración se aumenta la variación y por tanto se concluye que la distribución del estimador es poco precisa. Considerando un diseño de muestreo no balanceado, los resultados de la eficiencia de los estimadores son similares y concluyentes en términos de

que la distribución de estos estimadores conserva una amplia variabilidad. En conclusión, cuando se tiene este escenario, la mejor estrategia es $(MAS, \hat{t}_{y,\pi})$. Lo anterior es concluyente debido a que si en la práctica, los dos primeros escenarios son muy poco frecuentes y las características de información auxiliar raras veces presentan un comportamiento estructural proporcional al de la característica de interés, entonces lo mejor es dejar de lado los refinamientos teóricos y utilizar un simple y sencillo estimador de Horvitz-Thompson conservando el diseño aleatorio simple. Así, además de ser insesgado, el estimador de Horvitz-Thompson será más preciso que los estimadores de calibración.

Por otro lado, nótese que cuando se utiliza un diseño de muestreo sin reemplazo proporcional al tamaño de una característica de información auxiliar (πPT) , el estimador de calibración coincide plenamente con el estimador de Horvitz-Thompson puesto que las probabilidades de inclusión de primer están dadas por

$$\pi_k = n \frac{x_k}{t_x} \quad (17)$$

Por lo tanto, el estimador de Horvitz-Thompson es tal que

$$\hat{t}_{x,\pi} = \sum_{k \in S} \frac{x_k}{\pi_k} = t_x$$

y como consecuencia, el estimador de calibración toma la siguiente forma

$$\hat{t}_{y,cal} = \hat{t}_{y\pi} + (t_x - t_{x\pi})\hat{B} = \hat{t}_{y\pi}.$$

El mismo razonamiento se sigue para un diseño de muestreo balanceado con probabilidades de inclusión iniciales desiguales y dadas por la expresión 17. Sin embargo, dado que el tamaño de muestra para una estrategia $(\pi PT, \hat{t}_{y,\pi})$ o $(\pi PT, \hat{t}_{y,cal})$ es fijo, a diferencia de las estrategias $(BAL, \hat{t}_{y,\pi})$ o $(BAL, \hat{t}_{y,cal})$ en donde el tamaño de muestra es aleatorio, entonces los estimadores no necesariamente coinciden y, de la misma manera, la variabilidad en ambos escenarios no es la misma.

Con base en lo anterior, se repitió el ejercicio de simulación para diseños de muestreo sin reemplazo proporcionales al tamaño de x . Esta simulación sólo se realizó en el escenario en donde se tuvo una correlación baja. Todos los supuestos de las simulaciones anteriores se mantuvieron y se utilizó la función **S.pIPS** del paquete **TeachingSampling** (Gutiérrez 2010b) para la selección de muestras. En términos de eficiencia relativa, se encontró que las estrategias $(\pi PT, \hat{t}_{y,\pi})$ y $(\pi PT, \hat{t}_{y,cal})$ coinciden. Lo propio sucede con las estrategias $(BAL, \hat{t}_{y,\pi})$ y $(BAL, \hat{t}_{y,cal})$. Sin embargo, el cociente de las varianzas de estas estrategias con las varianzas de una estrategia con diseño simple y con el estimador de expansión $N\bar{y}_S = N/n \sum_S y_k$, denotada como $(MAS, N\bar{y}_S)$, es muy grande, como se puede ver en la Tabla 4. Por lo tanto, en estos casos es mejor abstenerse de utilizar estrategias sofisticadas y utilizar la estrategia más simple.

n	$N\bar{y}_S$	$\hat{t}_{y,\pi}$	\hat{t}_{cal}
10	1.000	1521.9	587.3
50	1.000	1125.1	515.3
200	1.000	845.6	651.1

Tabla 4: Eficiencia relativa de estrategias de muestreo para $corr(x, y) = 0.20$ con probabilidades de inclusión desiguales y proporcionales al tamaño de x .

5. Conclusiones

En términos del uso de diseños balanceados y estimadores de calibración, tal como lo afirman Deville & Tillé (2004), el uso de los estimadores de calibración ha tenido mucha acogida en las agencias estatales que hacen estudios por muestreo. Además, el diseño de muestreo balanceado ha venido siendo implementado en censos de países desarrollados como Francia (Tillé 2006). Por otra parte, Tillé (2006) sugiere esta combinación para solucionar problemas de ausencia de respuesta haciendo más sencillo el acercamiento del estadístico hacia los problemas prácticos que se presentan en la vida real, tales como deficiencias del marco de muestreo y errores de medición (Särndal 2007). En esta misma dirección, Deville (2005) propone una metodología estadística para el uso del método del cubo para balancear imputaciones causadas por la ausencia de respuesta.

Acudiendo al tema principal de este artículo, el uso del término muestra representativa se sigue utilizando indiscriminadamente en ambientes no técnicos, como medios de comunicación, revistas o periódicos. Por esta razón, el estadístico debe tener en cuenta que la representatividad es una característica propia de la estrategia de muestreo y no del subconjunto de unidades seleccionadas para el proceso de medición. La muestra debe ser una herramienta usada para obtener estimaciones y, por lo tanto, no admite el calificativo de representativa. Sin embargo, lo que sí admite este calificativo es la estrategia de muestreo utilizada para realizar inferencias sobre los parámetros de interés. Es así como el estadístico puede escoger entre un diseño de muestreo balanceado o un estimador de calibración o la combinación de ambos para obtener una estrategia representativa.

Como producto de los ejercicios realizados en este artículo, se recomienda utilizar, siempre que se pueda, un estimador de calibración, en la etapa de estimación junto con un diseño de muestreo balanceado, porque induce una estrategia representativa siempre y cuando se pueda garantizar que el comportamiento estructural de las características de información auxiliar es proporcional al de la característica de interés. Cuando no es posible sostener este supuesto, el estadístico debe abstenerse de utilizar diseños balanceados o estimadores de calibración, o incluso diseños de muestreo proporcionales al tamaño de la característica de información auxiliar. En este caso, es una mejor opción confiar en el principio de representatividad que involucra cualquier diseño de muestreo. En los anteriores casos, si se quiere ganar eficiencia es mejor utilizar un estimador de expansión sencillo con un diseño de muestreo simple que combinarlo con un diseño de muestreo balanceado.

Agradecimientos

Agradezco a Dios por la oportunidad que me da al escribir esta disertación empírica como una prueba factible de que en la práctica, el estadístico no debe dejarse confundir por términos sofisticados que poseen poco valor cuando no son bien utilizados. Además, agradezco a los dos árbitros anónimos por sus valiosos aportes al revisar y criticar la primera versión de este artículo, varios años atrás cuando fue sometido a otra revista científica.

Recibido: 19 de noviembre de 2010

Aceptado: 8 de febrero de 2011

Referencias

- Cassel, C.M., S. C. & Wretman, J. (1976), ‘Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations’, *Biometrika* **63**, 615–620.
- Chaudhuri, A. & Stenger, H. (2006), *Survey Sampling: Theory and Methods*, Springer-Verlag, New York.
- Cochran, W. G. (1977), *Sampling Techniques*, 3 edn, John Wiley.
- Deville, J. C. (2005), ‘Imputation stochastique et échantillonnage équilibre’, *Tech. Rep. Ecole Nationale de la Statistique et de l’Analyse de l’Information*.
- Deville, J. C. & Särndal, C. E. (1992), ‘Calibration Estimators in Survey Sampling’, *Journal of the American Statistical Association* **87**, 376–382.
- Deville, J. C. & Tillé, Y. (2004), ‘Efficient Balanced Sampling: The Cube Method’, *Biometrika* **91**, 893–912.
- Deville, J. C. & Tillé, Y. (2005), ‘Variance Approximation under Balanced Sampling’, *Journal of Statistical Planning and Inference* **128**(2), 569–591.
- Gutiérrez, H. A. (2009), *Estrategias de muestreo: diseño de encuestas y estimación de parámetros*, Universidad Santo Tomás.
- Gutiérrez, H. A. (2010a), ‘El concepto de representatividad en la escogencia de la mejor estrategia de muestreo’, *IB Revista de la Información Básica* **4**.
- Gutiérrez, H. A. (2010b), *TeachingSampling: Sampling Designs and Parameter Estimation in Finite Population*. R package version 1.7.9.
*<http://CRAN.R-project.org/package=TeachingSampling>
- Hájek, J. & Dupac, V. (1981), *Sampling from a Finite Population*, Statistics, textbooks and monographs, v. 37, M. Dekker.

- Hansen, M., Hurwitz, W. N. & Madow, W. G. (1953), *Sample Survey Methods and Theory*, John Wiley and Sons, New York.
- Horvitz, D. G. & Thompson, D. J. (1952), ‘A Generalization of Sampling Without Replacement from a Finite Universe’, *Journal of the American Statistical Association* **47**, 663 – 685.
- Kish, L. (1995), *Survey Sampling (Wiley Classics Library)*, Wiley-Interscience.
- R Development Core Team (2010), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
- Särndal, C.-E. (2007), ‘The Calibration Approach in Survey Theory and Practice’, *Survey Methodology* **33**(2), 99 – 119.
- Särndal, C. E., Swensson, B. & Wretman, J. (2003), *Model Assisted Survey Sampling*, Springer-Verlag.
- Tillé, Y. (2006), *Sampling Algorithms*, Springer-Verlag, New York.
- Tillé, Y. & Anne-Catherine, F. (2005), ‘Optimal Allocation in Balanced Sampling’, *Statistics and Probability Letters* **74**(1), 31–37.
- Tillé, Y. & Matei, A. (2011), *Sampling: Survey Sampling*. R package version 2.4. *<http://CRAN.R-project.org/package=sampling>
- Wu, C. (2003), ‘Optimal Calibration Estimators in Survey Sampling’, *Journal of the American Statistical Association* **77**, 89–96.
- Wu, C. & Luan, Y. (2003), ‘Optimal Calibration Estimators under two-Phase Sampling’, *Journal of Official Statistics* **19**(2), 119–131.

A. Apéndice

A.1. Programación de la simulación

En este apartado se presenta la programación de la simulación concerniente a la sección 4.

```
#####
## Comparación de algunas estrategias de muestreo ##
#####

library(sampling)
library(TeachingSampling)

#-----
sigma <- 17      ## Entre más pequeño sigma hay más correlación !!
```

```

## sigma = 0.1, 17, 120,
nsim <- 1000    ## Número de simulaciones
N <- 500       ## Tamaño poblacional
n <- 50        ## Tamaño de la muestra
               ## n = 10, 50, 200

#-----
x <- rgamma(N,1,2)
e <- rnorm(N)
y <- 1000+(30*x)+sigma*x*e
Ysum <- sum(y)
Xsum <- sum(x)

cor(y,x)

bias <- rep(0,5)
mse <- rep(0,5)

#####
###      La simulación empieza desde acá      ###
#####
for(m in 1:nsim){

## Muestra no balanceada !!

res <- S.piPS(n,x)
Pik_1 <- res[,2]
sam_1 <- res[,1]
XX_1 <- x[sam_1]
YY_1 <- y[sam_1]

#=====
# Cálculo del estimador de Horvitz - Thompson
#-----

YHT_1 <- HT(YY_1, Pik_1)
bias[1] <- bias[1]+Ysum-YHT_1
mse[1] <- mse[1]+(Ysum-YHT_1)^2

#=====
# Cálculo del estimador de calibración
#-----

Sigma_1 <- solve(diag(Pik_1*XX_1))
betas_1 <- solve(t(XX_1)%*%Sigma_1*%*XX_1,t(XX_1)%*%Sigma_1*%*YY_1)
XHT_1 <- HT(XX_1, Pik_1)

```

```

Ycal_1 <- YHT_1+as.numeric(betas_1)*(Xsum-XHT_1)
bias[2] <- bias[2]+Ysum-Ycal_1
mse[2] <- mse[2]+(Ysum-Ycal_1)^2

## Muestra balanceada !!

pikpps <- PikPPS(n,x)
sam_2 <- as.double(samplecube(x,pikpps))
XX_2 <- x[sam_2==1]
YY_2 <- y[sam_2==1]
Pik_2 <- pikpps[sam_2==1]

#=====
# Cálculo del estimador de Horvitz - Thompson
#-----

YHT_2 <- HT(YY_2, Pik_2)
bias[3] <- bias[3]+Ysum-YHT_2
mse[3] <- mse[3]+(Ysum - YHT_2)^2

#=====
# Cálculo del estimador de calibración
#-----

Sigma_2 <- solve(diag(Pik_2*XX_2))
betas_2 <- solve(t(XX_2)%*%Sigma_2*%*%XX_2,t(XX_2)%*%Sigma_2*%*%YY_2)
XHT_2 <- HT(XX_2, Pik_2)
Ycal_2 <- YHT_2+as.numeric(betas_2)*(Xsum-XHT_2)
bias[4] <- bias[4]+Ysum-Ycal_2
mse[4] <- mse[4]+(Ysum-Ycal_2)^2

## Muestra aleatoria simple !!

sam_3 <- sample(N,n)
XX_3 <- x[sam_3]
YY_3 <- y[sam_3]

#=====
# Cálculo del estimador de expansión
#-----

YHT_3 <- (N/n)*sum(YY_3)
bias[5] <- bias[5]+Ysum-YHT_3
mse[5] <- mse[5]+(Ysum-YHT_3)^2

#-----

```

```
if(floor(m/100)==m/100) print(m)
}

#####
#####

bias <- bias/(Ysum*nsim)
mse <- mse/nsim
mse.rel<- mse/mse[5]

bias
round(mse.rel,3)

par(mfrow=c(1,3))
plot(x,y)
hist(y)
hist(x)
```