UNA REVISIÓN DEL MÉTODO DE VEROSIMILITUD EMPÍRICA EN LAS ENCUESTAS POR MUESTREO

M. Rueda[1] y J. F. Muñoz[2], Departamento de Estadística e Investigación Operativa
Universidad de Granada, Granada, España

RESUMEN

En el ámbito del muestreo en poblaciones finitas han surgido en los últimos años nuevas metodologías para obtener estimadores más precisos usando información auxiliar. El método de verosimilitud empírica es el más reciente en este sentido. En este trabajo se describe este novedoso método definido en Chen y Sitter (1999), así como las distintas generalizaciones, extensión al muestreo estratificado y propiedades más importantes. El comportamiento de estos estimadores se ha comprobado con los datos de una población real y usando distintos diseños muestrales. En este estudio de simulación se muestra que estos estimadores son muy eficientes y por tanto, una alternativa valida a usar en las encuestas por muestreo.

ABSTRACT

In the last years have appeared new efficient methods to obtain estimators using auxiliary information in complex surveys. The pseudo empirical likelihood method is the most recent. In this paper, this method defined in Chen and Sitter (1999) is described. Furthermore, we show this approach under two points of view, under stratified sampling and show several properties. The real performance of these estimators is evaluated on the basis of data from a real population, using different sampling designs. This simulation study shows that these estimators are very efficient and are a valid option in complex surveys.

Key words: Empirical likelihood, superpopulation models, calibration estimator, generalized regression estimator.

MSC: 62D05

1. INTRODUCCIÓN

En la teoría de muestreo el objetivo principal de cualquier método para obtener estimadores o de cualquier diseño muestral es el de mejorar las estimaciones de los parámetros en estudio en el sentido de construir nuevos estimadores que, para el mismo tamaño muestral, tengan menor error de estimación, lo que implica mayor precisión en las estimaciones de los parámetros, o equivalentemente, tengan el mismo error que los ya conocidos pero con un menor tamaño muestral, lo que produce una disminución en el coste real de la realización de la encuesta.

Según la información que se utilice, se dispone de dos caminos para intentar mejorar la precisión de las estimaciones: por un lado, utilizando diseños muestrales más complejos (muestreos estratificados, por conglomerados, polietápicos, adaptativos, etc,) y por otro lado, emplear métodos basados en el uso de información auxiliar. Esta información auxiliar viene dada a través de un vector de variables auxiliares, $x = \{x_1,...,x_J\}$, que debe estar altamente correlacionado con la característica de interés. Se pueden combinar ambos caminos para obtener mejores estimaciones, es decir, usar diseños muestrales más complejos en métodos que utilicen información auxiliar. El método de verosimilitud empírica, que se va a describir en este trabajo, permite combinar las dos ideas anteriores y es bastante eficiente como se ha comprobado tanto teóricamente, como vía simulación.

Los primeros métodos que incorporan información auxiliar en la fase de estimación son los llamados métodos indirectos de estimación, entre los que destacan los conocidos métodos de razón, diferencia y regresión. Estos estimadores no siempre garantizan que se produzca una disminución del error de muestreo respecto a los estimadores que no usan información auxiliar. Esta ganancia en precisión depende en mayor medida de la relación entre las variables auxiliares y la variable objeto de estudio, del buen uso de las hipótesis que se supongan para emplear un procedimiento u otro, y de que dichas hipótesis se ajusten en mayor o menor medida al problema real.

E-mail: [1]mrueda@ugr.es
[2]jfmunoz@ugr.es

Todos los estimadores comentados hasta ahora se basan en un enfoque basado en el diseño muestral. Recientemente en muestreo ha surgido la perspectiva modelo-asistida, un nuevo enfoque basado en modelos de superpoblación (ver por ejemplo Pérez, 2002 y Sánchez-Crespo, 2002) y cuyo objetivo también es obtener mejores estimaciones. En este nuevo enfoque cobra especial importancia el uso de variables auxiliares cuyos valores tienen que ser conocidos para todos los individuos de la población.

Los novedosos estimadores de calibración (Deville y Särndal, 1992) y los de verosimilitud empírica (Chen y Qin, 1993 y Chen y Sitter, 1999) permiten aplicar este nuevo enfoque basado en los parámetros de un modelo de superpoblación, y concretamente usando un modelo de regresión. Estos estimadores se caracterizan por tener muy buenas propiedades teóricas y sin embargo son muy pocas las encuestas por muestreo que hacen uso de estas metodologías.

Este trabajo pretende dar a conocer los estimadores de verosimilitud empírica definidos en Chen y Qin (1993) para muestreo aleatorio simple y en Chen y Sitter (1999) para cualquier diseño muestral. El objetivo es mostrar de forma sencilla cómo se construye este estimador en distintos diseños muestrales y para los distintos enfoques existentes en muestreo, cuáles son sus propiedades más importantes y la relación que tiene con otros estimadores más conocidos. A pesar de ser una metodología bastante reciente y poco conocida, el uso de este método no está restringido a los teóricos del muestreo en poblaciones finitas, sino que su conocimiento se puede generalizar a un ambiente profesional más amplio, y en este trabajo se explica con todo detalle y de forma simple cómo se construyen estos nuevos estimadores.

2. CONSTRUCCIÓN DE ESTIMADORES DE VEROSIMILITUD EMPÍRICA

El marco de trabajo usado a lo largo de este texto es el descrito de forma muy breve a continuación.

La población U consta de N elementos distintos identificados a través de sus etiquetas i = $\{1,...,N\}$. Dentro de esta población se está interesado en estudiar ciertas características de una variable de estudio que denominaremos y. Asociado al elemento i de la población podemos conocer exactamente y sin error el valor de la característica de interés, esta cantidad se denotará como y_i. Observar el valor y_i en todas las unidades de la población va a resultar imposible o muy costoso, así que se utiliza una muestra para conocer los valores y_i de las unidades que pertenecen a la muestra. Una muestra es por tanto un subconjunto de n elementos (n < N), s, de U con sus valores asociados de y, es decir, (i, y_i), seleccionados de acuerdo con un diseño de muestreo específico que asigna una probabilidad conocida p(s) tal que p(s) > 0 para todo $s \in S$, conjunto de las posibles muestras s y $\sum_{s \in S} p(s) = 1$. p_i es la densidad de la observación y_i. Del vector de variables auxiliares, x, se van a conocer sus medias poblacionales, esto es, $\overline{X} = \overline{X}_1,...,\overline{X}_J)$. El objetivo que se persigue es estimar la media poblacional de la variable *y*, es decir, se va a estimar $\overline{Y} = N^{-1} \sum_{i \in U} y_i$.

La primera aplicación formal en muestreo para poblaciones finitas del método de verosimilitud empírica se debe a Chen y Qin (1993), que lo usaron para la estimación de la media poblacional, bajo muestreo aleatorio simple y utilizando información auxiliar. El estimador propuesto por estos autores no se puede extender a otros diseños muestrales más complejos, y por tanto surge la necesidad de plantear esta metodología para cualquier diseño muestral. Chen y Sitter (1999) proponen el estimador de verosimilitud pseudo empírica (*PEMLE*), que es aplicable a cualquier diseño muestral y coincide con el estimador propuesto por Chen y Qin (1993) bajo un muestreo aleatorio simple.

Por tanto, en los siguientes apartados nos centramos en este último estimador, daremos su definición y su extensión a los distintos enfoques y diseños de muestreo. En las secciones siguientes se comentan las propiedades más importantes y se hace un estudio de simulación para mostrar el comportamiento de estos estimadores.

2.1. Estimadores de verosimilitud empírica bajo un diseño muestral general

Como su nombre indica, este método de estimación de parámetros se basa en una función de verosimilitud para obtener los estimadores. El objetivo es maximizar la función L(p) = $\prod_{i=1}^{N} p_i$, pero como

sucede en este tipo de problemas, esto es equivalente a maximizar su logaritmo: log L(p) = l(p) = $\sum_{i=1}^{N} p_i$. Como sólo están disponibles los valores p_i para las unidades de la muestra, se debe buscar una estimación para l(p). Esta estimación viene dada por la función de verosimilitud pseudo empírica, $\hat{l}(p) = \sum_{i \in s} d_i \log p_i$, donde las cantidades d_i, i = $\{1,...,n\}$, son los pesos diseñados que hacen que la función $\hat{l}(p)$ sea insesgada bajo el diseño para l(p). Por ejemplo, en un muestreo sin reemplazo se tiene que $d_i = 1/\pi_i$, donde π_i son las probabilidades de inclusión de primer orden de las unidades de la muestra.

La información auxiliar se puede incorporar a través de la función $u_i = u(y_i, x_i)$, donde $u(\cdot)$ es una función conocida de y_i y de x_i que debe verificar: $N^{-1} \sum_{i=1}^{N} u_i = 0$.

Para construir el estimador de verosimilitud pseudo empírica hay que obtener los pesos $\hat{p}_i$ maximizando la función $\hat{l}(p)$ sujeta a las siguientes restricciones

$$\sum_{i \in s} p_i = 1 \qquad (0 \le p_i \le 1) \tag{1}$$

$$\sum_{i \in s} p_i u_i = 0 \tag{2}$$

Usando el método de los multiplicadores de Lagrange para resolver este problema, el estimador de verosimilitud pseudo empírica *(PEMLE)* propuesto en Chen y Sitter (1999) para la media poblacional es $\hat{\bar{Y}}_{PE} = \sum_{i \in S} \hat{p}_i y_i$, donde para todo i $\in$ s, $\hat{p}_i = d_i^* / (1 + \lambda' u_i)$. . El vector de multiplicadores de Lagrange, λ, se obtiene despejándolo del sistema de ecuaciones

$$g(\lambda) = \sum_{i \in s} \frac{d_i^* u_i}{1 + \lambda' u_i} = 0, \tag{3}$$

siendo $d_i^* = d_i \Big/ \sum_{j \in s} d_j$. En el caso de no disponer de información auxiliar, es decir, cuando $u_i = 0$, este método produce $\hat{p}_i = d_i^*$, obteniendo $\hat{\bar{Y}}_{PE} = \sum_{i \in s} d_i^* y_i = \bar{y}_w$. Este estimador se llama media muestral ponderada y es el estimador aproximadamente insesgado descrito en Rao (1966), Basu (1971) y Särndal et al. (1992). Para algunos diseños muestrales, este estimador y el de Horvitz-Thompson, $\hat{\bar{Y}}_{HT} = N^{-1} \sum_{i \in s} d_i y_i$, son idénticos, esto es, producen los mismos valores para cualquier muestra. Una ventaja de este estimador es que cuando N es desconocido, el estimador de Horvitz-Thompson no se puede usar en la estimación de la media poblacional. Esto autores dan otros argumentos a favor de la media muestral ponderada. En resumen, cuando N sea conocido o no, se ha comprobado que $\bar{y}_w$ es el estimador recomendado.

Otra ventaja de esta aproximación es que los pesos resultantes son estrictamente positivos, lo cual no se verifica en otros métodos como el de calibración. Esta propiedad hace que esta aproximación sea aplicable al problema de la estimación de la función de distribución y cuantiles, es decir, la función de distribución estimada mediante esta aproximación cumple todas las propiedades de una verdadera función de distribución, y por tanto, se van a obtener mejores estimaciones para los cuantiles, obtenidos mediante inversión directa de la función de distribución estimada.

Aún quedan dos aspectos muy importantes a tener en cuenta en esta metodología:

1. ¿Cómo resolver la ecuación $g(\lambda) = 0$ dada en (3)?

2. ¿Qué función u_i debemos seleccionar para obtener mejores estimaciones?

La primera cuestión no es nada simple al tratarse de ecuaciones no lineales. Se pueden utilizar métodos especiales para la resolución de ecuaciones no lineales, como el de bisección o el de Newton-Raphson. Además, en Chen, Sitter y Wu (2002) se ha propuesto un algoritmo que también calcula el *PEMLE* en caso de que exista una única solución. El algoritmo a seguir es el siguiente:

Paso 0. Sea $\lambda_0 = 0$, $k = 0$, $\gamma_0 = 1$ y $\varepsilon = 10^{-8}$.

Paso 1. Calcular $\Delta(\lambda_\kappa)$ donde

$$\Delta(\lambda_\kappa) = \left\{\frac{\partial}{\partial\lambda} g(\lambda)\right\}^{-1}; \qquad g(\lambda) = \left\{-\sum_{i\in s} \frac{d_i^* u_i u_i'}{(1+\lambda' u_i)^2}\right\}^{-1} \sum_{i\in s} \frac{d_i^* u_i}{1+\lambda' u_i}$$

Si $\|\Delta(\lambda)\| < \varepsilon$, se detiene el algoritmo con solución λ_κ. En otro caso ir al Paso 2

Paso 2. Calcular $\delta_k = \gamma_k \Delta(\lambda_k)$. Si $1 + (\lambda_k - \delta_k)' u_i \le 0$ para algún i o $\tilde{l}(\lambda_k - \delta_k) < \tilde{l}(\lambda_k)$, entonces se toma $\gamma_k = \gamma_k/2$ y se repite el Paso 2.

Paso 3. Se considera $\lambda_{k+1} = \lambda_k - \delta_k$, $k = k+1$ y $\gamma_{k+1} = (k+1)^{-1/2}$. Ir al Paso 1.

Se ha denotado $\|\cdot\|$ como la norma euclídea y $\tilde{l}(\lambda) = \sum_{i\in s} d_i^* \log(1+\lambda' u_i)$.

La demostración de este resultado se puede consultar en Chen, Sitter y Wu (2002). También se puede comprobar que este algoritmo es similar a la modificación del método de Newton descrito en Polyak (1987). Los cambios del paso 2 aseguran que la función cóncava $\tilde{l}(\lambda)$ se mueve alrededor del punto máximo. El algoritmo es simple, eficiente y la convergencia está garantizada.

Respecto a la segunda cuestión, se demuestra que para una relación lineal entre la característica de interés y el vector de variables auxiliares, considerar $u_i = x_i - \bar{X}$, para todo $i \in s$, , es la mejor opción para obtener estimaciones más precisas. En este caso, la restricción (2) se puede expresar como $\sum_{i\in s} p_i x_i = \bar{X}$, indicando que las cantidades p_i ponderadas sobre las variables auxiliares dan estimaciones perfectas para $\bar{X}$.

Para otro tipo de relación entre y y x, es necesario basarse en modelos de superpoblación para la búsqueda de las cantidades u_i que den estimaciones más eficientes. El siguiente apartado hace referencia a este caso.

2.2 Estimadores de verosimilitud empírica basados en modelos de superpoblación

El método que se acaba de comentar se basa en los datos muestrales para obtener el *PEMLE*, es decir, usa un enfoque basado en el diseño muestral. Este es el enfoque clásico que se ha usado en muestreo. Recientemente (Wu y Sitter, 2001) han usado el enfoque basado en modelos de superpoblación para aplicarlo al método de calibración y al de verosimilitud empírica. Se usa este enfoque para obtener mejores estimaciones y buscar las cantidades u_i óptimas en el sentido de minimizar la esperanza bajo un modelo de superpoblación de la varianza asintótica basada en el diseño.

Dependiendo del modelo considerado, se distinguen dos casos, la aproximación modelo-calibrada y la aproximación modelo-asistida.

En ambas aproximaciones, las funciones u_i vienen expresadas en la forma $u_i = w_i - N^{-1} \sum_{i=1}^{N} w_i$, donde w_i es una función conocida. Es fácil demostrar que bajo esta situación también se tiene que $N^{-1}\sum_{i=1}^{N} u_i = 0$, y por tanto se cumplen las condiciones necesarias para aplicar la metodología de verosimilitud empírica. Operando en la restricción (2) se llega a la restricción alternativa:

$$\sum_{i \in s} p_i w_i = \frac{1}{N} \sum_{i=1}^{N} w_i, \tag{4}$$

que es la usada en este enfoque. Por tanto, el problema de buscar la función u_i adecuada para obtener estimaciones eficientes, es similar al de encontrar la función w_i.

Para llegar a estos estimadores hay que tener en cuenta algunas definiciones y basarse en el siguiente esquema asintótico: asumimos que hay una secuencia de poblaciones finitas indexadas por ν. El tamaño poblacional y el tamaño muestral para la población ν-ésima se denotan como N_ν y n_ν. Cuando $\nu \to \infty$, $N_\nu \to \infty$ y $n_\nu \to \infty$. El índice ν se suprime para simplificar notación. Para un mayor detalle de este esquema asintótico se puede consultar Isaki y Fuller (1982).

En la aproximación modelo-calibrada, se considera que $y_1, y_2, \ldots, y_N$ es una muestra aleatoria de un modelo de superpoblación semiparamétrico ξ tal que

$$E_\xi(y_i \mid x_i) = \mu_i = \mu(x_i, \theta), \quad V_\xi(y_i \mid x_i) = \nu_i^2 \sigma_i^2, \text{ para } i = \{1,2,...,N\}, \tag{5}$$

y donde $\theta = (\theta_0, \theta_1, ..., \theta_J)'$ y σ^2 son parámetros poblacionales desconocidos, $\mu(x, \theta)$, es una función conocida de x y de θ, ν_i es una función conocida de x_i o bien de $\mu_i = \mu(x_i, \theta)$ y E_ξ y V_ξ denotan la esperanza y la varianza con respecto al modelo de superpoblación. Se asume también, que los pares $(y_1, x_1), (y_2, x_2), ..., (y_N, x_N)$ son mutuamente independientes.

Este modelo es bastante general, e incluye dos casos muy importantes: el modelo de regresión lineal y el modelo lineal generalizado.

Se considera un método basado en el diseño para la estimación de los parámetros del modelo. Cuando se emplea una aproximación basada en el modelo, (y_i, x_i) con $i \in s$ se puede ver como una muestra independiente idénticamente distribuida del modelo de superpoblación. Los parámetros θ se pueden estimar usando procedimientos estándares. Bajo el enfoque basado en el diseño, los datos muestrales pueden no seguir la misma estructura del modelo que la población finita completa bajo un esquema muestral complejo, y θ puede carecer de sentido desde el punto de vista del diseño. En este caso, θ se reemplaza por θ_N, una estimación de θ basada en los datos de la población completa. θ_N se reemplaza entonces por $\hat{\theta}$, una estimación basada en el diseño de los datos muestrales. Véase Godambe y Thompson (1986) para una mayor profundización.

Se define $\tilde{Y}_{C_w}$ como el estimador de verosimilitud pseudo empírica modelo-calibrado de $\overline{Y}$ cuando se usa $C_W = \{w_1, w_2, ..., w_n\}$ en la restricción (4). L será el conjunto de secuencias $C_W = \{w_1, w_2, ..., w_n\}$ que verifican $\frac{1}{N}\sum_{i=1}^{N} (w_i)^6 = O(1)$ y $\frac{1}{N}\sum_{i=1}^{N} (w_i)^2 \to c \neq 0$ cuando $N \to \infty$.

Estas condiciones sobre la secuencia $C_W \in L$ no son muy restrictivas y se usan para facilitar las demostraciones. Se asume que $\{\mu_1, ..., \mu_N\} \in L$.

Se dice que un diseño muestral es regular si el diseño que resulta de un tamaño de muestra indexado tiene probabilidades de inclusión π_i y π_{ij} independientes de la característica y_i dada x_i, y además satisface las siguientes condiciones:

(i) $\max_{i \in s}(nd_i / N) = O(1)$.

(ii) $\frac{1}{N}\sum_{i \in s} d_i w_i - \frac{1}{N}\sum_{i=1}^{N} w_i = O_p(n^{-1/2})$ para toda secuencia de funciones $(w_1,...,w_N) \in$ L.

El estimador de verosimilitud pseudo empírica modelo-calibrado *MCPE* se construye considerando $w_i = \mu(x_i, \hat{\theta})$, o dicho de otro modo, la función u_i viene dada por $u_i = \hat{\mu}_i - N^{-1}\sum_{i=1}^{N}\hat{\mu}_i$, donde $\hat{\mu}_i = \mu(x_i, \hat{\theta})$. Se observa que es necesario conocer por completo el vector de variables auxiliares para obtener esta aproximación, pero a cambio, las estimaciones van a ser más precisas respecto al *PEMLE* y respecto a otros métodos que tan sólo utilizan los datos muestrales de las variables auxiliares.

En Wu (2003) se demuestra que entre todas las clases de estimadores $\tilde{Y}_{C_w}$, donde $C_W = \{w_1, w_2,..., w_N\} \in$ L, el valor $C_\mu = \{\mu(x_1, \theta),...,\mu(x_N,\theta)\}$ como variable de calibración en (4) minimiza $E_\xi\left[AV_p(\tilde{Y}_{C_w})\right]$ bajo el modelo (5) y para cualquier diseño muestral regular. AV_p denota la varianza asintótica bajo el diseño.

Observaciones sobre la aproximación modelo-calibrada

1. En Wu y Sitter (2001) se demuestra que reemplazar θ por $\hat{\theta}$ en $\mu_i = \mu(x_i, \theta)$, no cambia asintóticamente el estimador resultante.

2. Con probabilidad tendiendo a uno, el *MCPE* existe y se puede calcular usando el algoritmo comentado en secciones anteriores de Chen, Sitter y Wu (2002).

3. El uso efectivo de la información auxiliar depende de los parámetros estimados y de la relación entre la variable respuesta y las covarianzas. Por tanto, usar ciegamente la calibración sobre las variables auxiliares no es usualmente una buena aproximación.

4. Ya se ha comentado que para una relación lineal entre *y* y el vector de variables auxiliares, se toma $u_i = x_i - \bar{X}$ para la construcción del *PEMLE*. En esta situación, el *PEMLE* y el *MCPE* son asintóticamente equivalentes si se considera $\hat{\mu}_i = x_i'\hat{\theta}$ como variable de calibración para el cálculo de la aproximación modelo-calibrada. La demostración puede consultarse en Wu y Sitter (2001).

5. Si la relación entre y y x es lineal, tan sólo el conocimiento de $\bar{X}$ es suficiente para obtener estimadores eficientes para la media o el total poblacional. Si dicha relación no es lineal o el parámetro de interés no es una función lineal, una información auxiliar completamente disponible y/o más datos sobre el modelo son esenciales para una estimación óptima.

6. Al igual que se ha comentado anteriormente, las cantidades p_i son positivas. Esta propiedad no se cumple ni en los estimadores de calibración ni en cálculo del *GREG* y juega un papel muy importante en la estimación de otros parámetros de interés en el muestreo, como son la función de distribución, cuantiles, varianza y otras funciones cuadráticas.

La aproximación modelo-asistida se diferencia de la anterior en que se considera otro modelo distinto, esto es $(y_1, y_2,...,y_N)$ es una muestra aleatoria de un modelo de superpoblación ξ tal que

$$E_\xi(y_i) = \mu_i,\ V_\xi(y_i) = \sigma_i^2,\ \text{para } i = \{1,2,...,N\}, \qquad (6)$$

y donde $(y_1,...,y_n)$ son independientes entre ellos.

En Wu (2003) también se puede comprobar que entre todas las clases de estimadores $\tilde{Y}_{C_w}$ con $C_W = \{w_1,w_2,...,w_N\} \in$ L, el valor $C_\mu = \{\mu_1,...,\mu_N\}$ como variable de calibración en (4) minimiza $E_\xi\left[AV_p(\tilde{Y}_{C_w})\right]$

bajo el modelo (6) y para cualquier diseño muestral regular. El estimador de verosimilitud pseudo empírica modelo-asistido se construye tomando $w_i = \mu_i$, o dicho de otro modo, considerando $u_i = \mu_i - N^{-1}\sum_{i=1}^{N}\mu_i$.

2.3. Estimadores de verosimilitud empírica bajo muestreo estratificado

El muestreo estratificado es una técnica muy poderosa y simple que es muy usada en la práctica por varias razones: administrativas (cuando un territorio está dividido en distritos geográficos), ganancia en eficiencia respecto a diseños no-estratificados (cuando los estratos están bien formados), etc.

La metodología de verosimilitud empírica también puede aplicarse para obtener estimadores en diseños muestrales más complejos, y en concreto en un muestreo estratificado. El esquema de muestreo a seguir es el siguiente. La población U consta de N unidades y es dividida en L subpoblaciones de tamaños $N_1,N_2,...,N_L$. Estas subpoblaciones, que reciben el nombre de estratos, no se superponen y juntas forman la totalidad de la población, es decir, $\sum_{h=1}^{L} N_h = N$. S_h es el conjunto de todas las posibles muestras de tamaño n_h para el estrato h. Una vez que han sido determinados los estratos se extrae una muestra $s_h \in S_h$ de tamaño n_h de cada uno mediante un diseño de muestreo específico que asigna una probabilidad conocida $p(s_h)$ a cada $s_h \in S_h$ tal que $p(s_h) > 0$ y $\sum_{s_h \in S_h} p(s_h) = 1$. La muestra final está compuesta por el conjunto de estas submuestras y su tamaño será $n = \sum_{h=1}^{L} n_h$. El proceso de muestreo se realiza de modo independiente en cada estrato, lo que permite la aplicación simultánea de métodos de muestreo diferentes. y_{hi} y x_{hi} son los valores en la unidad i del estrato h para la característica y y el vector x respectivamente. $W_h = N_h/N$, $\overline{Y}_h$ y $\overline{X}_h$ son las medias poblacionales del estrato h de la variable y y del vector x. Por último, p_{hi} es la densidad de la observación i en el estrato h.

Bajo este muestreo, se tiene que $l(p) = \sum_{h=1}^{L}\sum_{i=1}^{N_h} \log p_{hi}$, que se puede ver como un total poblacional, cuya estimación insesgada bajo el diseño muestral es

$$\hat{l}(p) = \sum_{h=1}^{L}\sum_{i\in s_h} d_{hi} \log p_{hi} \qquad (7)$$

En este caso, d_{hi} son los pesos diseñados básicos que hacen que $\hat{l}(p)$, denominada log-función de verosimilitud pseudo empírica, sea insesgada bajo el diseño para l(p). Por ejemplo, si se usa muestreo aleatorio simple en cada estrato, se tiene que $d_{hi} = N_h/n_H$.

En este caso el problema es maximizar (7) sujeta a las restricciones

$$\sum_{i\in s_h} p_{hi} = 1 \qquad (p_{hi} > 0),\ h = \{1,...,L\} \qquad (8)$$

$$\sum_{h=1}^{L} W_h \sum_{i\in s_h} p_{hi} x_{hi} = \overline{X} \qquad (9)$$

En la restricción (9) se ha considerado por comodidad que la relación entre y y x es lineal, aunque para cualquier otra relación es posible modificar esta restricción. Una vez obtenidas todas las soluciones $\hat{p}_{hi}$ de este problema, el *PEMLE* bajo muestreo estratificado es

$$\hat{\bar{Y}}_{PEst} = \sum_{h=1}^{L} W_h \sum_{i \in s_h} \hat{p}_{hi} y_{hi} \qquad (10)$$

.

Este planteamiento descrito no tiene una resolución simple, y el algoritmo propuesto en Chen, Sitter y Wu (2002) no puede ser aplicado para obtener las soluciones. En la actualidad, existen dos algoritmos para resolver este planteamiento. El más eficiente de ellos puede consultarse en Wu (2004).

En el caso particular de que las cantidades $\bar{X}_h$, con h = $\{1,...,L\}$, sean conocidas, el cálculo del *PEMLE* en muestreo estratificado es más simple. Bajo esta situación, se considera la restricción

$$\sum_{i \in s_h} p_{hi} x_{hi} = \bar{X}_h \qquad (11)$$

en cada estrato y se maximiza (7) sujeta a las restricciones (8) y (11). Entonces el estimador se puede obtener calculando el *PEMLE* para cada estrato, $\hat{\bar{Y}}_{PEh}$, y luego ponderando estas cantidades por el peso del estrato, esto es $\hat{\bar{Y}}_{PEst} = \sum_{h=1}^{L} W_h \hat{\bar{Y}}_{PEh}$

3. PROPIEDADES ASINTÓTICAS DE LOS ESTIMADORES DE VEROSIMILITUD EMPÍRICA

En esta sección se describen las principales propiedades asintóticas que se tienen en la actualidad sobre los estimadores de verosimilitud empírica descritos en las secciones anteriores.

3.1. Propiedades en un diseño muestral general

Por comodidad, se va a considerar una sola variable auxiliar y relación lineal entre ésta y la característica de interés, es decir, que $u_i = x_i - \bar{X}$. Se deben comprobar las siguientes condiciones

(i) $u^* = \max_{i \in s} |u_i| = o_p(n^{1/2})$

(ii) $\dfrac{\sum_{i \in s} d_i u_i}{\sum_{i \in s} d_i u_i^2} = O_p(n^{-1/2})$

Teorema 1. (Chen y Sitter, 1999) Bajo las dos condiciones anteriores, el PEMLE de $\bar{Y}$ cuando $\bar{X}$ es conocida, es asintóticamente equivalente a un estimador de regresión generalizado (GREG). Es decir, $\hat{\bar{Y}}_{PE} = \hat{\bar{Y}}_{GREG} + o_p(n^{-1/2})$, donde $\hat{\bar{Y}}_{GREG} = \sum_{i \in s} \tilde{d}_i y_i$,

$$\tilde{d}_i = d_i^* \left[1 - \frac{(x_i - \bar{x}_w)(\bar{x}_w - \bar{X})}{\sum_{i \in s} d_i^* (x_i - \bar{x}_w)^2} \right], \ \bar{y}_w = \sum_{i \in s} d_i^* y_i, \ \bar{x}_w = \sum_{i \in s} d_i^* x_i \ \text{ y } \ d_i^* \frac{d_i}{\sum_{j \in s} d_j}.$$

Las condiciones descritas anteriormente no son muy restrictivas, y los diseños muestrales más usados las satisfacen. Estas condiciones se cumplen en tres diseños comunes, como son, el muestreo con probabilidades proporcionales al tamaño con reemplazamiento, el método de Rao-Hartley-Cochran y el muestreo por conglomerados.

Otro punto importante es la estimación de la varianza de $\hat{\bar{Y}}_{PE}$. Según el Teorema anterior, es evidente que cualquier estimador de la varianza consistente para $\hat{\bar{Y}}_{GREG}$ será consistente para el *PEMLE*. Aunque esto es

asintóticamente válido, no es atractivo usar un estimador de la varianza del *GREG* para estimar la varianza del *PEMLE*. Una alternativa óptima es aplicar estimadores de remuestreo de la varianza, tales como jackknife, bootstrap y submuestras repetidas balanceadas (ver Shao y Wu ,1989, 1992, Chen y Qin,1993 y Shao, 1994) sobre $\hat{\bar{Y}}_{PE}$, recalculando p_i en cada remuestra.

3.2. Propiedades en la aproximación modelo-calibrada

En este apartado se asume que existe una secuencia de diseños muestrales y una secuencia de poblaciones finitas indexadas por ν. El tamaño muestral n_ν y el tamaño poblacional N_ν se aproximan a infinito cuando $\nu \to \infty$.

A continuación se detallan las condiciones necesarias para poder aplicar el Teorema 2.

(i) $\hat{\theta} = \theta_N + O_P(n^{-1/2})$ y $\theta_N \to \infty$.

(ii) Para cada x_i, $\partial\mu(x_i, t)/\partial t$ es continua en t y $|\partial\mu(x_i, t)/\partial t| \leq h(x_i, \theta)$, para t en un entorno de θ, y $N^{-1}\sum_{i=1}^{N} h(x_i,\theta) = O_p(1)$.

(iii) Los pesos básicos muestrales, d_i , hacen que los estimadores de Horvitz-Thompson para ciertas medias muestrales estén asintóticamente normalmente distribuidos.

(iv) $u^* = \max_{i\in s}|u_i| = o_p(n^{1/2})$, donde $u_i = \mu(x_i, \theta_N) - N^{-1}\sum_{i=1}^{N}\mu(x_i,\theta_N)$.

(v) $$\frac{\sum_{i\in s} d_i u_i}{\sum_{i\in s} d_i u_i^2} = O_p(n^{-1/2}).$$

(vi) $h^* = \max_{i\in s}|h_i| = o_p(n)$, siendo $h_i = h(x_i, \theta_N)$.

Teorema 2. Bajo el esquema asintótico descrito y las condiciones (i) ~ (vi) anteriores, se tiene que $\hat{\bar{Y}}_{MCPE} = \hat{\bar{Y}}_{MC} + o_p(n^{-1/2})$, donde $\hat{\bar{Y}}_{MC}$ es el estimador modelo-calibrado para la media obtenido mediante el método de calibración y cuya expresión es la siguiente

$$\hat{\bar{Y}}_{MC} = \hat{\bar{Y}}_{HT} + \left\{\frac{1}{N}\sum_{i=1}^{N}\hat{\mu}_i - \frac{1}{N}\sum_{i\in s} d_i\hat{\mu}_i\right\}\hat{B}_N$$

con

$$\hat{B}_N = \frac{\sum_{i\in s} d_i q_i(\hat{\mu}_i - \bar{\mu})(y_i - \bar{y})}{\sum_{i\in s} d_i q_i(\hat{\mu}_i - \bar{\mu})^2}, \quad \bar{y} = \frac{\sum_{i\in s} d_i q_i y_i}{\sum_{i\in s} d_i q_i}, \; \bar{\mu} = \frac{\sum_{i\in s} d_i q_i \hat{\mu}_i}{\sum_{i\in s} d_i q_i}.$$

Las cantidades q_i son constantes positivas.

Puesto que $\hat{\bar{Y}}_{MCPE}$ es asintóticamente equivalente al $\hat{\bar{Y}}_{MC}$, la misma expresión de la varianza y del estimador de la varianza de $\hat{\bar{Y}}_{MC}$ se puede usar para $\hat{\bar{Y}}_{MCPE}$. Estas varianzas asintóticas vienen dadas por

$$V\left(\hat{\bar{Y}}_{MCPE}\right) = \frac{1}{N^2}\sum_{i<j}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{U_i}{\pi_i} - \frac{U_j}{\pi_j}\right)^2,$$

donde π_{ij} son las probabilidades de inclusión de segundo orden, $U_i = y_i - \mu_i B_N$, $\mu_i = \mu(x_i, \theta_N)$,

$$B_N = \frac{\sum_{i=1}^{N} q_i(\hat{\mu}_i - \overline{\mu}_N)(yi - \overline{Y})}{\sum_{i=1}^{N} q_i(\hat{\mu}_i - \overline{\mu}_N)^2} \text{ y } \overline{\mu}_N = \frac{1}{N}\sum_{i=1}^{N}\mu_i$$

y la estimación es

$$\hat{V}(\hat{\overline{Y}}_{MCPE}) = \frac{1}{N^2}\sum_{i<j}^{n}\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{u_i}{\pi_i} - \frac{u_j}{\pi_j}\right)^2,$$

donde $u_i = y_i - \hat{\mu}_i\hat{B}_N$

Aunque estas aproximaciones son asintóticamente válidas, resulta más atractivo usar estimadores de varianzas remuestreados sobre el MCPE.

3.3. Propiedades en muestreo estratificado

Paralelamente al Teorema 1, se puede llegar a que el *PEMLE* bajo muestreo estratificado es asintóticamente equivalente a un *GREG*, pero esta no es la mejor aproximación posible, puesto que se sabe que el estimador de regresión óptimo (*ORE*), definido en Rao (1994), funciona mejor que el *GREG* en muestreo estratificado. Por este motivo, en Chen y Sitter (1999) se busca una mejor aproximación. El siguiente corolario se basa en un muestreo aleatorio estratificado y asume que existe una secuencia de poblaciones finitas indexadas por ν, tal que cuando $\nu \to \infty$ se verifican las condiciones

(i) $0 \le c_1 \le \sum_{h=1}^{L} W_h\sigma_h^2 \le c_2 \le \infty.$

(ii) $\max\left\{n_h^{-1}W_h\right\} = O(n^{-1}).$

(iii) $N^{-1}\sum_{h=1}^{L}\sum_{i=1}^{N_h}\left|x_{hi}\right|^3 = O(1).$

(iv) $N^{-1}\sum_{h=1}^{L}\sum_{i=1}^{N_h}\left|y_{hi}\right|^3 = O(1).$

Corolario 1. Bajo muestreo aleatorio estratificado y las cuatro condiciones anteriores, el PEMLE de $\overline{Y}$, cuando $\overline{X}$ es conocida y la información del tamaño del estrato es incorporada, es asintóticamente equivalente al estimador lineal óptimo dado en Rao (1994).

Bajo otros diseños muestrales en cada estrato, las comparaciones entre dos métodos de estimación distintas son demasiado dificultosas y se recurre a la simulación para realizar las comparaciones. En este caso, la estimación de la varianza se obtiene también a través de estimadores de la varianza remuestreados. En Chen y Sitter (1999), se demuestra que bajo muestreo aleatorio estratificado el estimador de la varianza jackknife para el PEMLE es consistente.

4. ESTUDIO EMPÍRICO

Con el fin de comprobar el comportamiento real de los estimadores de verosimilitud pseudo empírica frente a otras aproximaciones, en esta sección se muestra un resumen del estudio de simulación que hemos obtenido sobre la población Fam1500 para el parámetro media muestral, $\overline{Y}$. Esta población consta de 1500 familias de Andalucía y puede consultarse en Fernández y Mayor (1994). La variable de interés, y, denota los gastos de alimentación de dichas familias. Además, se dispone de dos variables auxiliares (ingresos familiares y otros gastos) que permiten obtener los distintos estimadores indirectos bajo los métodos de muestreo más usuales.

Los estimadores indirectos que se consideran para contrastar la eficiencia de los estimadores de verosimilitud pseudo empírica son el estimador de razón y el GREG. La eficiencia de estos estimadores se compara en términos de eficiencia relativa (RE), siendo

$$\text{RE} = \frac{\text{ECM}(\hat{\bar{Y}}_j)}{\text{ECM}(\hat{\bar{Y}}_{base})};\ \text{ECM}(\hat{\bar{Y}}_j) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\bar{Y}}_j(b) - \bar{Y})^2,$$

donde j denota el estimador utilizado, B = 1000 son las muestras independientes generadas, b se refiere a la muestra b-ésima, y $\hat{\bar{Y}}_{base}$ es el estimador estándar del diseño muestral considerado que se utiliza para la comparación.

Por tanto, se obtendrá la RE de estos tres estimadores bajo distintos esquemas de muestreo y tamaños muestrales. Estos resultados pueden consultarse en las Tablas 1,...,6. Los coeficientes de correlación lineal entre la variable de interés y las variables ingresos y otros gastos son de 0,848 y 0.546 respectivamente. Se han utilizado ambas variables auxiliares para la construcción de los distintos estimadores, pudiéndose comprobar cómo le afectan a estos el grado de linealidad.

Las generaciones aleatorias y todos los estimadores están calculados con el programa R. Esta programación se puede solicitar por correo electrónico a los autores.

Tabla 1.Muestreo aleatorio simple. Se utiliza la variable auxiliar ingresos para obtener estimadores.

T. muestral:	**25**	**50**	**75**	**100**	**125**	**150**	**175**	**200**	**250**	**300**
PEMLE	0.288	0.273	0.258	0.290	0.279	0.281	0.277	0.284	0.288	0.282
Razón	0.281	0.270	0.256	0.291	0.277	0.279	0.277	0.281	0.287	0.282
GREG	0.286	0.273	0.258	0.290	0.280	0.281	0.277	0.284	0.288	0.282

Tabla 2. Muestreo aleatorio simple. Se utiliza la variable auxiliar otros gastos para obtener estimadores.

T. muestral:	**25**	**50**	**75**	**100**	**125**	**150**	**175**	**200**	**250**	**300**
PEMLE	0.768	0.699	0.719	0.735	0.749	0.701	0.707	0.745	0.721	0.675
Razón	1.356	1.304	1.371	1.398	1.515	1.528	1.417	1.429	1.314	1.366
GREG	0.744	0.698	0.718	0.736	0.749	0.701	0.708	0.745	0.721	0.675

Tabla 3. Muestreo con probabilidades desiguales y sin reemplazo. Se extraen muestras por el método de Lahiri utilizando la variable auxiliar otros gastos y se utiliza la variable auxiliar ingresos para obtener estimadores.

T. muestral:	**25**	**50**	**75**	**100**	**125**	**150**	**175**	**200**	**250**	**300**
PEMLE	0.151	0.113	0.082	0.064	0.054	0.051	0.039	0.036	0.032	0.025
Razón	0.129	0.101	0.080	0.063	0.054	0.051	0.039	0.035	0.032	0.025
GREG	0.129	0.101	0.080	0.063	0.054	0.051	0.039	0.035	0.032	0.025

Tabla 4. Muestreo con probabilidades desiguales y sin reemplazo. Se extraen muestras por el método de Midzuno utilizando la variable auxiliar ingresos y se utiliza la variable auxiliar otros gastos para obtener estimadores.

T. muestral:	25	50	75	100	125	150	175	200	250	300
PEMLE	0.754	0.755	0.735	0.717	0.718	0.733	0.741	0.721	0.649	0.687
Razón	1.401	1.408	1.383	1.302	1.353	1.427	1.456	1.356	1.326	1.364
GREG	1.355	1.364	1.338	1.260	1.310	1.380	1.409	1.313	1.278	1.317

Tabla 5. Muestreo aleatorio estratificado. Se utiliza la variable auxiliar ingresos para obtener estimadores.

T. muestral:	80	100	125	150	175	200	250	300	400	500
PEMLE	0.975	0.992	0.991	0.997	0.992	0.992	0.988	0.997	0.996	0.999
Razón	0.976	0.991	0.990	0.995	0.991	0.991	0.989	0.996	0.996	0.999
GREG	0.978	0.998	0.996	1.000	0.996	0.995	0.991	0.999	0.998	1.000

Tabla 6. Muestreo estratificado con probabilidades iguales y con reemplazo. Se utiliza la variable auxiliar ingresos para obtener estimadores.

T. muestral:	80	100	125	150	175	200	250	300	400	500
PEMLE	0.305	0.298	0.290	0.300	0.319	0.268	0.300	0.293	0.287	0.336
Razón	0.295	0.289	0.279	0.294	0.316	0.267	0.300	0.295	0.286	0.336
GREG	0.299	0.297	0.287	0.299	0.317	0.267	0.299	0.293	0.287	0.335

Las principales observaciones que se deducen de estos resultados son las siguientes:

- En general, las estimaciones de los tres estimadores son bastante similares, obteniéndose una ganancia en eficiencia notable respecto al estimador estándar. Por ejemplo, en la Tabla 1 la *RE* de los estimadores está en torno a 0.280, suponiendo esto una mejora en eficiencia respecto al estimador estándar del 72 %. Esta ganancia es aún mayor en la Tabla 3, donde se obtiene hasta un 97.5 % de mejoría respecto al estimador base. Estos datos hacen recomendar el uso de cualquier estimador basado en información auxiliar para obtener mayor precisión en la estimación de parámetros lineales.

- En la mayoría de los casos el *PEMLE* tiene un comportamiento equivalente al *GREG*, siendo esto más significativo conforme aumenta el tamaño de la muestra. Este hecho se ha demostrado en Chen y Sitter (1999).

- En todos los casos el *PEMLE* es más eficiente que el estimador estándar, es decir, su *RE* está siempre por debajo de 1. Esto no ocurre ni con los estimadores de razón ni con el *GREG*.

- Cuando existe poca linealidad entre la variable de interés y la auxiliar que se utiliza para construir los estimadores, esto es, cuando se ha utilizado la variable otros gastos (Tablas 2 y 4), o bien el estimador de razón, o bien el *GREG*, son mucho peores que el estimador estándar, mientras que el *PEMLE* mejora siempre en eficiencia al estimador estándar.

Estos datos son un ejemplo del estudio de simulación llevado a cabo sobre otros esquemas de muestreo y otras poblaciones reales y simuladas. En todos los casos, se llega a conclusiones similares que las observadas en este caso.

En resumen, los estimadores de verosimilitud empírica se ven menos afectados que los estimadores de razón y regresión generalizada a la falta de linealidad, obteniéndose en cualquier caso unas estimaciones muy precisas. Se puede decir, por tanto, que los estimadores de verosimilitud empírica son una alternativa válida a los estimadores de razón y regresión generalizada.

REFERENCIAS

BASU, D. (1971): Foundations of statistical inference. A Symposium, eds. V.P. Godambe and D. A. Sprott, Holt Rinehart and Winston. Toronto.

CHEN, J. and J. QIN (1993): "Empirical likelihood estimation for finite populations and the effective usage of auxiliary information", **Biometrika**, 80, 107-116.

CHEN, J. and R.R. SITTER (1999): "A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys", **Statistica Sínica**, 9, 385-406.

CHEN, J.; R.R. SITTER and C. WU (2002): "Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys", **Biometrika**, 89, 230-237.

DEVILLE, J. and C. SÄRNDAL (1992): "Calibration estimators in survey sampling", **Journal of the American Statistical Association**, 87, 376-382.

FERNÁNDEZ, F.R. and J.A. MAYOR (1994): Muestreo en Poblaciones Finitas: Curso Básico. P.P.U, Barcelona.

GODAMBE, V.P. and M.E. THOMPSON (1986): "Parameters of superpopulation and survey population: Their relationships and estimation", **International Statistical Review**, 54, 127-138.

ISAKI, C.T. and W.A. FULLER (1982): "Survey design under the regression superpopulation model", **Journal of the American Statistical Association**, 77, 89-96.

PÉREZ, R.A. (2002): "¿Qué es un modelo de superpoblación?", **Metodología de Encuestas**, 4 (1), 79-86.

POLYAK, B.T. (1987): Introduction to Optimization. Optimization Software, Inc. Publications Division. New York.

RAO, J.N.K. (1966): "Alternative estimators in PPS sampling for multiple characteristics", **Sankhya, Ser**. A, 28, 47-60.

RAO, J.N.K. (1994): "Estimating totals and distribution functions using auxiliary information at the estimation stage", **Journal of Official Statistics**, 10, 153-165.

SÁNCHEZ-CRESPO, G. (2002): "Introducción a los modelos de superpoblación en las técnicas de muestreo con probabilidades desiguales", **Metodología de Encuestas**, 4 (1), 87-104.

SÄRNDAL, C.E.; B. SWENSSON and J.H. WRETMAN (1992): Model Assisted Survey Sampling. Springer-Verlag, New York.

SHAO, J. (1994): "L-statistics in complex survey problems", **The Annals of Statistics**, 22, 946-967.

SHAO, J. and C.F.J. WU (1989): "A general theory for jackknife variance estimation", **The Annals of Statistics**, 17, 1176-1197.

_____ (1992): "Asymptotic properties of the balanced repeated replication method for sample quantiles", **The Annals of Statistics**, 20, 1571-1593.

WU, C. (2003): "Optimal calibration estimators in survey sampling", **Biometrika**, 90, 937-951.

_____ (2004): "Some algorithmic aspects of the empirical likelihood method in survey sampling", **Statistica Sinica**, 14. In press.

WU, C. and R.R. SITTER (2001): "A model-calibration approach to using complete auxiliary information from survey data", **Journal of the American Statistical Association**, 96, 185-193.