Metodología de Encuestas
Vol 4, Núm 1, 2002, 87-104

ISSN: 1575-7803

INTRODUCCIÓN A LOS MODELOS DE SUPERPOBLACIÓN EN LAS TÉCNICAS DE MUESTREO CON PROBABILIDADES DESIGUALES

Gonzalo Sánchez-Crespo
Instituto Nacional de Estadística

RESUMEN

Los modelos de superpoblación pueden concebirse como estrategias para la cuantificación de la representatividad del error cuadrático medio de una estimación. La importancia que tiene el estudio de la representatividad en los entes estadísticos hace aconsejable utilizar los modelos de superpoblación en lugar de los clásicos, para las investigaciones por muestreo. En este trabajo se presentan los modelos de superpoblación orientados al muestreo con probabilidades de selección desiguales.

Palabras clave: *modelos de superpoblación, muestreo con probabilidades desiguales, reposición parcial, POSDEM.*

Introducción

En este artículo vamos a realizar una introducción sobre la forma de evaluar distintas técnicas de probabilidades desiguales al tiempo que mostraremos la necesidad de utilizar modelos de superpoblación al diseñar una investigación por muestreo, para cuantificar la representatividad de la varianza o del error cuadrático medio de una estimación y poder elegir el método de muestreo que mejor se adapte a una determinada investigación.

La intuición nos hace desconfiar de la estimación obtenida del error de muestreo (Error cuadrático Medio *ECM*) con una única muestra, entre otros motivos porque la estimación del *ECM* tiene a su vez un error de estimación. Para ver esto de forma sencilla, pensemos en que si, por casualidad, en una determinada investigación se obtiene una muestra completamente homogénea, su error de muestreo, calculado con los datos obtenidos de esa muestra concreta, será cero. Esto nos alerta sobre los procedimientos de cálculo del error de muestreo utilizando métodos basados únicamente en el contenido de una sola muestra, sabiendo que, en este caso, el problema tampoco se resuelve utilizando técnicas de remuestreo. Por otro lado, también tendremos que preguntar por la representatividad de la estimación del error de muestreo obtenida cuando podemos acceder a formar todo el espacio muestral en base a una única población. En este caso se deja de considerar que la población marco está sujeta a variaciones o errores aleatorios. Por ello, necesitamos calcular el *ECM* teniendo en cuenta su heterogeneidad y la posible variabilidad aleatoria de la población, esto es sus posibles "cambios aleatorios". De ahí que obtengamos un modelo que define y que genera poblaciones con lo que podríamos llamar el molde de las poblaciones que se pueden generar de una determinada clase (que por comodidad llamaremos superpoblación pero que debiera llamarse suprapoblación). Con este modelo sí puede calcularse el *ECM* y su estabilidad. Esto nos proporciona una cota superior de error que incluye ambos conceptos y que nos permite tener un indicador de la representatividad de determinadas estimaciones obtenidas con determinados métodos de selección y estimación.

Esto hay que acompañarlo con que, además, existen métodos de muestreo especialmente sensibles a los cambios en la variabilidad señalada y a los cambios en los supuestos en los que se basan. En ellos, aplicar un esquema de superpoblación permite disponer de estimaciones más robustas, eligiendo el más idóneo. Para ilustrar estos conceptos y definir alguno de los indicadores propuestos, se va a desarrollar este tema utilizando dos ejemplos hipotéticos.

Por último, se presenta abierta una vía de investigación para la formulación teórica de los resultados obtenidos mediante procesos de simulación, referentes al valor esperado respecto del modelo de superpoblación de la varianza del estimador del *ECM*.

Evaluación de métodos de muestreo

Para evaluar los métodos de muestreo, utilizaremos un modelo de superpoblación adaptado a la población que queremos investigar. Con ello, se evita obtener resultados poco robustos o incluso anecdóticos. En el caso de los estimadores utilizados en las

ilustraciones de este artículo, al ser insesgados coinciden los conceptos de error cuadrático medio del estimador y varianza del estimador, por lo que utilizaremos ambos términos:

$$ECM\left(\hat{\bar{x}}\right) = E\left(\hat{\bar{x}} - \bar{\bar{X}}\right)^2$$

El programa POSDEM (Sánchez-Crespo y Lezcano, 1999; Mateo, 2000) obtiene la estructura de la población marco ajustando los datos con el método de los polinomios ortogonales. Posteriormente es posible generar poblaciones definidas por el modelo y calcular el *ECM* para cada población y especificaciones del método de muestreo considerado. La media del *ECM* calculada sobre el conjunto de las poblaciones generadas, se puede considerar como una aproximación del valor esperado del *ECM* bajo un enfoque de superpoblación.

$$E^*\left[ECM\left(\hat{\bar{x}}\right)\right] \cong \frac{\sum_{g=1}^{G} ECM\left(\hat{\bar{x}}\right)_g}{G}$$

donde $g = 1,2,...G$ representa el conjunto de poblaciones finitas generadas con el modelo.

Este procedimiento de evaluación está en relación con el trabajo de Bellhouse y Rao (1975). Los resultados obtenidos para modelos con grados uno y dos son coincidentes con los derivados teóricamente por éstos.

La varianza sobre el modelo se considera la medida de la acuracidad del *ECM*. Viene dada por la expresión:

$$V^*\left[ECM\left(\hat{\bar{x}}\right)\right] \cong \frac{\sum_{g=1}^{G}\left(ECM\left(\hat{\bar{x}}\right)_g - E^*\left[ECM\left(\hat{\bar{x}}\right)\right]\right)^2}{G}$$

Para evaluar los métodos de muestreo, utilizaremos un límite superior de confianza que tiene en cuenta el valor esperado del *ECM*

$$E^*\left[ECM\left(\hat{\bar{x}}\right)\right]$$

y su desviación

$$\sqrt{V^*\left[ECM\left(\hat{\bar{x}}\right)\right]}$$

Es decir:

$$E^*\left[ECM\left(\hat{\bar{x}}\right)\right] + 1{,}96\sqrt{V^*\left[ECM\left(\hat{\bar{x}}\right)\right]}$$

que permite disponer de un límite de confianza que contendrá el *ECM*, al menos en un 95 de cada 100 poblaciones generadas o potencialmente posibles de encontrar en el trabajo de campo.

Muestreo con probabilidades proporcionales al tamaño

Consideremos una población compuesta por N unidades, que queremos investigar mediante una muestra. Representaremos cada unidad por u_i, con i=1,2,...N. Si dentro de cada una de estas unidades, interesa medir una característica desconocida, X_i , que tiene asociada otra característica conocida, M_i , podemos utilizar la información auxiliar para mejorar el diseño de la investigación.

Para este caso ilustrativo, consideramos una agencia que debe inspeccionar cuatro expedientes de gasto, para conocer cuál es la cantidad de dinero que se ha gastado indebidamente. Para esta labor inspectora, debido a serios recortes presupuestarios y dado que el trabajo debe ser terminado en un breve espacio de tiempo, la agencia no dispone de los medios necesarios para investigar la totalidad de los expedientes. Debido a ello únicamente podrá investigar dos de los cuatro mencionados. Existen distintas formas de elegir qué unidades deben ser investigadas. En función del método de selección que se elija, el estudio será más o menos preciso. Por tanto el objetivo inicial es elegir entre distintos métodos de selección de muestras, de forma que el error cometido —debido a que no se cubre toda la población sino sólo una parte—, sea el menor posible.

Para responder a esta cuestión sobre qué método de selección es preferible, se plantea un escenario simulado. Para ello, suponemos que la información buscada, X_i, está ya disponible, y que los datos de los cuatro expedientes son los que figuran en el cuadro 1.

X_i	1	2	3	4
M_i	1	1	2	2

Cuadro 1: *relación entre unidades y tamaños.*

donde X_i representa el número de millones que se han gastado indebidamente en cada uno de esos expedientes y M_i representa una variable de clasificación según el tamaño del gasto total del expediente i-ésimo. Tanto el tamaño de población como los datos de X_i y de M_i constituyen una simplificación con fines ilustrativos. En la última parte de este artículo, se hará una simulación más general.

Elegir las unidades que van a formar parte de la muestra con probabilidades proporcionales a una variable auxiliar —en este caso el tamaño de cada unidad—, es un proceso probabilístico que puede representarse con un esquema de urnas (véase figura 1): por cada unidad, se introducen en la urna tantas bolas como nos indique la variable auxiliar. Después, cada unidad es seleccionada en función de si la bola que la representa es o no extraída. Si el esquema de selección es con reposición, después de cada extracción devolvemos la bola a la urna y, por tanto, ésta permanecerá invariable. Si el esquema de selección fuese sin reposición, entonces evitamos que una unidad pueda volver a ser seleccionada, sacando de la urna todas las bolas que la representan.

Para comparar los resultados que vamos a obtener con el caso más simple de muestreo aleatorio —donde las probabilidades de selección de cada unidad son iguales—, tendremos que, según se observa en el cuadro 2, M_i=1, $\forall i$. En tal caso, los cuatro expedientes están igualmente representados por una única bola en la urna. Por tanto, la probabilidad de ser seleccionados es idéntica para cualquiera de ellos.

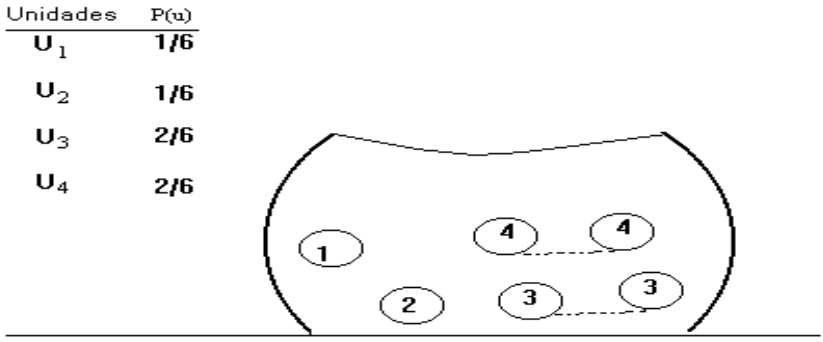

Figura 1: *esquema de urnas con selección proporcional al tamaño.*

X_i	1	2	3	4
M_i	1	1	1	1

Cuadro 2: *relación entre unidades y tamaños.*

Para este caso, se puede comprobar fácilmente que tanto la varianza del estimador como su propia varianza toman los valores que figuran en la tabla 1.

Tabla 1: *varianzas en función del modelo de selección con probabilidades iguales.*

Métodos	$V(\hat{x})$	$V\left[\hat{V}(\hat{x})\right]$
Con reposición	10,00	132,00
Sin reposición	6,67	32,89

Los cálculos realizados para obtener la varianza del estimador del total se basan en la expresión

$$V(\hat{x}) = \sum_{j=1}^{k} \frac{(\hat{x}_j - x)^2}{k}$$

Mientras que, para la varianza del estimador de la varianza, se recurre a:

$$V\left(\hat{V}(\hat{x})\right) = \sum_{j=1}^{k} \frac{\left(\hat{V}(\hat{x}_j) - V(\hat{x})\right)^2}{k}$$

siendo *k* el número de muestras posibles.

Para responder a la pregunta sobre si puede disminuir la varianza del estimador al utilizar la información auxiliar sobre el tamaño de los expedientes, cambiando las probabilidades de selección, vamos a suponer que consideramos todas las M_i=1 menos M_4=2. Los nuevos resultados se muestran en la tabla 2.

Tabla 2: *varianzas en función del modelo de selección con probabilidades proporcionales al tamaño.*

Métodos	$V(\hat{x})$	$V\left[\hat{V}(\hat{x})\right]$
Con reposición	5,00	43,75
Sin reposición	4,17	75,35

Como puede observarse, la varianza del estimador ha disminuido respecto a los procedimientos con probabilidades iguales. Además, podemos observar que, en contra de lo que ocurría en el caso de muestreo con probabilidades iguales, la varianza del estimador de la varianza es menor para el modelo con reposición que para el modelo sin reposición.

Vemos, con este caso ilustrativo, que es posible obtener una menor varianza del estimador modificando las probabilidades de selección. No obstante, si no disponemos de una información auxiliar conveniente, es posible que se presenten resultados indeseables, al partir de un modelo inadecuado. Por ejemplo, si por alguna razón contamos con tamaños asociados con cada unidad —que figuran en el cuadro 3—, se generarían los resultados que se muestran en el cuadro 3.

X_i	1	2	3	4
M_i	1	2	1	1

Cuadro 3: *unidades de tamaño desigual.*

Tabla 3: *resultados para el cuadro 3.*

Métodos	$V(\hat{x})$	$V\left[\hat{V}(\hat{x})\right]$
Con reposición	20,00	512,50
Sin reposición	12,50	384,38

Vemos como, en este caso, al utilizar probabilidades proporcionales al tamaño, los resultados han sido peores que si hubiésemos utilizado probabilidades iguales. Esto es, si la relación existente entre la variable auxiliar y la variable de estudio no es proporcional, puede ser mejor utilizar probabilidades iguales.

Muestreo con reposición y probabilidades proporcionales al tamaño

Si usamos un esquema de selección con reposición y probabilidades proporcionales al tamaño (*Cr_ppt*), la probabilidad de cada unidad de pertenecer a una muestra de n unidades es será np_i. Con ello, el estimador del total será:

$$\hat{X}_{Cr\_ppt} = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i}{p_i}$$

Consideremos el ejemplo del cuadro 1. En tal caso, las posibles muestras —junto con sus valores asociados X_i y M_i—, el estimador obtenido para el total y la probabilidad de cada muestra serán los que constan en la tabla 4.

Tabla 4: *resultados de las realizaciones muestrales para un modelo con reposición y probabilidades proporcionales al tamaño.*

U_i,U_j	X_i, X_j	M_i, M_j	$\hat{X}_{Cr\_ppt}$	P(U_i,U_j)
U_1, U_1	1, 1	1, 1	6,00	1/36
U_1, U_2	1, 2	1, 1	9, 00	1/18
U_1, U_3	1, 3	1, 2	7,50	1/9
U_1, U_4	1, 4	1, 2	9,00	1/9
U_2, U_2	2, 2	1, 1	12,00	1/36
U_2, U_3	2, 3	1, 2	10,50	1/9
U_2, U_4	2, 4	1, 2	12,00	1/9
U_3, U_3	3, 3	2, 2	9,00	1/9
U_3, U_4	3, 4	2, 2	10,50	2/9
U_4, U_4	4, 4	2, 2	12,00	1/9
	Total			1

La varianza del estimador del total viene dada por la expresión:

$$V\left(\hat{X}_{Cr\_ppt}\right) = \frac{1}{n}\sum_{i=1}^{N} p_i\left(\frac{X_i}{p_i} - X\right)^2$$

Así, en cada muestra podemos utilizar el estimador insesgado para la varianza dado por:

$$\hat{V}\left(\hat{X}_{Cr\_ppt}\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{X_i}{p_i} - \hat{X}_{Cr\_ppt}\right)^2$$

Cuyos resultados correspondientes constan en la tabla 5.

Tabla 5: *varianzas del estimador según la selección mediante un modelo con reposición y probabilidades iguales al tamaño.*

U_i,U_j	$\hat{V}\left(\hat{X}_{Cr\_ppt}\right)$	P(U_i,U_j)
U_1, U_1	0,00	1/36
U_1, U_2	9,00	1/18
U_1, U_3	2,25	1/9
U_1, U_4	9,00	1/9
U_2, U_2	0,00	1/36
U_2, U_3	2,25	1/9
U_2, U_4	0,00	1/9
U_3, U_3	0,00	1/9
U_3, U_4	2,25	2/9
U_4, U_4	0,00	1/9
	Total	1

Sobre el conjunto de todas las muestras posibles tenemos los siguientes resultados:

$$E\left(\hat{X}_{Cr\_ppt}\right) = X \qquad E\left[\hat{V}\left(\hat{X}_{Cr\_ppt}\right)\right] = V\left(\hat{X}_{Cr\_ppt}\right)$$

La tabla 6 muestra estas operaciones en formato de hoja de cálculo. En ella vemos, como principales resultados, que $V\left(\hat{X}_{Cr\_ppt}\right)$=2,5 y que $V\left[\hat{V}\left(\hat{X}_{Cr\_ppt}\right)\right]$= 9,5.

Tabla 6: *resultados intermedios en formato de hoja de cálculo.*

A, B	C, D	E	F	G	H=F*E	I= E*(F-ΣH)^2	J=G*E	K=E*(G-ΣJ)^2
X_i, X_j	M_i, M_j	p(U_i, U_j)	$\hat{X}_{Cr}$	$\hat{V}\left(\hat{X}_{Cr}\right)$	$E\left(\hat{X}_{Cr}\right)$	$V\left(\hat{X}_{Cr}\right)$	$E\left[\hat{V}\left(\hat{X}_{Cr}\right)\right]$	$V\left[\hat{V}\left(\hat{X}_{Cr}\right)\right]$
1, 1	1, 1	1/36	6,00	0,00	0,17	0,44	0,00	0,17
1, 2	1, 1	1/18	9,00	9,00	0,50	0,06	0,50	2,35
1, 3	1, 2	1/9	7,50	2,25	0,83	0,69	0,25	0,01
1, 4	1, 2	1/9	9,00	9,00	1,00	0,11	1,00	4,69
2, 2	1, 1	1/36	12,00	0,00	0,33	0,11	0,00	0,17
2, 3	1, 2	1/9	10,50	2,25	1,17	0,03	0,25	0,01
2, 4	1, 2	1/9	12,00	0,00	1,33	0,44	0,00	0,69
3, 3	2, 2	1/9	9,00	0,00	1,00	0,11	0,00	0,69
3, 4	2, 2	2/9	10,50	2,25	2,33	0,06	0,50	0,01
4, 4	2, 2	1/9	12,00	0,00	1,33	0,44	0,00	0,69
Total		1,00			10,00	2,50	2,50	9,50

Muestreo sin reposición y probabilidades proporcionales al tamaño

Ahora, si usamos un esquema de muestreo sin reposición y probabilidades proporcionales al tamaño, con las condiciones de Brewer, tendremos

$$\hat{X}_{Sr\_ppt} = \sum_{i}^{n} \frac{X_i}{np_i} \qquad \text{con} \qquad E\left[\hat{X}_{Sr\_ppt}\right] = X$$

Con ello, la varianza del estimador del total es

$$V\left[\hat{X}_{Sr\_ppt}\right] = \sum_{i}^{N}\sum_{j>i}\left(p_i p_j - p_{ij}\right)\left(\frac{X_i}{p_i} - \frac{X_j}{p_j}\right)^2$$

Y el estimador insesgado para la varianza con muestras de tamaño dos:

$$\hat{V}\left[\hat{X}_{Sr\_ppt}\right] = \sum_{i}^{2}\sum_{j>i}\frac{\left(p_i p_j - p_{ij}\right)}{p_{ij}}\left(\frac{X_i}{p_i} - \frac{X_j}{p_j}\right)^2$$

donde $p_{ij} = \dfrac{2p_i p_j}{D}\dfrac{1-\left(p_i+p_j\right)}{\left(1-2p_i\right)\left(1-2p_j\right)}$ $\qquad p_i = \dfrac{M_i}{M}$ $\qquad D = \dfrac{1}{2}\left(1+\sum_{i}^{N}\dfrac{p_i}{1-2p_j}\right)$

En el ejemplo que estamos siguiendo las muestras posibles junto con sus correspondientes cálculos y estimaciones vienen dadas por la tabla 7, donde vemos, como principales resultados, que

$$V\left[\hat{X}_{Sr\_ppt}\right] = 1{,}79 \text{ y que } V\left(\hat{V}\left[\hat{X}_{Sr\_ppt}\right]\right) = 7{,}74$$

Tabla 7: *resultados intermedios para el modelo de muestreo sin reposición, con probabilidades proporcionales al tamaño.*

A, B	C, D	E	F	G	H=F*E	I= E*(F-ΣH)^2	J=G*E	K=E*(G-ΣJ)^2
X_i, X_j	M_i, M_j	p_i , p_j, $p(U_i,U_j)$	$\hat{X}_{Cr}$	$\hat{V}(\hat{X}_{Cr})$	$E(\hat{X}_{Cr})$	$V(\hat{X}_{Cr})$	$E\left[\hat{V}(\hat{X}_{Cr})\right]$	$V\left[\hat{V}(\hat{X}_{Cr})\right]$
1, 2	1, 1	1/3 1/3 1/7	9,00	12,00	0,43	0,05	0,57	4,97
1, 3	1, 2	1/3 2/3 1/7	7,50	1,25	1,07	0,89	0,18	0,04
1, 4	1, 2	1/3 2/3 1/7	9,00	5,00	1,29	0,14	0,71	,48
2, 3	1, 2	1/3 2/3 1/7	10,50	1,25	1,50	0,04	0,18	0,04
2, 4	1, 2	1/2 2/3 1/7	12,00	0,00	1,71	0,57	0,00	0,46
3, 4	2, 2	2/3 2/3 3/8	10,50	0,38	4,00	0,10	0,14	0,76
Total		1,00			10,00	1,79	1,79	7,74

El principal inconveniente de este método es que, en ciertos casos, el estimador de la varianza puede ser indeterminado debido a que p_{ij} puede tomar el valor cero. Además, pueden presentarse estructuras de población que hagan que el estimador de la varianza tome valores negativos, lo que no tiene sentido. Como ejemplo, puede comprobarse que para los valores: M_i = 1,1,1,3 en el caso que estamos desarrollando, tendremos una varianza indeterminada y para M_i = 1,1,1,6 estimaciones de varianza negativas.

Muestreo con reposición parcial y probabilidades proporcionales al tamaño

En este caso, se puede hacer la selección muestral con el siguiente esquema: en la primera selección se utiliza la medida original del tamaño M_i, suponiendo que la unidad U_i es seleccionada con probabilidad:

$$p_i = \frac{M_i}{M} \quad \text{donde} \quad M = \sum_i^N M_i$$

Para la segunda selección, se utiliza la medida reducida M_i -b, donde la constante b se define como:

$$b = \left[\frac{\min M_i}{n-1}\right]^* = \left[\frac{Mo}{n-1}\right]$$

donde n es el tamaño de muestra y $[\ ]^*$ representa el entero más próximo.

La probabilidad de la unidad U_i en la primera selección es

$$p\left(\frac{u_i}{1^{\text{a}}\text{selección}}\right) = \frac{M_i}{M} = p_i$$

Para la segunda selección, tenemos para la probabilidad incondicional de u_i:

$$p\left(\frac{u_i}{u_j \neq 1^{\text{a}}\text{selección}}\right) + p\left(\frac{u_i}{u_i \text{ en la } 1^{\text{a}}\text{selección}}\right) =$$

$$= \frac{M - M_i}{M}\frac{M_i}{M-b} + \frac{M_i}{M}\frac{M_i - b}{M-b} = \frac{M \cdot M_i - {M_i}^2 + {M_i}^2 - bM_i}{M(M-b)} =$$

$$= \frac{M_i(M-b)}{M(M-b)} = p_i$$

Con ello, la probabilidad de que la unidad u_i pertenezca a la muestra de tamaño dos es $2p_i$. Así pues, el estimador para el total es

$$\hat{X}_{Crp} = \sum_{i}^{n} \frac{X_i}{np_i}$$

Su varianza

$$V\left(\hat{X}_{Crp}\right) = \frac{M-nb}{M-b}\frac{1}{n}\sum_{i}^{N} p_i\left(\frac{X_i}{p_i} - X\right)^2$$

Finalmente, un estimador insesgado de la expresión anterior para n=2, es

$$\hat{V}\left(\hat{X}_{Crp}\right) = \frac{M-2b}{M-b}\frac{1}{4}\sum_{i}^{N}\left(\frac{X_1}{p_1} - \frac{X_2}{p_2}\right)^2$$

Los resultados correspondientes a los datos del ejemplo se muestran en la tabla 8, en donde vemos, como principales resultados, que

$$V\left[\hat{X}_{Crp}\right] = 2{,}0 \text{ y que } V\left(\hat{V}\left[\hat{X}_{Crp}\right]\right) = 4{,}4$$

Tabla 8: *resultados intermedios para el modelo de muestreo con reposición parcial.*

A, B	C, D	E	F	G	H=F*E	I= E*(F-ΣH)^2	J=G*E	K=E*(G-ΣJ)^2
X_i, X_j	M_i, M_j	$p(U_i,U_j)$	$\hat{X}_{Cr}$	$\hat{V}\left(\hat{X}_{Cr}\right)$	$E\left(\hat{X}_{Cr}\right)$	$V\left(\hat{X}_{Cr}\right)$	$E\left[\hat{V}\left(\hat{X}_{Cr}\right)\right]$	$V\left[\hat{V}\left(\hat{X}_{Cr}\right)\right]$
1, 2	1, 1	1/30	9,00	6,00	0,30	0,03	0,20	0,53
1, 3	1, 2	1/15	7,50	1,50	0,50	0,42	0,10	0,02
1, 4	1, 2	1/15	9,00	6,00	0,60	0,07	0,40	1,07
2, 1	1, 1	1/30	9,00	6,00	0,30	0,03	0,20	0,53
2, 2	1, 1	0	12,00	0,00	0,00	0,00	0,00	0,00
2, 3	1, 2	1/15	10,50	1,50	0,70	0,02	0,10	0,02
2, 4	1, 2	1/15	12,00	0,00	0,80	0,27	0,00	0,27
3, 1	2, 1	1/15	7,50	1,50	0,50	0,42	0,10	0,02
3, 2	2, 1	1/15	10,50	1,50	0,70	0,02	0,10	0,02
3, 3	2, 2	1/15	9,00	0,00	0,60	0,07	0,00	0,27
3, 4	2, 2	2/15	10,50	1,50	1,40	0,03	0,20	0,03
4, 1	2, 1	1/15	9,00	6,00	0,60	0,07	0,40	1,07
4, 2	2, 1	1/15	12,00	0,00	0,80	0,27	0,00	0,27
4, 3	2, 2	2/15	10,50	1,50	1,40	0,03	0,20	0,03
4, 4	2, 2	1/15	12,00	0,00	0,80	0,27	0,00	0,27
Total		1,00	150,0		10,00	2,00	2,00	4,40

En resumen y considerando los tres métodos en conjunto, cuyos resultados obtenidos hasta el momento se encuentran en la tabla 9, se puede concluir que la varianza del estimador del total es más pequeña para *Sr_ppt*. Sin embargo, si observamos la varianza del estimador de la varianza, entonces el mejor resultado corresponde a *Crp_ppt*. Esto es

consecuencia de que el método *Sr_ppt* es más inestable al estimar la varianza. Si comparamos los valores obtenidos al aplicar la expresión *media más tres veces la desviación típica*, con el máximo valor que toma el estimador de la varianza, podemos observar el efecto descrito de comportamiento errático del estimador de la varianza para el método *Sr_ppt*.

Tabla 8: *comparación entre los tres métodos.*

Métodos	$\hat{V}(\hat{x})$	$V\left[\hat{V}(\hat{x})\right]$	Máximo $\hat{V}(\hat{x})$	$\hat{V}(\hat{x})+3\sqrt{V\left[\hat{V}(\hat{x})\right]}$
Cr_ppt	2,5	9,5	9	11,7
Sr_ppt	1,7	7,7	12	10,1
Crp_ppt	2,0	4,4	6	8,2

Un modelo de superpoblación sencillo

Hasta el momento, hemos supuesto que conocíamos los valores que toma X_i y que éstos eran constantes. Es posible que, en realidad, estos valores esten afectados por errores en la observación, en la medición, cambios por el paso del tiempo en ciertos casos u otras causas. La cuestión es que estos valores de X_i pueden considerarse afectados por una variabilidad aleatoria.

En nuestro ejemplo, para ilustrar que los valores de X_i son, en realidad, los resultados de una realización aleatoria, vamos a suponerlos modelados por la siguiente expresión: DISTR.NORM.INV(random();i;0.i) con i=1,2,3,4. Esta expresión proporciona unos valores aleatorios que se distribuyen según una normal de parámetros media *i* y desviación típica 0.*i*

Una realización nos proporcionaría los valores contenidos en el cuadro 4.

X_i	0,9856	2,0167	2,7025	4,0679
M_i	1	1	2	2

Cuadro 4: *realizaciones aleatorias de unidades.*

Si repetimos la generación aleatoria ocho veces, tendremos ocho poblaciones similares en cuanto a estructura pero que presentan ligeras variaciones aleatorias unas de otras. Los resultados para los métodos de selección considerados se muestran en la tabla 10.

Gráficamente, con la figura 2, podemos observar que no es posible establecer en qué medida un método es preferible a otro con el análisis de los datos de una sola población, puesto que esa medida depende de la población concreta que se analiza. Las conclusiones obtenidas en cuanto a la varianza del estimador, tras analizar las poblaciones finitas tres o cinco (PF3, PF5), serían muy diferentes. Queremos resaltar que, en este ejemplo, y a pesar de ser muy similares unas poblaciones a otras, la octava población, que es la que hemos puesto en la tabla inicial de este apartado, presenta un comportamiento diferente en cuanto a estabilidad que las otras siete poblaciones consideradas. Si nuestras conclu-

siones se basaran únicamente en esta realización concreta, nuestras conclusiones estarían afectadas por un componente anecdótico que no representa un comportamiento general.

Tabla 10: *simulación de resultados (modelo de superpoblación)*

$\hat{V}(\hat{x})$	PF1	PF2	PF3	PF4	PF5	PF6	PF7	PF8	Valores esperados
Cr_ppt	3,51	2,46	0,74	2,16	4,31	3,08	3,36	3,36	2,87
Sr_ppt	2,41	1,57	0,59	1,74	2,73	1,95	2,42	2,23	1,96
Crp_ppt	2,81	1,97	0,59	1,73	3,45	2,47	2,69	2,69	2,30
$V\left[\hat{V}(\hat{x})\right]$	PF1	PF2	PF3	PF4	PF5	PF6	PF7	PF8	Valores esperados
Cr_ppt	13,78	6,55	17,50	7,03	21,32	3,73	13,03	18,74	17,71
Sr_ppt	14,26	5,49	17,40	4,58	10,77	2,84	10,40	7,31	9,13
Crp_ppt	6,45	3,02	8,38	3,09	9,32	1,83	6,11	8,79	5,87

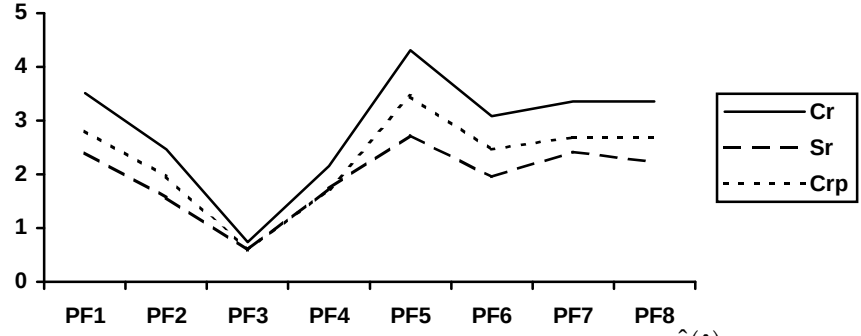

Figura 2: *representación gráfica de los valores de* $\hat{V}(\hat{x})$.

Podemos confirmar estos resultados empíricos con los obtenidos teóricamente mediante la expresión:

$$R = 1 - \frac{E^*\left[V\left(\hat{X}_{Crp\_ppt}\right)\right]}{E^*\left[V\left(\hat{X}_{Cr\_ppt}\right)\right]} = \frac{b(n-1)}{M-b} = \frac{n-1}{\frac{M}{b}-1}$$

Sustituyendo los valores de *n*, *M* y *b* se obtiene que la ganancia de *Crp* sobre *Cr*, en varianza del estimador, será igual al 20%. Al sustituir por los valores empíricos de las esperanzas respecto del modelo estaríamos en un 19,86%. Con ello, podemos considerar una confirmación de los valores teóricos casi exacta. Para la comparación entre *Sr* y *Cr* los valores obtenidos serían, respectivamente, 33% y 31,7%

Gráficamente, con la figura 3, utilizando la segunda parte de la tabla 10, podemos observar los resultados calculados para la varianza del estimador de la varianza en cada método, para el conjunto de las ocho generaciones aleatorias de la población finita utilizada como patrón.

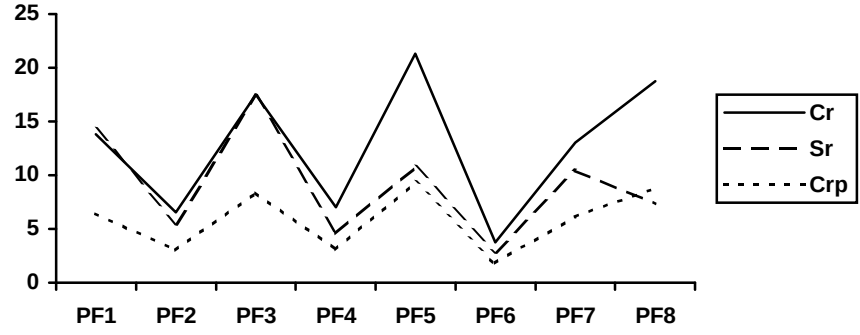

Figura 3: *representación gráfica de los valores de* $V\left[\hat{V}(\hat{x})\right]$.

El método *Crp_ppt* mejora a los otros en cuanto a estabilidad de la varianza en todas las poblaciones menos en la octava. Ésta es la población que hemos puesto de ejemplo en este trabajo para ver que no es suficiente el análisis de una sola población y que es necesario plantear un enfoque *supra*poblacional. Esta octava población "especial" se ha obtenido después de un conjunto amplio de pruebas, simplemente con el fin de mostrar que es físicamente posible.

De forma análoga al caso de varianza del estimador, si bien no está disponible un estudio teórico, podemos llevar a cabo una cuantificación de la ganancia de un método respecto de otro, pero esta vez en términos de varianza del estimador de la varianza.

Así, para *Crp* sobre *Cr*, tendremos:

$$R = 1 - \frac{E^*\left(V\left[\hat{V}\left(\hat{X}_{Crp\_ppt}\right)\right]\right)}{E^*\left(V\left[\hat{V}\left(\hat{X}_{Cr\_ppt}\right)\right]\right)} = 1 - \frac{5{,}87}{17{,}71} = 66{,}8\ \%$$

y para *Sr* sobre *Cr*, será

$$R = 1 - \frac{E^*\left(V\left[\hat{V}\left(\hat{X}_{Crp\_ppt}\right)\right]\right)}{E^*\left(V\left[\hat{V}\left(\hat{X}_{Cr\_ppt}\right)\right]\right)} = 1 - \frac{9{,}13}{17{,}71} = 48{,}4\ \%$$

Una población y un tamaño de muestra mayor

Aquí vamos a utilizar una población algo mayor que la considerada hasta ahora y realizaremos los cálculos con un programa de ordenador confeccionado a medida de estas necesidades. Un tamaño de muestra de dos unidades es debido a que tiene importantes ventajas y no supone una limitación. Pero aquí ampliaremos estas dimensiones, mediante una población de tamaño 8, y muestras de tamaño 4 con los diferentes esquemas de probabilidades desiguales estudiados. Para obtener muestras de mayor tamaño, recurrimos a considerar la población dividida en 2 grupos, estratos, de 4 unidades cada uno. En cada grupo, seleccionamos una muestra independiente de tamaño igual 2. De

esta forma, por agregación, tenemos un total de 4 unidades muestrales. Este esquema puede extenderse a cualquier tamaño de población y de muestra.

Las fórmulas de aplicación, para el estimador del total, dado que la selección es independiente en cada estrato, son:

$$\hat{X}_{st} = \sum_{h}^{L} \hat{X}_h$$

$$\hat{V}\left(\hat{X}_{st}\right) = \sum_{h}^{L} \hat{V}\left(\hat{X}_h\right)$$

Obtenidos los valores de $\hat{X}_{st}$ y de $\hat{V}\left(\hat{X}_{st}\right)$, con tamaños de muestra en cada estrato igual a dos unidades y con las fórmulas desarrolladas para los distintos esquemas de selección y estimación con probabilidades proporcionales al tamaño de los apartados anteriores, podremos obtener las estimaciones para el total y su varianza, correspondientes a tamaños de muestra superiores a dos.

Al recurrir al programa de ordenador POSDEM, podemos utilizar como población marco cualquier población finita. Por simplicidad, en el ejemplo utilizaremos la población del cuadro 5, cuyos principales resultados se encuentran en la tabla 11.

Xi	1	2	3	4	5	6	7	8
Mi	1	1	1	2	2	2	3	3

Cuadro 5: *simulación de una población.*

Tabla 11: *comparación entre los tres métodos.*

Métodos	$\hat{V}\left(\hat{X}_{st}\right)$	$V\left[\hat{V}\left(\hat{X}_{st}\right)\right]$
Cr_ppt	7,7352	62,6526
Sr_ppt	5,7139	123,9088
Crp_ppt	5,8514	21,7006

En los ejemplos anteriores hemos calculado todas las muestras posibles. Cuando la población y el tamaño de muestra son pequeños, esto no supone una gran dificultad. No obstante, cuando estos aumentan, para realizar los cálculos necesarios es más conveniente utilizar una representación lo más amplia posible del espacio muestral. En este ejemplo, hemos representado el espacio de todas las muestras que es posible obtener, con un determinado procedimiento de selección, generando un conjunto de cuatrocientas muestras con cada procedimiento para cada población finita.

Ahora, vamos a analizar este mismo resultado enfocándolo desde un modelo aleatorio de superpoblación sencillo. Se trata de evitar el posible componente anecdótico que podría estar ligado más que a la forma de la población y a su estructura básica, a ciertas combinaciones de valores. Para esto, podemos representar los valores de la población mediante la ecuación:

$$X^{*}i = 0.99 * i + e$$

donde

X^*i son los valores de la variable de estudio obtenidos ajustando una función lineal a los datos originales.

0.99 es la pendiente de la recta.

i = 1,2,...8 representa el índice de cada unidad.

e es un término aleatorio distribuido normal de media cero y desviación típica 0,1

Se recurre a la aplicación POSDEM para modelar, mediante la técnica de polinomios ortogonales, cualquier población, independientemente de la forma que tenga. Es posible también, si la forma de la población es muy compleja, obtener ecuaciones por tramos y utilizarlas conjuntamente para representar la estructura de una población.

En el ejemplo, hemos obtenido un conjunto de 200 poblaciones que siguen el mismo patrón de la población original. En cada una de ellas hemos obtenido la varianza del estimador del total. En la tabla 12 pueden observarse los resultados obtenidos en las 20 primeras realizaciones y un tamaño de muestra igual a cuatro.

Tabla 12: *resultados de la simulación con POSDEM para la varianza.*

$v(\hat{X})$	1	2	3	4	5	6	7	8	9	10
Cr_ppt	6.427	7.131	10.24	6.957	11.15	11.21	9.792	7.842	6.806	10.51
Sr_ppt	6.670	5.506	8.326	6.401	7.210	4.949	5.528	5.822	5.670	5.798
Crp_ppt	6.174	4.737	6.358	5.844	7.806	4.901	5.959	6.888	6.322	6.658

$v(\hat{X})$	11	12	13	14	15	16	17	18	19	20
Cr_ppt	9.008	6.922	9.100	6.116	6.570	8.666	7.826	8.702	4.837	7.583
Sr_ppt	6.592	3.860	7.967	5.511	5.253	6.650	4.989	6.735	6.345	6.327
Crp_ppt	5.805	4.680	7.740	5.746	6.516	6.818	5.400	5.773	6.395	5.863

Estos resultados, respecto del modelo, pueden observarse gráficamente para el conjunto de las doscientas poblaciones finitas aleatorias mediante la figura 4.

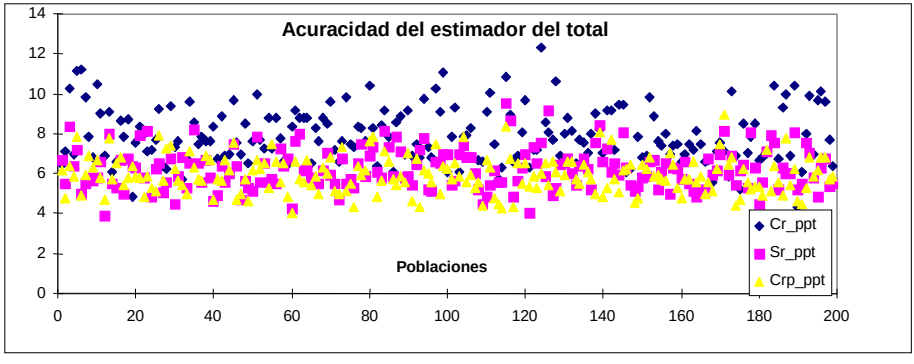

Figura 4: *valores para la acuracidad del estimador del total, en una simulación de 200 poblaciones.*

Igualmente, pueden a su vez resumirse mediante la esperanza respecto del modelo de la varianza del estimador, según consta en la tabla 13.

Tabla 13: *comparación entre los tres métodos.*

Métodos	$E^*\left[\hat{V}\left(\hat{X}_{st}\right)\right]$
Cr_ppt	7,88
Sr_ppt	6,20
Crp_ppt	5,94

Si en lugar de observar en cada población el parámetro varianza del estimador, observamos el parámetro varianza de la varianza del estimador, podremos obtener una medida de la representatividad del estimador de la varianza para cada método. Los resultados obtenidos se muestran en la tabla 14 y representados gráficamente en la figura 5.

Tabla 14: *resultados de la simulación con POSDEM para la varianza.*

$V\left(\hat{V}\right)$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Cr_ppt	51.45	39.747	40.339	78.985	52.786	70.509	74.730	42.653	72.674	39.173
Sr_ppt	94.55	66.299	140.40	200.64	85.657	114.08	216.17	61.527	128.70	115.80
Crp_ppt	20.03	18.553	22.741	40.542	23.748	30.374	40.757	19.126	28.664	14.062

$V\left(\hat{V}\right)$	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Cr_ppt	84.987	50.946	22.435	91.789	35.901	50.955	56.293	62.044	60.575	37.121
Sr_ppt	134.17	119.05	117.73	96.043	124.77	121.22	81.562	88.788	129.35	69.099
Crp_ppt	34.799	18.347	21.276	18.342	19.546	18.593	21.849	28.296	24.258	21.822

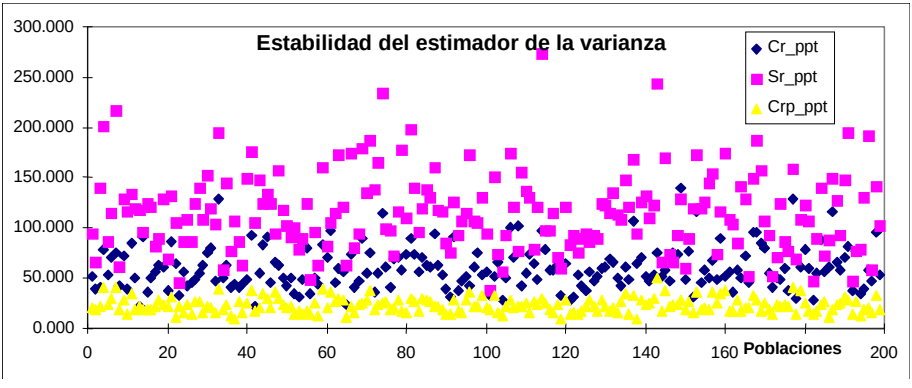

Figura 5: *valores para la estabilidad del estimador de la varianza, en una simulación de 200 poblaciones.*

El resumen de los valores esperados respecto del modelo para la varianza del estimador de la varianza de los métodos considerados se encuentra en la tabla 15.

Tabla 15: *comparación entre los tres métodos.*

Métodos	$E^*\left(V\left[\hat{V}\left(\hat{X}_{st}\right)\right]\right)$
Cr_ppt	59,93
Sr_ppt	115,02
Crp_ppt	23,05

En este ejemplo ilustramos cómo el método con reposición parcial puede presentar una ventaja en cuanto a precisión de las estimaciones respecto del método con reposición y, también, puede suponer una mejora en cuanto al método sin reposición en el sentido de que el estimador de la varianza que propone presenta ganancias en cuanto a representatividad del verdadero valor.

Conclusión

Con un ejemplo ilustrativo hemos introducido al lector en las técnicas de muestreo con probabilidades desiguales al tiempo que hemos mostrado las diferencias, ventajas e inconvenientes, de unos métodos con otros. Poniendo de manifiesto la necesidad de realizar diseños muestrales adaptados a cada investigación por muestreo y la necesidad de la utilización de los modelos de superpoblación en las investigaciones empíricas. Para facilitar esta tarea se ha utilizado el software POSDEM desarrollado con el propósito de optimizar la elección entre planes de muestreo alternativos. Por último dejamos abierta una vía de investigación para la formulación teórica de los resultados que hemos obtenido mediante procesos de simulación referentes al valor esperado respecto del modelo de superpoblación de la varianza del estimador del error cuadrático medio.

Los resultados muestran de forma inequívoca la necesidad de trabajar con modelos de superpoblación como estrategia para llegar a estimaciones del *ECM* y para salvar el inconveniente del posible carácter anecdótico de la configuración poblacional del momento.

Referencias

Bellhouse, D.R. y Rao, J.N.K. (1975) Systematic sampling in the presence of a trend. *Biometrika*, 62, 694-697.

Mateo, M. (2000) Posdem: selección entre planes de muestreo probabilístico, de G. Sánchez-Crespo y A. Lezcano. *Metodología de Encuestas*, 2 (1) 167-170.

Sanchez-Crespo Benitez, G; Lezcano Lastra, A. (1999): POSDEM. *Revista Electrónica de Metodología Aplicada*, 4 (2) 12 -36.