Marí, Gonzalo[*]
Barbará, Gabriela[**]
Mitas, Gerardo[**]
Passamonti, Sergio[**]
**Instituto de Investigaciones Teóricas y Aplicadas en Estadística, Escuela de Estadística*
***Dirección de Metodología Estadística, Instituto Nacional de Estadística y Censos*

MUESTRAS EQUILIBRADAS EN POBLACIONES FINITAS: UN ESTUDIO COMPARATIVO EN MUESTRAS DE EXPLOTACIONES AGROPECUARIAS[1]

1. INTRODUCCIÓN

En el presente trabajo, se considera el estudio y la difusión de los diseños muestrales en poblaciones finitas que invocan, en la etapa de selección, estrategias de calibración o balanceo. El objetivo del trabajo es realizar una síntesis crítica de las bondades de las técnicas que emplean información auxiliar en la etapa de selección para alcanzar una muestra equilibrada. En particular el trabajo consiste en comparar, en términos de eficiencia, el estimador de Horvitz & Thompson (HT) cuando se emplea una muestra balanceada a través del método del cubo (calibración-ante). En la discusión y la comparación se emplearán datos provenientes del Censo Nacional Agropecuario 2002 para definir los marcos de muestreo (unidades de muestreo, información auxiliar y variables a estudiar), recurriendo a simulaciones para comprobar el comportamiento del mencionado estimador.

El objetivo de este trabajo consistió en comparar las estimaciones provenientes de un diseño muestral tradicional, como el estratificado simple al azar (Cochran, 1977), y las que se obtienen por el método del cubo estratificado simple al azar (Deville y Tillé, 2004).

Las comparaciones se presentan a través de la ganancia relativa lograda en términos de las variancias, sobre el estimador de Horvitz-Thompson para totales.

2. MÉTODO DEL CUBO

Un diseño muestral balanceado tiene la propiedad de que los estimadores de Horvitz-Thompson de totales de un conjunto de variables auxiliares son iguales a los totales poblacionales de dichas variables. Esto permite una reducción en la variancia de las variables de interés, que depende de la correlación de las mismas con las variables de control

Dada una población finita U de tamaño N, se desea estimar $Y = \sum_{k \in U} y_k$. Supongamos que se cuenta con p variables auxiliares de las que se conocen sus valores para toda la población, y sea $x_k = (x_{k1}.....x_{kj}.....x_{kp})$ el valor de dichas variables para el k-ésimo individuo de la población, k = 1,....,N.

[1] Trabajo realizado en el marco del proyecto de investigación "Diseño de Muestras Balanceadas en Poblaciones Finitas". 1ECO54. Secretaría de Ciencia y Tecnología. Universidad Nacional de Rosario.

Un diseño muestral p(s) se dice balanceado sobre las variables auxiliares $x_1,\ldots, x_p$, si satisface las ecuaciones de balanceo

$$\sum_{k\in s} \frac{x_{kj}}{\pi_k} = \sum_{k\in U} x_{kj} \qquad (2.1)$$

para toda $s \in S$ tal que $p(s) > 0$, y para todo $j = 1,\ldots,p$. De esta forma, es posible obtener para estas variables, estimaciones del error muestral nulas, o sea, var ($\hat{X}$) = 0.

Supongamos además que la población se divide en H estratos $U_1,.....,U_H$. Un diseño muestral es balanceado por estrato sobre las variables auxiliares $x_1,....., x_p$ si

$$\sum_{k\in s_h} \frac{x_{kj}}{\pi_k} = \sum_{k\in U_h} x_{kj} \qquad (2.2)$$

para h=1,…,H.

El método del Cubo es un procedimiento que permite seleccionar muestras balanceadas, con probabilidades de inclusión iguales o distintas, y sin ninguna restricción sobre la cantidad de variables auxiliares a utilizar. En este método los totales estimados por Horvitz-Thompson para las variables auxiliares son iguales o aproximadamente iguales a los totales poblacionales de ellas.

El método se basa en una representación geométrica de un diseño muestral. La muestras posibles de U pueden representarse por 2^N vectores $s = \{s_k\}$ de $\Re^N$, donde

$$s_k = \begin{cases} 1 & si\ k \in s \\ 0 & si\ k \notin s \end{cases}$$

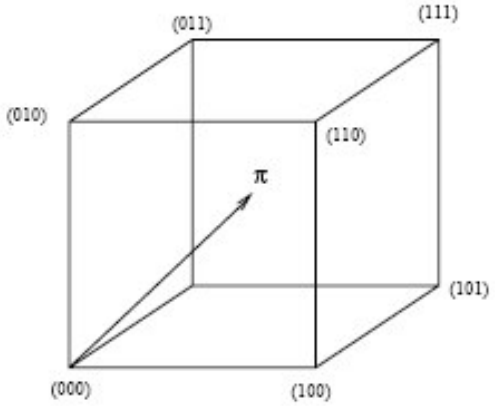
Cada vector s es un vértice de un cubo N-dimensional y el número de muestras posibles es el número de vértices del mismo. Un diseño muestral con probabilidades de inclusión π_k, $k \in U$, consiste en asignar una probabilidad p(s) a cada vértice del N-cubo que representan las muestras posibles de manera tal que

$$E(s) = \sum_{s\in S} p(s)\, s = \pi,$$

donde $\pi = \{\pi_k\}$ es el vector que contiene las probabilidades de inclusión para los N elementos de la población.

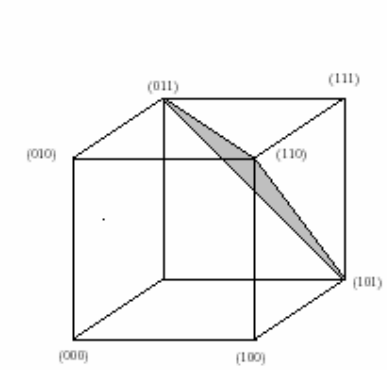
Geométricamente, un diseño muestral consiste en expresar al vector π como una combinación lineal convexa de los vértices del N-cubo. Un algoritmo muestral puede ser visto como un camino aleatorio para alcanzar un vértice del N-cubo desde un vector π de manera tal que se satisfagan las ecuaciones de balanceo (2.1).

La siguiente figura muestra la representación geométrica de todas las muestras posibles para una población con tamaño N=3

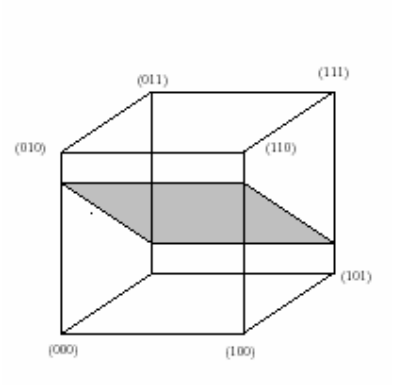


En una primera etapa, el método del cubo tiene como objetivo que las ecuaciones de balanceo se satisfagan exactamente, redondeando a 0 o 1 todas las probabilidades de inclusión. Si esto no es posible, en una segunda etapa el método trata de controlar lo mejor posible el hecho de que dichas ecuaciones no se satisfagan.

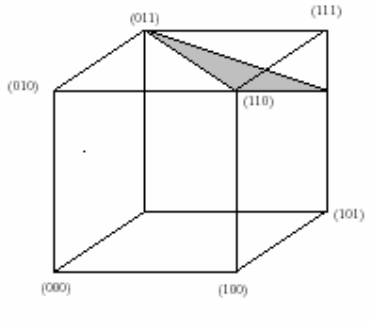
En la figura siguiente se considera un diseño muestral en una población de tamaño N=3 donde la única restricción es el tamaño de muestra n=2 y así $x_k=\pi_k$, $k\in U$. Las probabilidades de inclusión satisfacen $\pi_1+\pi_2+\pi_3=2$, y por lo tanto las ecuaciones de balanceo se satisfacen exactamente



En la siguiente figura el hiperplano de restricciones no pasa a través de ningún vértice del cubo. Las probabilidades de inclusión son $\pi_1=\pi_2=\pi_3=0.5$. La única restricción esta dada por $x_1=0$, $x_2=6\pi_2$ y $x_3=4\pi_3$. Es imposible satisfacer exactamente las ecuaciones de balanceo y por lo tanto estas se cumplen aproximadamente.



Por último, la siguiente figura muestra el caso en que el hiperplano de restricciones pasa a través de dos vértices del cubo pero uno de los vértices de la intersección no es un vértice del cubo. Las probabilidades de inclusión son $\pi_1=\pi_2=\pi_3=0.8$ y la única restricción está dada por la variable auxiliar $x_1=\pi_1$, $x_2=3\pi_2$ y $x_3=\pi_3$. En este caso, la ecuación de balanceo se satisface para algunas muestras. Existen muestras balanceadas pero no existe un diseño muestral balanceado exactamente para las probabilidades de inclusión dadas, es decir, aunque existan muestras balanceadas uno debe aceptar seleccionar muestras aproximadamente balanceadas para satisfacer las probabilidades de inclusión dadas.



En el caso de muestreo estratificado, el método consiste primero en balancear sobre las variables auxiliares independientemente en cada estrato.

Cuando no es posible balancear por estrato, se agrupan las unidades que no han sido muestreadas o rechazadas durante la primera etapa en el estrato y entonces se lleva a cabo nuevamente la primera etapa del método sobre todas estas unidades antes de comenzar la segunda etapa.

3. SIMULACIÓN

Para evaluar la utilización del método del cubo para la selección de muestras provenientes de poblaciones finitas, se empleó como marco de muestreo un conjunto de establecimientos agropecuarios (EAP's) de una provincia agrícola ganadera. El universo de unidades (N=12083) fue estratificado en 5 estratos después de tratar dos estratificaciones marginales que constaban de 4 estratos cada una. Las variables de estratificación utilizadas son: superficie y Cantidad de Trabajadores de la EAP.

Para la determinación de los límites en cada estratificación marginal se empleó el método de Horgan (Gunning et al., 2004). Esta estrategia recurre a una progresión geométrica para

determinar los límites de los estratos, bajo el supuesto de obtener coeficientes de variación (CV) aproximadamente iguales en cada uno de los estratos .

Los límites para la variable Superficie son [50ha,200ha],(200ha,700ha],(700ha,2600ha], (2600ha, ∞) y para Cantidad de Trabajadores, [1,3],(3,8],(8,25], (25, ∞). La estratificación definitiva quedó definida al combinar los intervalos de ambas variables:

Tabla 1. Estratos de EAPs de una provincia Agrícola Ganadera

Estrato	Superficie	Cantidad de Trabajadores	N_h	n_h
1	[50ha,200ha]	[1,3]	2994	30
2	(200ha,700ha]	[1,3]	4846	100
3	(700ha,∞)	[1,3]	2376	600
4	(50hs,2600ha]	(3,∞)	1615	300
5	(2600ha,∞)	(3,∞)	391	391

Como se observa en la Tabla 1, se fijó al estrato 5 como un estrato autorrepresentado. El tamaño de la muestra definido en los estratos no autorrepresentados fue de 1030 EAP's (n=1030) realizando la asignación de éste tamaño en cada estrato a través del método de Neymann.

Las variables que se emplearon como control en el método del cubo fueron 5:

- Superficie Total de la EAP en has.
- Superficie Implantada de Oleaginosas
- Superficie Implantada Total
- Superficie dedicada a la Ganadería
- Cantidad de Trabajadores.

Se estimaron los siguientes totales:

- Superficie Implantada de:
  - Girasol
  - Trigo
  - Soja 1ra
  - Soja 2da
  - Avena
  - Maíz
  - Forrajeras
  - Forrajeras Perennes
- Superficie apta para Cultivo
- Superficie no apta para Cultivo

- Cantidad de:
    - Bovinos para cría
    - Bovinos para Tambo
    - Porcinos.

Para evaluar el comportamiento del estimador $\hat{t}_{HT}$ que origina el método del cubo, se seleccionaron 10000 muestras usando este procedimiento con las variables de control señaladas en los párrafos anteriores. En cada una de las muestras se calcularon los 13 totales. Por otro lado se estimaron los totales correspondientes a las variables de control, que en teoría deberían ser aproximadamente iguales a los poblacionales en virtud del método del cubo. Los mismos cálculos fueron realizados a partir de un diseño estratificado simple al azar considerando los tamaños muestrales antes descriptos.

Se consideró la variabilidad de los 10000 totales como estimación de la variancia del estimador correspondiente, computándose adicionalmente la ganancia relativa correspondiente

Tabla 2. Variancias y Ganancias Relativas de Variables de Control y no controladas para muestras balanceadas (Método del Cubo) y sin balancear (MESA)

	Variable	Variancia		Ganancia relativa
		Método del Cubo	Muestreo Estratificado Simple al Azar	
Variables de Control	Oleaginosas	51868041	8147662720	**99.36**
	Superficie Ganadera	114083789	16554391325	**99.31**
	Superficie Implanada	107205909	12285942995	**99.13**
	Superficie total	180545714	14801078507	**98.78**
	Cantidad de Trabaj.	6274	295713.6521	**97.88**
Variables no controladas	Soja1ra	1442556264	4593460734	68.60
	Forrajeras Perennes	785798557	2318363661	66.11
	Bovinos para Cría	7140404278	20365117662	64.94
	Forrajeras	866248069	2074263159	58.24
	Bovinos para Tambo	4060828518	7838017117	48.19
	Trigo	1287963803	2395705185	46.24
	Soja2da	1382722849	2558817907	45.96
	Girasol	104716282	151289615.6	30.78
	Avena	279966917	387741606.9	27.80
	Superficie no apta	378250069	523514288.9	27.75
	Maíz	141680006	179470451.1	21.06

	Superficie apta	317358055	381255175.2	16.76
	Porcinos	5972095970	6246960962	4.40

4. CONCLUSIONES

Los resultados hallados en la tabla 2, que surgen del proceso de simulación descripto en el punto 3 resultan ser los esperados. Por un lado, las variables que fueron utilizadas como control en el método de cubo, resultaron con una variancia muy pequeña, siendo la ganancia relativa de las variancias de los totales estimados cercana al 100%.

En el resto de las estimaciones, las correspondientes a características que no fueron utilizadas como control, las ganancias resultaron ser variables, obteniéndose porcentajes que van desde el 5% hasta el 70% aproximadamente. Esto se debe a las distintos grados de relación existente entre las variables estimadas y las utilizadas para control en el método de cubo. En el caso de los totales provenientes de características poco correlacionadas con las variables de control, la ganancia no resulta ser grande.

En estudios futuros se intentará aplicar éstos métodos de selección de muestras balanceadas a diseños muestrales en varias etapas, con distintas probabilidades de inclusión, y se continuará la búsqueda de estimadores de variancia que resulten ser asintóticamente correctos.

5. BIBLIOGRAFÍA

Cochran, W.G., (1977). *Sampling Techniques*, *3rd Edition*, New York: Wiley.

Deville, J.-C., Tillé, Y., (2004). Efficient balanced sampling: The Cube Method. *Biometrika*; 91 (4), pp. 893-912.

Gunning, P., Horgan, J.M., (2004). A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations. *Survey Methodology* Vol. 30, No 2, pp. 159-166. Statistics Canada, Catalogue No 12-001.

Särndal, C.E., Swenson, B., Wretman, J.H., (1992). *Model Assisted Survey Sampling*, New York, Springer-Verlag.

Tillé, Y. (2001). *Théorie des Sondages: échantillonnage et estimation en populations finies*. Dunod, Paris

Tillé,Y., Favre, A.-C. (2004). Coordination, combination and extension of balanced samples. *Biometrika*; 91 (4), pp. 913 -928