Estadística Española • Volumen 55, número 182 / 2013, pp. 323-335

Aplicación de técnicas de calibración en la estimación de líneas de pobreza

Sergio Martínez Puertas
Departamento de Matemáticas. Universidad de Almería. CITE III
Almería

Helena Martínez Puertas
Departamento de Matemáticas. Universidad de Almería. CITE III
Almería

Resumen

El objetivo principal de este artículo es la estimación de líneas de pobreza a través de la estimación de ratio de percentiles. Para ello, el artículo propone estimadores calibrados para los percentiles y con ello estimar el ratio correspondiente. Emplearemos técnicas de explotación intensiva de la muestra para la estimación de la varianza, así como para la obtención de intervalos de confianza. Una aplicación con datos reales basados en datos procedentes de la Encuesta de Presupuestos Familiares realizada por el Instituto Nacional de Estadística (INE) correspondiente al año 2006, es incluida en el artículo para ilustrar cómo las técnicas propuestas funcionan mejor que las técnicas ya existentes.

Palabras Clave: Información Auxiliar, Estimadores de Calibración, Métodos Jackknife, Líneas de pobreza.

Clasificación AMS: 62D05

Application of calibration techniques in estimating poverty lines

Abstract

The article's goal is the estimation of the poverty measures by developing percentile ratio estimation. For it, the article propose calibration estimators for the percentiles and then consider the estimation of the corresponding ratio. For the variance estimation and confidence interval construction, the authors consider the use of resampling techniques. Finally, a numerical application with real data extracted from the Encuesta de Presupuestos Familiares 2006, a survey conducted by INE, is included to illustrate how suggested procedures can perform better than existing ones.

Keywords: Auxiliary information, Calibration estimators, Jackknife methods, Poverty measures.

AMS Classification: 62D05

1. Introducción

En la actualidad hay un creciente interés en los estudios sobre el análisis de la pobreza. Entre las razones por las cuales puede explicarse esta tendencia actual nos encontramos con el hecho de que las líneas de pobreza pueden ser un indicador del bienestar económico de un país, también pueden poner de manifiesto problemas de equidad social o de estratificación social.

Por estas razones, tales estudios se centran en el análisis de la distribución de ingresos, así como en el estudio de la desigualdad social, ya que esta última puede ser un factor de sustentación de diferencias generalizadas de los ingresos en países desarrollados (ver European Commision (1998), Kahn (1998)).

Entre las medidas empleadas para la medición de la desigualdad salarial podemos encontrar el ratio del percentil 95 con el percentil 20, el ratio de percentiles 95/50, el ratio 50/10 o los ratios 50/5 y 50/25 analizados por Dickens and Manning (2004).

El principal problema que presentan todas las medidas anteriormente mencionadas, es que no son funciones lineales de los datos pues dependen de los cuantiles y esto hace muy difícil la incorporación de información auxiliar ya que los trabajos relacionados con la estimación de la mediana u otros cuantiles que incorporan información auxiliar son menos numerosos, que por ejemplo para el caso de la estimación de la media y por tanto un modo adecuado de mejorar la estimación de las medidas de pobreza, se puede alcanzar con la incorporación eficiente de la información auxiliar a la hora de definir nuevos estimadores de cuantiles. Para conseguir este objetivo, lo que nos proponemos en primer lugar, es la estimación de cuantiles a partir de la función de distribución, esto es, incorporar la información auxiliar en la estimación de dicha función para obtener nuevos estimadores más eficientes y a partir de estos últimos, mediante su inversión, obtener estimaciones de los cuantiles.

Deville and Särndal (1992) propusieron el método de calibración en la estimación de la media poblacional, con el objetivo de incorporar la información auxiliar de un modo más eficiente y Rueda et al (2007a) (2007b) proponen estimadores de calibración para la función de distribución, que a su vez son auténticas funciones de distribución. El propósito de este trabajo, es emplear las técnicas de calibración, para en primer lugar obtener estimadores de la función de distribución, que incorporen de manera eficaz la información auxiliar, en segundo lugar a través de estos estimadores vamos a poder estimar el cuantil de cualquier orden de manera eficiente y por tanto es de esperar que los ratios de percentiles y cualquier otra medida de pobreza que dependa de cuantiles puede ser estimada de manera eficiente.

Otra cuestión que se plantea es, debido a la complejidad de los ratios de percentiles y a los diseños muestrales usados en los estudios oficiales de pobreza (diseños estratificados, diseños con probabilidades de inclusión desiguales, etc), el cálculo de la varianza de los estimadores de las medidas de pobreza propuestos, para lo cual se hace necesario proponer métodos alternativos para la estimación de la varianza de los estimadores empleados. Con este propósito, consideraremos el uso de técnicas de explotación intensiva de la muestra y más concretamente emplearemos el método Jackknife para la estimación de la varianza.

El trabajo queda estructurado de la siguiente manera, en la Sección 2 revisaremos distintos estimadores de la función de distribución que emplean información auxiliar y posteriormente definiremos nuevos estimadores de la misma por medio de técnicas de calibración. Estos nuevos estimadores permiten definir estimadores para los cuantiles y para los ratios de percentiles. En la sección 3 tratamos el problema de la estimación de la varianza de los estimadores de las medidas de pobreza propuestos en la Sección 2. Finalmente en la Sección 4 mostramos mediante una aplicación con datos reales la actuación de los nuevos estimadores propuestos en comparación con los ya existentes, siendo los resultados bastante satisfactorios.

2. Estimadores de Calibración para Cuantiles

La técnica usual para la estimación de cuantiles consiste en la estimación eficiente de la función de distribución y mediante su inversión poder estimar el cuantil de interés. A continuación, debido a la cantidad existente de técnicas para estimar la función de distribución, resumiremos algunas de las más importantes.

Sea U una población finita formada por N diferentes unidades, y sean $y_1, y_2, \dots y_N$ los valores que una variable de estudio y toma en los distintos elementos de U. Sea $\mathbf{x_i} = \left(x_{i1}, \dots x_{ij}, \dots x_{iJ}\right)'$ el valor de un vector de variables auxiliares en la unidad i de la población. Supongamos que una muestra s, es extraída de la población U de acuerdo con un determinado diseño muestral, cuyas probabilidades de inclusión de primer orden vienen dadas por $\pi_i = P[i \in s]$. Además supondremos que el valor y_i de la variable de estudio y sólo está disponible para las unidades i incluidas en la muestra s, mientras que el valor $\mathbf{x_i}$ del vector auxiliar es conocido para todas los elementos de la población U.

La función de distribución poblacional de la variable y viene dada por:

$$F_y(t) = \frac{1}{N}\sum_{i\in U} \Delta\left(t - y_i\right)$$

donde

$$\Delta\left(t - y_i\right) = \begin{cases} 1 \text{ si } t \geq y_i \\ 0 \text{ si } t < y_i \end{cases}$$

El cuantil de orden β viene definido por:

$$Q_y(\beta) = \inf\left\{t : F_y(t) \geq \beta\right\}$$

y para poder estimar $Q_y(\beta)$, debemos obtener una estimación $F_y(t)$ de la función de distribución, para después estimar $Q_y(\beta)$ de la siguiente manera:

$$\hat{Q}_y(\beta) = \inf \left\{t: \hat{F}_y(t) \geq \beta\right\}$$

Una vez establecido el mecanismo para la estimación de cuantiles, podemos pasar a estimar nuestro auténtico objetivo, el ratio

$$r_{0.95,0.5} = \frac{Q_y(0.95)}{Q_y(0.5)} \qquad [1]$$

que puede estimarse mediante:

$$\hat{r}_{0.95,0.5} = \frac{\inf \left\{t: \hat{F}_y(t) \geq 0.95\right\}}{\inf \left\{t: \hat{F}_y(t) \geq 0.5\right\}}$$

De este modo, a la hora de la estimación de $r_{0.95,0.5}$, debemos plantearnos la estimación de la función de distribución $F_y(t)$, de manera que podamos incorporar la información auxiliar proporcionada por el vector $\mathbf{x_i}$. Para ello, disponemos en la literatura reciente de diversas técnicas, como por ejemplo los estimadores de razón y estimadores de tipo diferencia (ver Rao et al 1990). Estos estimadores presentan algunos inconvenientes, siendo uno de ellos, el hecho de que este tipo de estimadores pueden tomar valores fuera del rango [0,1] y también que no son por sí mismos funciones de distribución. Más aún, no siempre son funciones monótonas no decrecientes y por tanto su empleo a la hora de estimar cuantiles no es recomendable.

Chambers y Dunstan (1986) propusieron un estimador de $F_y(t)$, basado en el enfoque predictivo y que por tanto se basa fundamentalmente en la existencia de un modelo de superpoblación, que liga la variable de estudio con las variables auxiliares. Entre las principales desventajas de este estimador se encuentra que tiene un comportamiento pobre si no se respeta el modelo de superpoblación y su dificultad computacional. Rao et al (1990) propusieron un estimador basado en el diseño, asintóticamente insesgado bajo el modelo de superpoblación y asintóticamente insesgado bajo el diseño. Su principal desventaja es su alto coste computacional.

Recientemente, el método de calibración ha sido empleado en la estimación de la media poblacional y en la estimación de la función de distribución. Concretamente en Rueda et al (2007a,b) se modifica la técnica usual de calibración, de manera que el estimador así obtenido posee propiedades deseables tales como que es una auténtica función de distribución. Los estimadores de $F_y(t)$ que emplearemos en la estimación de líneas de pobreza, se basan en estimadores calibrados, y es por ello que pasamos a desarrollar esta técnica.

El estimador usual de $F_y(t)$, es el estimador de Horvitz-Thompson, el cual vienen dado por:

$$\hat{F}_{YHT}(t) = \frac{1}{N}\sum_{i\in s} d_i \Delta\left(t - y_i\right) \qquad [2]$$

Para incorporar la información auxiliar proporcionada por el vector $\mathbf{x_i}$, siguiendo a Rueda et al (2007a,b), en primer lugar vamos a definir una pseudo variable g, dada por $g_i = \beta'\mathbf{x}_i$ donde

$$\beta = \left(\sum_{i\in s} d_i\, q_i\, x_i\, x_i'\right)^{-1} . \sum_{i\in s} d_i\, q_i\, x_i\, y_i$$

El método de calibración sustituye los pesos básicos d_i por unos nuevos pesos ω_i, que minimizan la distancia

$$\Phi_s = \frac{1}{2}\sum_{i\in s}\frac{\left(\omega_i - d_i\right)^2}{d_i\, q_i} \qquad [3]$$

sujeto a las siguientes condiciones de calibración:

$$\frac{1}{N}\sum_{i\in s}\omega_i \Delta\left(\mathbf{t_g} - g_i\right) = F_g\left(\mathbf{t_g}\right) \qquad [4]$$

donde los valores q_i son constantes positivas arbitrariamente elegidas y donde $\mathbf{t_g} = \left(t_{g1}, \ldots t_{gP}\right)'$ es un vector de puntos.

El primer estimador de calibración para $F_y(t)$ que proponemos, toma por vector $\mathbf{t_g} = (t_{g50}, t_{g95})'$ donde $t_{g50} = Q_g(0.5)$ y $t_{g95} = Q_g(0.95)$. De este modo, el estimador cumple las siguientes condiciones:

$$\frac{1}{N}\sum_{i\in s}\omega_i \Delta\left(t_{g50} - g_i\right) = F_g\left(t_{g50}\right) \qquad [5]$$

$$\frac{1}{N}\sum_{i\in s}\omega_i \Delta\left(t_{g95} - g_i\right) = F_g\left(t_{g95}\right) \qquad [6]$$

La razón de elegir el vector $\mathbf{t_g} = \left(t_{g50}, t_{g95}\right)$ radica en el hecho de que así el estimador de la función de distribución obtenido proporciona estimaciones perfectas para los valores t_{g50} y t_{g95}. Los pesos obtenidos con este estimador vienen dados por:

$$\omega_i = d_i + d_i q_i N\left(F_g(\mathbf{t_g}) - F_{GHT}(\mathbf{t_g})\right)' T^{-1}\Delta(\mathbf{t_g} - g_i) \qquad [7]$$

con $\Delta(\mathbf{t_g} - g_i)' = \Delta(t_{g50} - g_i), \Delta(t_{g95} - g_i)$ y con

$$T = \sum_{i \in s} d_i \, q_i \Delta\left(\mathbf{t_g} - g_i\right) \Delta\left(\mathbf{t_g} - g_i\right)' \quad [8]$$

y el estimador resultante viene dado por:

$$\hat{F}_{YC1}(t) = \hat{F}_{YHT}(t) + \left(F_g(\mathbf{t_g}) - \hat{F}_{GHT}(\mathbf{t_g})\right)' \hat{D} \quad [9]$$

donde

$$\hat{D} = T^{-1} \sum_{i \in s} d_i \, q_i \Delta\left(\mathbf{t_g} - g_i\right) \Delta\left(t - y_i\right)$$

Otra posibilidad a la hora de obtener un nuevo estimador, es seleccionar el vector $\mathbf{t_g} = \left(t_{g25}, t_{g50}, t_{g75}\right)'$. Así, las condiciones que cumple el estimador, en este caso, son:

$$\frac{1}{N} \sum_{i \in s} \omega_i \Delta\left(t_{g25} - g_i\right) = F_g\left(t_{g25}\right) \quad [10]$$

$$\frac{1}{N} \sum_{i \in s} \omega_i \Delta\left(t_{g50} - g_i\right) = F_g\left(t_{g50}\right) \quad [11]$$

$$\frac{1}{N} \sum_{i \in s} \omega_i \Delta\left(t_{g95} - g_i\right) = F_g\left(t_{g95}\right) \quad [12]$$

Entre las razones por las cuales puede justificarse esta elección, está el hecho de que para que el proceso de calibración tenga solución los valores seleccionados no deben estar muy próximos, ya que puede haber problemas de condiciones de calibración incompatibles.

Los pesos obtenidos con esta elección viene dado por la expresión [7] pero sustituyendo el vector $\mathbf{t_g} = (t_{g50}, t_{g95})'$ por el vector $\mathbf{t_g} = (t_{g25}, t_{g50}, t_{g75})'$ y el estimador calibrado obtenido, que denotaremos por $\hat{F}_{YC2}(t)$ tiene la misma expresión que el estimador $\hat{F}_{YC1}(t)$ pero realizando la misma sustitución anteriormente comentada.

El principal problema que podemos encontrarnos con los estimadores $\hat{F}_{YC1}(t)$ y $\hat{F}_{YC2}(t)$ es que no son auténticas funciones de distribución. Concretamente, no cumplen las propiedades de monotonía no decreciente ni la propiedad de que el límite en $+\infty$ sea igual a la unidad, ya que:

$$\hat{F}_{YC1}(+\infty) = \lim_{t \to +\infty} \hat{F}_{YC1}(t) = \frac{1}{N} \sum_{i \in s} \omega_i$$

y en general esta cantidad no es igual a la unidad, lo que también pasaría en el caso del estimador $\hat{F}_{YC2}(t)$. Ahora bien, la propiedad fundamental a la hora de la estimación de cuantiles, es la propiedad de monotonía no decreciente, siendo la propiedad del límite en $+\infty$ igual a la unidad menos relevante.

Para que el estimador obtenido con las condiciones de calibración [4] sea una auténtica función de distribución, siguiendo a Rueda et al (2007a,b) vamos a suponer que los puntos $t_{g1}, t_{g2}, \dots t_{gP}$ están ordenados, es decir:

$$t_{g1} < t_{g2} < \cdots t_{gP}$$

Además vamos a asumir las dos siguientes condiciones:

1. El punto $t_{g\,P} = t_{g\,M} = \max\limits_{k \in U} g_k$

2. $q_i = c$ para $i \in U$

La primera condición está relacionada con la propiedad del límite en $+\infty$, mientras que la segunda condición es para alcanzar la monotonía no decreciente, fundamental en la estimación de cuantiles. Así, vamos a considerar nuevamente los estimadores $\hat{F}_{YC1}(t)$ y $\hat{F}_{YC2}(t)$, pero con la elección $q_i = c \ \forall i \in U$, ya que con esta elección, podemos emplear los estimadores $\hat{F}_{YC1}(t)$ y $\hat{F}_{YC2}(t)$ en la estimación de cuantiles. Para analizar con más detalle, el motivo de la elección $q_i = c$, trataremos el caso de un estimador calibrado, en general con J puntos, esto es $\mathbf{t_g} = (t_1, t_{2,} \dots t_J)$, de manera que:

$$t_1 < t_2 < \cdots < t_J$$

y dada la muestra s, sea

$$g_1 \leq g_2 < \cdots \leq g_n$$

los valores que la variable auxiliar g toma en las unidades de la muestra s, ordenados de menor a mayor. Vamos a suponer también:

El valor t_1 deja por debajo los primeros k_1 valores de la variable g en la muestra
El valor t_2 deja por debajo los primeros k_2 valores de la variable g en la muestra
⋮ ⋮ ⋮
El valor t_J deja por debajo los primeros k_J valores de la variable g en la muestra

Y supondremos también que los $n - k_J$ restantes valores de la variable g en la muestra s son superiores a t_J

Con las suposiciones anteriores, los pesos calibrados ω_i toman la siguiente forma (Rueda et al (2007a,b))

$$\omega_i = d_i + d_i\, q_i\, N \frac{\left(F_g(t_1) - \hat{F}_{GHT}(t_1)\right)}{\sum\limits_{i=1}^{k_j} d_i q_i} \qquad i = 1, 2, \dots, k_1$$

y en general para las unidades $i = k_{j-1} + 1, \dots, k_j$, los pesos tienen la expresión:

$$\omega_i = d_i + d_i\, q_i\; N \frac{\left(F_g\left(t_{j-1}\right) - \hat{F}_{GHT}\left(t_{j-1}\right)\right)}{\sum_{i=k_{j-1}+1}^{k_j} d_i q_i} + d_i\, q_i\; N \frac{\left(F_g\left(t_j\right) - \hat{F}_{GHT}\left(t_j\right)\right)}{\sum_{i=k_{j-1}+1}^{k_j} d_i q_i}$$

Y finalmente para las unidades muestrales $i = k_J + 1, \dots, n$ tenemos que los pesos $\omega_i = d_i$

De este modo, para las unidades muestrales $i = k_J + 1, \dots, n$ los pesos ω_i son positivos pues coinciden con los pesos d_i. En cuanto a las unidades $i = 1, 2, \dots, k_1$, si consideramos la opción $q_i = c$, tenemos que:

$$\omega_i = d_i + d_i\, c\; N \frac{\left(F_g\left(t_1\right) - \hat{F}_{GH}\left(t_1\right)\right)}{c \sum_{i=1}^{k_1} d_i} = \frac{N\, d_i\, F_g\left(t_1\right)}{\sum_{i=1}^{k1} d_i} \geq 0$$

De manera similar, para las unidades $k_{j-1} + 1, \dots, k_j$ con $j = 2, \dots, J$, con la elección $q_i = c$, tenemos:

$$d_i + d_i\, c\; N \frac{\left(F_g\left(t_j\right) - \hat{F}_{GH}\left(t_j\right)\right)}{c \sum_{i=k_{j-1}+1}^{k_j} d_i} - d_i\, c\; N \frac{\left(F_g\left(t_{j-1}\right) - \hat{F}_{GH}\left(t_{j-1}\right)\right)}{c \sum_{i=k_{j-1}+1}^{k_j} d_i} = \frac{N\, d_i \left(F_g\left(t_j\right) - F_g\left(t_{j-1}\right)\right)}{\sum_{i=k_{j-1}+1}^{k_j} d_i} \geq 0$$

De este modo, con la elección $q_i = c$ tenemos que los pesos calibrados $\omega_i \geq 0$ para toda unidad i de la muestra s y en consecuencia, el estimador calibrado correspondiente

$$\hat{F}_{YC}\left(t\right) = \frac{1}{N} \sum_{i=1}^{n} \omega_i\, \Delta\left(t - y_i\right)$$

es claramente monótono no decreciente.

Por otro lado, si queremos que el estimador calibrado obtenido cumpla todas las propiedades de función de distribución, debemos incorporar en el proceso de calibración también la condición 1) y es por ello, que a continuación consideremos como elección nueva para el vector $\mathbf{t_g}$, la siguiente selección $\mathbf{t_g} = \left(t_{g50}, t_{g95}, t_{gM}\right)'$, con lo que las condiciones que cumple el nuevo estimador son:

$$\frac{1}{N} \sum_{i \in s} \omega_i \Delta\left(t_{g50} - g_i\right) = F_g\left(t_{g50}\right) \qquad [13]$$

$$\frac{1}{N}\sum_{i\in s}\omega_i\Delta\left(t_{g95}-g_i\right)=F_g\left(t_{g95}\right) \qquad [14]$$

$$\frac{1}{N}\sum_{i\in s}\omega_i\Delta\left(t_{gM}-g_i\right)=\frac{1}{N}\sum_{i\in s}\omega_i=F_g\left(t_{gM}\right)=1 \qquad [15]$$

El estimador obtenido con esta nueva elección, que denotaremos por $\hat{F}_{YC3}(t)$ y cuya expresión es similar a la de los estimadores $\hat{F}_{YC1}(t)$ y $\hat{F}_{YC2}(t)$, es una auténtica función de distribución.

Finalmente podemos considerar el estimador $\hat{F}_{YC4}(t)$, que se obtiene al seleccionar como vector $\mathbf{t_g}=(t_{g25},t_{g50},t_{g75},t_{gM})'$. De este modo, el estimador $\hat{F}_{YC4}(t)$ tiene una expresión similar a los anteriores, pero en este caso dicho estimador cumple las siguientes condiciones:

$$\frac{1}{N}\sum_{i\in s}\omega_i\Delta\left(t_{g25}-g_i\right)=F_g\left(t_{g25}\right) \qquad [16]$$

$$\frac{1}{N}\sum_{i\in s}\omega_i\Delta\left(t_{g50}-g_i\right)=F_g\left(t_{g50}\right) \qquad [17]$$

$$\frac{1}{N}\sum_{i\in s}\omega_i\Delta\left(t_{g95}-g_i\right)=F_g\left(t_{g95}\right) \qquad [18]$$

$$\frac{1}{N}\sum_{i\in s}\omega_i\Delta\left(t_{gM}-g_i\right)=\frac{1}{N}\sum_{i\in s}\omega_i=F_g\left(t_{gM}\right)=1 \qquad [19]$$

Así, disponemos de cuatro estimadores nuevos para la función de distribución $F_y(t)$, todos ellos monótonos no decrecientes (y por tanto aptos para la estimación de cuantiles), mientras que sólo dos de ellos (los estimadores $\hat{F}_{YC3}(t)$ y $\hat{F}_{YC4}(t)$) son aunténticas funciones de distribución. A continuación, podemos definir los correspondientes estimadores para el cuantil $Q_y(\beta)$, de la siguiente manera:

$$\hat{Q}_{YC1}(\beta)=\inf\left\{t:\hat{F}_{YC1}(t)\geq\beta\right\} \qquad [20]$$

$$\hat{Q}_{YC2}(\beta)=\inf\left\{t:\hat{F}_{YC2}(t)\geq\beta\right\} \qquad [21]$$

$$\hat{Q}_{YC3}(\beta)=\inf\left\{t:\hat{F}_{YC3}(t)\geq\beta\right\} \qquad [22]$$

$$\hat{Q}_{YC4}(\beta) = \inf\{t : \hat{F}_{YC4}(t) \geq \beta\} \qquad [23]$$

Una vez, que disponemos de estimadores para la estimación de cuantiles, podemos definir los estimadores para la línea de pobreza $r_{0.95,0.5}$, que vienen dados por:

$$\hat{r}_{0.95,0.5}^{YC1} = \frac{\hat{Q}_{YC1}(0.95)}{\hat{Q}_{YC1}(0.5)} = \frac{inf\{t:\hat{F}_{YC1}(t) \geq 0.95\}}{inf\{t:\hat{F}_{YC1}(t) \geq 0.5\}} \qquad [24]$$

$$\hat{r}_{0.95,0.5}^{YC2} = \frac{\hat{Q}_{YC2}(0.95)}{\hat{Q}_{YC2}(0.5)} = \frac{inf\{t:\hat{F}_{YC2}(t) \geq 0.95\}}{inf\{t:\hat{F}_{YC2}(t) \geq 0.5\}} \qquad [25]$$

$$\hat{r}_{0.95,0.5}^{YC3} = \frac{\hat{Q}_{YC3}(0.95)}{\hat{Q}_{YC3}(0.5)} = \frac{inf\{t:\hat{F}_{YC3}(t) \geq 0.95\}}{inf\{t:\hat{F}_{YC3}(t) \geq 0.5\}} \qquad [26]$$

$$\hat{r}_{0.95,0.5}^{YC4} = \frac{\hat{Q}_{YC4}(0.95)}{\hat{Q}_{YC4}(0.5)} = \frac{inf\{t:\hat{F}_{YC4}(t) \geq 0.95\}}{inf\{t:\hat{F}_{YC4}(t) \geq 0.5\}} \qquad [27]$$

3. Estimación de la Varianza y Obtención de Intervalos de Confianza

En los estudios sobre líneas de pobreza es interesante obtener la varianza de las diferente medidas estimadas, pues permite decidir acerca del grado de validez o confianza de los datos en relación con el uso que se va a hacer de ellos. El problema que presenta la medida de pobreza $r_{0.95,0.5}$, es que al ser un ratio, no es una función lineal de los datos lo que dificulta la obtención de una expresión para la varianza de sus correspondientes estimadores. Por este motivo, debemos hacer uso de técnicas para el cálculo aproximado de la varianza de un estimador. Concretamente, en esta sección emplearemos una técnica de explotación intensiva de la muestra, el método Jackknife, que adaptaremos para la estimación de la varianza de un estimador de $r_{0.95,0.5}$ y para la obtención de intervalos de confianza.

Si suponemos que disponemos de un estimador $\hat{r}_{0.95,0.5}$ para el ratio $r_{0.95,0.5}$ y queremos estimar su varianza, vamos a considerar las muestras s_i resultantes de eliminar el elemento i-ésimo de la muestra s, obteniendo así n muestras de tamaño n–1. A continuación si obtenemos el estimador $\hat{r}_{0.95,0.5}$ para cada muestra s_i obtenemos los estimadores $\hat{r}_{0.95,0.5}(s_i)$ y a partir de estos valores definimos los pseudo valores:

$$\hat{r}_{0.95,0.5}^{(i)} = n\,\hat{r}_{0.95,0.5} - (n-1)\,\hat{r}_{0.95,0.5}\,(s_i) \qquad [28]$$

con los que definimos el estimador jackknife de la siguiente manera:

$$\hat{r}_{0.95,0.5}^{J} = \frac{1}{n}\sum_{i=1}^{n} \hat{r}_{0.95,0.5}^{(i)} \qquad [29]$$

y finalmente el estimador de la varianza para $r_{0.95,0.5}$ viene dado por:

$$\hat{V}_J = \frac{1}{n\,(n-1)} \sum_{i=1}^{n} \left(\hat{r}^{(i)}_{0.95,0.5} - \hat{r}^{J}_{0.95,0.5} \right) \qquad [30]$$

El estimador jackknife [29] no es un estimador consistente, pero es asintóticamente normal si el diseño muestral empleado es muestreo aleatorio simple. Por tanto, el intervalo de confianza jackknife (1−γ)100% de $r_{0.95,0.5}$ viene dado por:

$$I_J = \left(r^J_{0.95,0.5} - z_{1-\gamma/2}.\hat{V}_J^{1/2}, r^J_{0.95,0.5} + z_{1-\gamma/2}.\hat{V}_J^{1/2} \right) \qquad [31$$

Donde $z_{\gamma/2}$ es el cuantil $\gamma/2$ de una distribución normal estándar.

4. Aplicación con Datos Reales

En esta sección, proponemos un estudio empírico de los estimadores de medidas de pobreza propuestos en secciones anteriores, para analizar su comportamiento y precisión en comparación con otras técnicas ya existentes. También hemos analizado el comportamiento de los métodos de explotación intensiva de la muestra, propuestos en la sección 3, para la estimación de la varianza y construcción de intervalos de confianza.

Concretamente, el estudio que se ha llevado a cabo consiste en una aplicación con datos reales correspondientes a N=5000 familias españolas y que han sido extraídos de la Encuesta de Presupuestos Familiares realizada por el Instituto Nacional de Estadística (INE) correspondiente al año 2006 y donde hemos considerado como variable de estudio y="Ingresos mensuales familiares" y donde solamente hemos considerado una variable auxiliar x="Gastos totales anuales familiares". Para llevar a cabo el estudio, se seleccionó, mediante muestreo aleatorio simple, una muestra de tamaño n=550 y se obtuvo la estimación del ratio $r_{0.95,0.5}$ mediante los estimadores *YC*1; *YC*2; *YC*3 y *YC*4. Además de los estimadores propuestos, también se obtuvieron los siguientes estimadores:

- El estimador de Horvitz-Thompson *HT*.
- El estimador diferencia *D*
- El estimador de razón *R*.
- El estimador de Chambers-Dunstan *CD*
- El estimador de Rao-Kovar-Mantel *RKM*

Para todos los estimadores incluidos en el estudio se calculó, la estimación del ratio $\hat{r}_{0.95,0.5}$, la estimación de la varianza mediante Jackknife descrita en la sección 3, el intervalo de confianza y la longitud de dicho intervalo de confianza. Los resultados obtenidos pueden consultarse en la tabla 1.

Tabla 1

Medidas para analizar el comportamiento de los estimadores del ratio $r_{0.95,0.5}$ basados en una muestra de tamaño n=550

Estimador	Medida			
	$\hat{r}_{0.95,0.5}$	$\hat{V}(\hat{r}_{0.95,0.5})$	$I_{95\%}$	$L_{95\%}$
HT	2,4571	0,0747	(1,9214-2,9928)	1,0714
D	2,4498	0,0611	(1,9653-2,9343)	0,9690
R	2,7954	0,5790	(1,3040-4,2869)	2,9828
CD	2,5981	0,0325	(2,2448-2,9515)	0,7067
RKM	2,5305	0,0334	(2,1723-2,8888)	0,7164
*YC*1	2,3994	0,4542	(1,3482-3,9900)	2,6419
*YC*2	2,6140	0,4546	(1,2924-3,9355)	2,6430
*YC*3	2,4724	0,0372	(2,0944-2,8504)	0,7561
*YC*4	2,4724	0,0319	(2,1223-2,8225)	0,7001

En los resultados obtenidos pueden observarse que el estimador con menor varianza estimada y por tanto con una menor longitud de intervalo de confianza es el estimador *YC*4, con un comportamiento similar a los estimadores *CD* y *RKM*, siendo las diferencias de varianzas prácticamente inapreciables. En cuanto a los estimadores propuestos, también destaca el estimador *YC*3 cuya varianza es menor que la mayoría de los estimadores incluidos en el estudio, excepto en el caso de los estimadores *YC*4; *CD* y *RKM*, si bien con todos ellos la diferencia de varianza es escasa y por tanto presenta un comportamiento similar. De este modo, si tenemos en cuenta que los estimadores de calibración propuestos tienen un bajo coste computacional en comparación a los estimadores *CD* y *RKM*, podemos concluir que los estimadores *YC*4 e *YC*3 son una gran alternativa que produce resultados similares con menor coste computacional. Por otro lado, los resultados obtenidos con los estimadores *YC*1 e *YC*2 son pobres y únicamente presentan mejor resultados si lo comparamos con el estimador de razón. La mala actuación de estos estimadores puede deberse al hecho de que no son auténticas funciones de distribución. Así concluimos que el estimador que presenta los mejores resultados es el estimador *YC*4, que si bien no incluye en sus condiciones de calibración el cuantil 0.95, al incluir el valor máximo en sus restricciones, cumple todas las propiedades de función de distribución.

Referencias

CHAMBERS, R.L., DUNSTAN, R., (1986). «Estimating distribution function from survey data». *Biometrika*. 73, 597-604.

DEVILLE, J.C., SÄRNDAL, C.E., (1992). «Calibration Estimators in Survey Sampling». *Journal of the American Statistical Association*. 87, 376-382.

DICKENS, R., MANNING, A., (2004). «Has the national minimun wage reduced UK wage inequality?». *J.R. Stat. Soc. Ser. A.* 167, 613-626.

EUROPEAN COMMISION, (1998). «Social protection in Europe». European Commission, Brusssels.

HARMS, T., DUCHESNE, P., (2006). «On calibration estimation for quantiles». *Survey Methodology*. 32, 37-52.

KAHN, L., (1998). «Collective bargaining and the interindustry wage structure: international evidence». *Economica*. 65, 507-534.

MARTÍNEZ, S., RUEDA, M., ARCOS, A. AND MARTÍNEZ, H., (2010). «Optimum calibration points estimating distribution functions». *Journal of Computational and Applied Mathematics*. 233, 2265-2277.

RAO, J.N.K., KOVAR, J.G., MANTEL, H.J., (1990). «On estimating distribution function and quantiles from survey data using auxiliary information». *Biometrika*. 77, 365-375.

RUEDA, M., MARTÍNEZ, S., MARTÍNEZ, H., ARCOS, A., (2007). «Estimation of the distribution function with calibration methods». *Journal of Statistical Planning and Inference*. 137, 435-448.

RUEDA, M., MARTÍNEZ, S., MARTÍNEZ, H. AND ARCOS, A., (2007). «Calibration methods for estimating quantiles». *Metrika*. 66, 355-371.

SÄRNDAL, C.E., (2007). «The calibration approach in survey theory and practice». *Survey Methodology*. 33, 99-119.

SINGH, S., (2003). «Advanced sampling theory with applications: How Michael ”selected” Amy». The Netherlands. Kluwer Academic Publisher.