

EL MUESTREO POR CUOTAS: ALGUNAS CONSIDERACIONES.

Carmelo García Pérez

Ana Isabel Zamora Sanz

Dpto. Estadística, Estructura y O.E.I

Fac. CC. Económicas y Empresariales

Universidad de Alcalá

RESUMEN

En gran cantidad de estudios actuales se emplean los métodos de muestreo no probabilístico, tales como el muestreo por cuotas. Es posible, bajo algunas condiciones, realizar inferencias a partir del muestreo no probabilístico, utilizando estimadores de predicción, basándonos únicamente en la aleatoriedad derivada de la adopción de un modelo de superpoblación.

Palabras clave: Muestreo por cuotas, Modelos de Superpoblación.

1. INTRODUCCIÓN

Sabemos que dentro del muestreo en poblaciones finitas podemos diferenciar entre muestreo aleatorio y no aleatorio. El muestreo por cuotas es uno de los más comunes del segundo grupo.

La diferencia esencial que presentan los métodos aleatorios con el muestreo por cuotas es que en los primeros la selección de las unidades muestrales se realiza por algún método impersonal estrictamente determinado, mientras que en el segundo, una vez que se ha analizado y decidido la composición general de la muestra (ej: número de individuos en los distintos grupos de edad, sexo, clases sociales...) y se han dado las indicaciones correspondientes a cada entrevistador, la elección

de las unidades muestrales dentro de este entorno se deja a discreción de los entrevistadores. Es éste elemento humano en la selección muestral el aspecto crucial y más controvertido del muestreo por cuotas ya que se podrían introducir sesgos de selección.

La realización de inferencias en el muestreo no aleatorio ha sido muy criticada en el pasado aunque también cuenta con algunos defensores, entre ellos Smith (1983). Estos autores presentan, bajo ciertas condiciones, la justificación del uso de algunos esquemas, tales como el de cuotas, basándose en la inferencia condicional en casos en los que se pueda ignorar el plan de muestreo.

En muchas ocasiones, las técnicas del muestreo por cuotas, teóricamente más débiles, se prefieren a las del muestreo aleatorio debido a sus ventajas prácticas, su coste y a la facilidad de su implementación. Sin embargo también presentan desventajas importantes como los sesgos, la necesidad de información externa para fijar las cuotas y el complicado manejo de las no respuestas. Pocos estudios experimentales se han dedicado a comparar ambos métodos.

El muestreo aleatorio no está exento de dificultades conceptuales. Las inferencias y los cálculos de la varianza aleatoria se realizan mediante la utilización del teorema Central del Límite. Como se sabe este teorema no se puede aplicar a todas las poblaciones (Smith, 1979) por lo que se necesitarían especificar restricciones para cada población. Por otra parte el resultado de Godambe (1955) demuestra que en el caso de inferencias sobre cualquier función conocida de los valores de la variable no existe un estimador insesgado y uniformemente de mínima varianza.

La introducción de la modelización podría en algunos casos permitirnos el estudio teórico del muestreo no aleatorio. Por una parte podemos plantear hipótesis sobre un modelo que sigue la población de la que extraemos la muestra, introduciendo así una fuente de aleatoriedad que, como veremos en algunos casos, nos permite por sí sola realizar inferencias. Se puede utilizar aquí todo el bagaje teórico de los modelos de superpoblación en los que se parte de las probabilidades obtenidas bajo el modelo estimado.

Otra alternativa que nos permitiría abrir otro frente de trabajo sería partir de la base probabilística que podría aportar un modelo de selección en la recogida de la muestra. Este modelo, que afectaría a la actividad de los entrevistadores, nos proveería de una distribución muestral a partir de la cual se obtendrían desarrollos teóricos. Este enfoque no se considera en el presente trabajo.

2. NOTACIÓN Y DEFINICIONES

Supongamos una población finita compuesta por N unidades, $U=\{u_1, \dots, u_N\}$. Para facilitar la

notación cada unidad u_k será identificada por su índice k ($k=1,...,N$).

Sea $\mathbf{y}=(y_1,...,y_k,...,y_N)$ el vector de valores desconocidos objeto de estudio, donde y_k es el valor asociado a la unidad u_k , y sea $\mathbf{x}_k=(x_{k1},...,x_{kj},...,x_{kq})'$ el vector conocido de información auxiliar correspondiente a dicha unidad. Las q variables que constituyen esta información previa pueden ser de tipo cualitativo o cuantitativo. Por tanto, la información auxiliar quedará recogida en la matriz de la población $\mathbf{x}=(\mathbf{x}_1,...,\mathbf{x}_N)$.

Llamaremos s al subconjunto de unidades de la población que están contenidas en la muestra y $S=\{s/s \subset U\}$ al conjunto de todas las posibles muestras.

Se define el vector de indicadores de selección en la muestra s por $\mathbf{I}_s=(I_1,...,I_N)'$ donde cada I_k toma los valores

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

Así \mathbf{I}_s puede representar a cualquier muestra seleccionada mediante muestreo aleatorio o no aleatorio sin repetición.

Con d_k denotamos un dato u observación de forma que $d_k = \{(k, y_k)/k \in s\}$

Un diseño o esquema de muestreo se puede definir como una forma de obtener las diferentes muestras $s \in S$ mediante un proceso de selección. Esta regla puede depender de la información auxiliar contenida en \mathbf{X} , de los valores \mathbf{y} objeto de la estimación, y de un vector de parámetros desconocidos, ψ , que representan los factores que se escapan al control del que realiza la selección. Por tanto la función de cuantía asociada al esquema de selección podríamos escribirla como $p(s)=p(\mathbf{I}_s|\mathbf{x},\mathbf{y};\psi)$.

Para los métodos aleatorios tradicionales de muestreo esta función tomaría la forma $p(\mathbf{I}_s|\mathbf{x})$. En un muestreo no aleatorio, como por ejemplo el de cuotas, la regla de selección podría depender no sólo de la información auxiliar previa sino también de algunos valores de la propia variable que se desea investigar. El problema de la falta de respuesta se podría considerar como otro esquema de selección muestral en el que además de la posible dependencia en \mathbf{x} e \mathbf{y} , influirían otras variables desconocidas.

3. LA INFERENCIA BASADA EN MODELOS

En este enfoque, el vector de valores $\mathbf{y}=(y_1, \dots, y_N)$ no se considera fijo, como en el caso del muestreo tradicional en poblaciones finitas, sino que ahora se considera una realización muestral generada a partir del vector aleatorio $\mathbf{Y}=(Y_1, \dots, Y_N)$. Es decir se tendrían N variables aleatorias cuya distribución conjunta se denota por $g(\mathbf{y}; \theta)$; el modelo de superpoblación sería el conjunto de condiciones que define una clase de distribuciones a la que pertenece $g(\mathbf{y}; \theta)$. En el caso de distribuciones absolutamente continuas $g(\mathbf{y}; \theta)$ será la función de densidad conjunta de (Y_1, \dots, Y_N) .

Así pues, en general, se pueden considerar dos fuentes de aleatoriedad, la debida al diseño de muestreo y la originada por el propio modelo de superpoblación.

Nos encontramos con una dicotomía en la forma de abordar la inferencia:

- la inferencia clásica, donde los valores de la variable de interés de la población se consideran cantidades fijas, aunque desconocidas, y las probabilidades de selección, introducidas con el diseño, se utilizan para determinar las esperanzas, varianzas, errores y otras propiedades de los estimadores y,

- la inferencia basada en modelos de superpoblación (o simplemente modelos), en la cual los valores de la variable de interés de la población se consideran generados por variables aleatorias; con ellas se describe la incertidumbre sobre los valores particulares que aparecerán, a través de un modelo probabilístico. Las propiedades de los estimadores dependen de la distribución conjunta de estas variables aleatorias. Por lo tanto sería ésta una vía de análisis formal alternativo aplicable, en algunos casos, incluso a esquemas de selección no aleatorios.

4. LOS ESQUEMAS DE MUESTREO Y LA INFERENCIA BASADA EN MODELOS DE SUPERPOBLACIÓN

Las dos fuentes de aleatoriedad citadas en el epígrafe anterior, quedarían representadas por las distribuciones de las variables aleatorias \mathbf{Y} , \mathbf{I}_s que quedan expresadas, respectivamente, como $g(\mathbf{y}|\mathbf{x}; \theta)$ y $p(\mathbf{I}_s|\mathbf{x}, \mathbf{y}; \psi)$. La distribución conjunta de estas dos variables sería

$$h(\mathbf{y}, \mathbf{I}_s|\mathbf{x}; \theta, \psi) = p(\mathbf{I}_s|\mathbf{x}, \mathbf{y}; \psi) \cdot g(\mathbf{y}|\mathbf{x}; \theta)$$

Si realizamos una partición en \mathbf{Y} de forma que $\mathbf{Y}=(\mathbf{Y}_s, \mathbf{Y}_{U-s})$, siendo \mathbf{Y}_s las variables objeto de estudio que corresponden a las unidades de la muestra s , podemos obtener la marginal

$$\begin{aligned} h(\mathbf{y}_s, \mathbf{I}_s|\mathbf{x}; \theta, \psi) &= \int h(\mathbf{y}_s, \mathbf{y}_{U-s}, \mathbf{I}_s|\mathbf{x}; \theta, \psi) d\mathbf{y}_{U-s} = \\ &= \int p(\mathbf{I}_s|\mathbf{y}_s, \mathbf{y}_{U-s}, \mathbf{x}; \psi) g(\mathbf{y}_s, \mathbf{y}_{U-s}|\mathbf{x}; \theta) d\mathbf{y}_{U-s} \end{aligned}$$

entendiendo que esta notación se aplica de manera general, pudiendo particularizarse según el tipo

concreto de las variables consideradas.

Si se satisface la condición

$$p(\mathbf{I}_s|\mathbf{y},\mathbf{x};\psi)=p(\mathbf{I}_s|\mathbf{x};\psi) \quad [1]$$

que implicaría que el esquema de selección no está influido por \mathbf{Y} , sino únicamente por la información previa, \mathbf{x} , lo que se denomina diseño no informativo, entonces se tiene:

$$\begin{aligned} h(\mathbf{y}_s, \mathbf{I}_s|\mathbf{x};\theta, \psi) &= p(\mathbf{I}_s|\mathbf{x};\psi) \int g(\mathbf{y}_s, \mathbf{y}_{U-s}|\mathbf{x};\theta) d\mathbf{y}_{U-s} = \\ &= p(\mathbf{I}_s|\mathbf{x};\psi) g(\mathbf{y}_s|\mathbf{x};\theta) \end{aligned}$$

que indicaría la independencia de tipo condicional entre las variables \mathbf{I}_s e \mathbf{Y}_s . Así Smith (1983) llega a la conclusión de que si la selección de la muestra no depende de los valores de \mathbf{Y} , el diseño muestral no tendría influencia sobre las inferencias realizadas a partir del modelo de superpoblación considerado. El muestreo aleatorio simple satisface esta condición [1]; Pero es posible que también otros diseños, incluidos algunos no aleatorios, la verifiquen con lo cual su utilización quedaría justificada bajo el modelo de superpoblación.

5. MUESTREO POR CUOTAS E INFERENCIA BAJO EL MODELO DE SUPERPOBLACIÓN

Supongamos que utilizamos varios criterios h, i, \dots, j ($h=1, \dots, H$; $i=1, \dots, I$; $j=1, \dots, J$), que dividen a la población en grupos con características diferentes sobre los que se diseña la muestra por cuotas.

Sea $N_{hi\dots j}$ el número de unidades de la población que verifican las características h, i, \dots, j de los criterios anteriores. La muestra estaría formada por la suma de las $n_{hi\dots j}$ unidades elegidas entre los individuos que pertenezcan a esa categoría.

La característica diferencial del muestreo por cuotas es que la elección de los $n_{hi\dots j}$ unidades se deja a juicio del entrevistador.

Si el esquema \mathbf{I}_s representara al muestreo por cuotas, deberíamos comprobar que este esquema de selección es independiente de los valores de \mathbf{Y} para aplicar las conclusiones del punto anterior. En otras palabras, habría que demostrar que la aleatoriedad de la población a partir del modelo y la aleatoriedad en el proceso de recogida son independientes. En el muestreo aleatorio esto se cumple siempre pero en el muestreo por cuotas no hay una garantía total. A continuación supondremos que se verifica esta propiedad, lo que implica el uso de diseños no informativos.

Por tanto, el modelo de superpoblación proporcionará una relación entre los valores de \mathbf{Y} y los parámetros del modelo. Los valores observados \mathbf{y}_k , $k \in s$ permitirían conseguir estimaciones para

los parámetros y predicciones de los valores y_k que no pertenecen a la muestra s . De este modo el estimador para el total poblacional (T) de la característica en estudio, basado en estos modelos, será de la forma:

$$\hat{T} = \sum_s y_k + \sum_{U-s} \hat{y}_k$$

que utiliza los valores observados en la muestra elegida y una predicción en base al modelo para las unidades no observadas. A partir de este estimador se deduce sin dificultad el correspondiente a la media poblacional, como bien es conocido.

Para que este estimador sea insesgado bajo el modelo se debería verificar que $E(\hat{T} - T) = 0$, donde E representa la esperanza calculada a partir de la distribución conjunta del modelo de superpoblación.

5.1. Modelo de superpoblación bajo un criterio de clasificación

Consideramos un único criterio $h=1, \dots, H$ para establecer las cuotas; conocemos los tamaños poblacionales N_h correspondientes al grupo de la población cuyos individuos reúnen la característica h . El modelo de superpoblación que supondremos, en este caso, es el siguiente

$$Y_k = \alpha_h + \epsilon_k \quad k=1, \dots, N$$

donde α_h es una constante para todos las unidades que pertenecen al grupo h , y ϵ_k son variables aleatorias independientes, centradas y de varianza σ_h^2 .

Los estimadores por mínimos cuadrados de los α_h serán las medias muestrales, \bar{y}_h , en cada grupo h .

El estimador utilizando este modelo será

$$\hat{T} = \sum_{h=1}^H n_h \bar{y}_h + \sum_{h=1}^H (N_h - n_h) \bar{y}_h$$

Además para juzgar la precisión del estimador se puede calcular su varianza bajo el modelo (Deville, 1991).

5.2. Modelo de superpoblación bajo dos criterios de clasificación

En el siguiente modelo utilizaremos dos criterios para clasificar los individuos (la generalización a más de dos criterios únicamente presenta las dificultades derivadas de la notación).

Sean h e i ($h=1,\dots,H$; $i=1,\dots,I$) estos dos criterios. El número de individuos que presentan la característica h e i sería N_{hi} y n_{hi} los seleccionados en la muestra.

Definimos:

$$N_{h.} = \sum_{i=1}^I N_{hi} \quad N_{.i} = \sum_{h=1}^H N_{hi} \quad n_{h.} = \sum_{i=1}^I n_{hi} \quad n_{.i} = \sum_{h=1}^H n_{hi}$$

El modelo que se utilizará ahora será

$$Y_k = \alpha_k + \beta_k + \epsilon_k \quad \begin{matrix} k=1,\dots,N \\ h=1,\dots,H \\ i=1,\dots,(I-1) \\ \beta_I=0 \end{matrix}$$

donde de nuevo los ϵ_k son centrados e independientes; la varianza de ϵ_k es $\sigma_h^2 + \gamma_i^2$.

Los estimadores de α_h y β_i por mínimos cuadrados se obtienen resolviendo el siguiente sistema de ecuaciones

$$\begin{aligned} \sum_{i=1}^I n_{hi} \bar{y}_{hi} &= n_{h.} \hat{a}_h + \sum_{i=1}^I n_{hi} \hat{b}_i & (h=1,\dots,H) \\ \sum_{h=1}^H n_{hi} \bar{y}_{hi} &= n_{.i} \hat{b}_i + \sum_{h=1}^H n_{hi} \hat{a}_h & (i=1,\dots,I-1) \end{aligned}$$

El estimador por predicción sería:

$$\hat{T} = \sum_{hi} (N_{hi} - n_{hi}) (\hat{a}_h + \hat{b}_i) + \sum_{hi} n_{hi} \bar{y}_{hi} = N \bar{y}$$

Se puede demostrar que este estimador es insesgado y analizar también su precisión mediante la varianza calculada bajo el modelo de superpoblación.

6. CONCLUSIONES

Si se trabaja con esquemas de muestreo no informativos, es decir aquellos en los que la selección de las unidades muestrales es independiente de los valores de la variable objeto de estudio, se podría ignorar el mecanismo de selección al realizar inferencias. Probar esta condición es lo que se requiere si queremos basar nuestras inferencias únicamente en un modelo de superpoblación. Lo que implicaría en los casos no aleatorios eliminar el efecto subjetivo de selección de los entrevistadores.

En el caso de que un diseño concreto de muestreo por cuotas verificara esta condición, se analizan dos modelos de superpoblación y se presentan los estimadores por predicción obtenidos en base a este modelo.

Vemos pues que una vía de investigación abierta sobre los principios de Smith y que ha

recibido aportaciones como las de Deville, puede servir de análisis formal del muestreo no aleatorio rechazado por la teoría tradicional.

Sin embargo, es preciso enfatizar que el trabajo con modelos siempre implica la formulación de hipótesis, bien sobre el comportamiento de los individuos de una población, bien sobre un plan de muestreo. Se introduce así una subjetividad de la que está exenta la teoría clásica del muestreo aleatorio.

El estudio del muestreo por cuotas, sobre la base de una modelización del proceso de selección de los individuos, aunque no se trate en el presente trabajo, ya ha sido abordado por diversos autores. El modelo, en este caso, se establece para las probabilidades de selección.

BIBLIOGRAFÍA

CASSEL,C.M.,SÄRNDALL,C.E., et WRETMAN,J.H. (1977). Foundations of Inference in Survey Sampling. New York: Wiley & Sons.

COCHRAN, W.G. (1980) Sampling Techniques. New York: Wiley & Sons.

DESABIE, J. (1965). Théorie et pratique des sondages. Paris: Dunod.

DEVILLE, J.C (1991). Une théorie des enquêtes par quotas. Techniques de enquêtes, 1991, Vol. 17, n° 2, pp. 177-195 Statistique Canada.

GODAMBE, V.P. (1955). A unified theory of sampling from finite populations. Journal of the Royal Statistical Society. Serie B. Vol 17, 269-278.

GORIÉROUX, C. (1981). Théorie des sondages. Paris: Economica.

HANSEN, HURWITZ and MADOW (1993). Sample Survey Methods and Theory. Wiley Classics Library.

LITTLE, R.J.A. (1982). Models for non-response in sample surveys. Journal of American Statistical Association. Vol 77, 237-250

MOSER, C.A. and STUART, A. (1953). An Experimental Study of Quota Sampling. Journal of the Royal Statistical Society. Serie A. Vol CXVI.

SMITH, T.M.F. (1979). Statistical sampling in auditing: a statistician's viewpoint. The statistician. Vol 28, 4, 267-280.

SMITH, T.M.F. (1983). On the validity of inferences from non-random samples. Journal of the Royal Statistical Society, A, 146, 394-403.