Bortolotto, Eugenia
Marí, Gonzalo
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística

UNA REVISIÓN DE LOS DISTINTOS ESTIMADORES ROBUSTOS PARA MUESTREO EN POBLACIONES FINITAS

RESUMEN

Uno de los objetivos de las encuestas por muestreo es la estimación de parámetros de variables de interés. Una de las soluciones viene del lado de los estimadores clásicos que gozan de buenas propiedades distribucionales como por ejemplo el insesgamiento. El problema surge cuando en la encuesta se presentan en algunas variables, valores alejados del común de los datos. Ante esta situación, los estimadores clásicos presentan dificultades que se ven traducidas en un desempeño pobre respecto a medidas relacionadas a la precisión. Se presenta una revisión de estimadores denominados robustos, los cuales son menos sensibles a la aparición de valores outliers. Por otro lado, se muestran estimadores de variancia para los estimadores planteados, así como los códigos disponibles en el programa estadístico R.

Palabras claves: muestreo en poblaciones finitas, estimador de Horvitz-Thompson, estimador de Hájek, estimadores robustos

ABSTRACT

One of the objectives of sample surveys is the estimation of parameters of variables of interest. One solution is the use of the classical estimators that have good distributional properties such as, for example, unbiasedness. The problem arise when in a survey, values from some variables are far from the common data. Given this situation, the classical estimators present difficulties that are translated in a poor performance with respect to precision measures. We present a review of robust estimators, which are less sensitive to the appearance of outliers. Furthermore, we show variance estimators for the proposed estimators, as well as the available codes in the statistical program R.

Keywords: sampling from finite populations, Horvitz-Thompson estimator, Hájek estimator, robust estimators

INTRODUCCIÓN

Cuando se está interesado en conocer características de una determinada población, como totales, medias o proporciones, existen distintas formas de recolectar la información, pudiéndose mencionar entre las más importantes, a los censos y a las encuestas. Los primeros constituyen el método de recolección de datos más antiguo. Los mismos contemplan la enumeración completa de la población de interés, siendo los más conocidos los Censos de Población, Hogares y Viviendas, los Censos Agropecuarios, y los Censos Económicos, desarrollados en nuestro país por el Instituto Nacional de Estadística y Censos (INDEC) como entidad rectora del Sistema Estadístico Nacional (SEN). Debido al nivel de cobertura, los censos poseen la ventaja de obtener datos que permiten brindar información a niveles muy desagregados de la población en cuestión. Como desventaja, se puede mencionar que los mismos son operativos muy grandes, los cuales son costosos y muy difíciles de controlar, y como consecuencia de esta falta de control, brindan resultados que en ciertas situaciones pueden tener un nivel de precisión pobre.

La segunda fuente mencionada considera la recolección de datos a partir de encuestas por muestreo. A diferencia de los censos, las unidades sobre las cuales se recolectan los datos son un subconjunto de la población. El objetivo es, a partir de esta muestra, poder inferir a toda población. Existen dos tipos de muestras, las probabilísticas y las no probabilísticas. Las primeras son las que aseguran la representatividad de las mismas y permiten realizar inferencias válidas para la población considerada. Esto se justifica a partir del hecho de asignar una probabilidad a cada unidad de la población no nula de ser seleccionada. Estas probabilidades son las que luego se utilizan en la etapa inferencial la cual debe contemplar el método de selección utilizado y diversas características como la no respuesta. En cambio, en el muestreo no probabilístico se desconoce la probabilidad de selección de las unidades, no se puede evaluar la precisión en términos probabilísticos y no garantiza la representatividad de las muestras sobre la población.

En esta investigación, sólo se tienen en cuenta la recolección de datos a través de muestreos probabilísticos. Entre los estimadores clásicos más utilizados para estimar medias y totales de las variables de interés se pueden mencionar el estimador de Horvitz-Thompson y, para el caso de la media, el estimador de Hajek. Con respecto a la estimación de la media, el primero se utiliza cuando el tamaño de la población sobre la cual se va a inferir es conocido, mientras que el segundo se emplea generalmente cuando se desconoce esta cantidad.

Una de las dificultades que surgen en los datos recolectados en encuestas para una amplia gama de aplicaciones, es que los mismos contienen frecuentemente una o más observaciones atípicas llamadas outliers, que son observaciones que están separadas de la mayoría de los datos. En estos casos los estimadores clásicos de la media y el total, pueden estar muy influenciados por los outliers y no arrojar estimaciones precisas.

Una de las soluciones a este problema es la utilización de diseños muestrales que consideren información auxiliar que permita identificar estas unidades. Una opción es utilizar un muestreo estratificado, agrupando a la población de acuerdo al tamaño de los valores de la variable auxiliar y relevando a todas las unidades del estrato que contiene a las observaciones de mayor tamaño. Sin embargo, algunas unidades con valores outliers en algunas variables aún pueden ser seleccionados de forma inesperada en la muestra debido a información auxiliar poco precisa en el momento de seleccionar la muestra.

Una segunda solución al problema planteado proviene de la estadística robusta. La misma contempla un conjunto de técnicas y herramientas que resultan menos sensibles a la aparición de estas observaciones atípicas. Existen una serie de estimadores considerados robus-

tos que son menos sensibles a los estimadores clásicos ante la aparición de observaciones extremas. Entre sus características se puede mencionar que los mismos son generalmente no lineales, y precisan de la definición de términos constantes para su aplicación. Se analizará cómo la media ponderada, la media winsorized y la media truncada deben ser adaptadas a los ponderadores muestrales, luego se discutirá el estimador univariado a un paso, la formulación del estimador de razón a un paso que contiene como caso especial la robustificación a un paso del estimador Horvitz-Thompson (HT-estimador), que puede ser expresado como una media ponderada. Se presenta para cada uno de los estimadores, un método para estimar la variancia.

En la sección 2 de este trabajo se detallará la teoría de los estimadores clásicos de Horvitz-Thompson y de Hajek. En la sección 3 se presentarán los estimadores robustos mencionados anteriormente para la estimación de parámetros que se basen en variables con observaciones atípicas con sus correspondientes estimadores de variancia. En la sección 4, la descripción de los algoritmos existentes en el programa estadístico R, tanto para la estimación de los parámetros como para la obtención de una medida de precisión estimada para cada uno de ellos. Y por último se describirán los estudios futuros a desarrollar.

ESTIMADORES CLÁSICOS

Sea una muestra s seleccionada a través de un diseño $p_d(.)$ de una población $U=\{u_1, u_2, \dots, u_N\} = \{1,2, \dots, N\}$. Estamos interesados en estimar parámetros sobre la característica y, como el total poblacional $t = \sum_{k\in U} y_k$, o la media poblacional $\bar{y}_U = \frac{t}{N} = \sum_U y_k/N$.

Una característica interesante que tienen las poblaciones finitas de N elementos, es que los mismos pueden tener probabilidades de inclusión diferentes en la muestra. Sea $p_d(.)$ un diseño fijo, la inclusión de un elemento k en la muestra es un evento aleatorio indicado por una variable aleatoria I_k, definida como

$$I_k = \begin{cases} 1 \text{ si } k\epsilon S \\ 0 \text{ en otro caso} \end{cases}$$

En un muestreo sin reemplazo, π_k es la probabilidad de inclusión de primer orden, es decir, la probabilidad de que la unidad k-ésima pertenezca a la muestra y se obtiene como $\pi_k = P(k\epsilon S) = E(I_k) = P(I_k = 1) = \sum_{s\ni k} p(s)$. La probabilidad de inclusión π_k puede ser calculada como la suma de las probabilidades de todas las muestras que contienen la k-ésima unidad y tiene la propiedad que $\sum_{k=1}^{N} \pi_k = n$.

La probabilidad que las unidades k y l estén ambas en la muestra se denota como π_{kl} y se obtiene de un diseño $p_d(.)$ dado, de la siguiente manera

$$\pi_{kl} = P(k \wedge l\epsilon S) = E(I_k I_l) = P(I_k I_l = 1) = \sum_{s\ni k\&l} p(s)$$

Se tiene que $\pi_{kl} = \pi_{lk}$ para todo k y l. Además si $k = l$, entonces $\pi_{kk} = P(I^2{}_k = 1) = P(I_k = 1) = \pi_k$.

Se asumen que las $\pi_k > 0$, para todo $k \in U$, esta condición es necesaria para que el diseño muestral sea un diseño muestral probabilístico. Otra propiedad importante de un diseño ocurre cuando $\pi_{kl} > 0$, para todo $k \neq l \in U$. Si la condición anterior y esta son satisfecha, el diseño muestral se dice medible, esto permite el cálculo de estimaciones de variancias e intervalos de confianza válidos basados en los datos observados.

Para cada elemento de $k \in s$ se tiene una ponderación $w_k = 1/\pi_k$ que refleja la probabilidad

de inclusión en el diseño muestral.

El estimador de Horvitz-Thompson es un estimador lineal que emplea las probabilidades de inclusión de primer orden (π_k). Para el caso de la media, este requiere conocer el tamaño de la población (N). Una alternativa para estimar parámetros poblacionales como la media cuando se desconoce N, es el estimador de Hajek. Este estimador es frecuentemente mejor que el estimador de Horvitz-Thompson, en el caso de que N sea conocido, dado que presenta mejores propiedades. El estimador de Hajek es no lineal, y por lo tanto aproximadamente insesgado, además no se puede obtener una expresión exacta para la variancia.

Estimador de Horvitz-Thompson.

El estimador de Horvitz-Thompson del total poblacional es $\hat{t}_\pi = \sum_S \frac{y_k}{\pi_k}$, mientras que la media poblacional es estimada a través del estimador $\bar{y}_\pi = \frac{\hat{t}_\pi}{N} = \frac{1}{N}\sum_{k\epsilon S}\frac{y_k}{\pi_k}$. Estos estimadores son insesgados del total y la media poblacional respectivamente.

La variancia para el total es

$$V(\hat{t}_{HT}) = \sum\sum_U (\pi_{kl} - \pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

La variancia para la media es

$$V(\hat{\bar{y}}_{U\pi}) = \frac{V(\hat{t}_{HT})}{N^2} = \frac{1}{N^2}\sum\sum_U (\pi_{kl} - \pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

Siempre que $\pi_{kl} > 0$, para todo $k,l \in U$, un estimador insesgado de la variancia para el total es

$$\hat{V}(\hat{t}_\pi) = \sum\sum_S \left(\frac{\pi_{kl}}{\pi_k\pi_l} - 1\right) y_k y_l = \sum\sum_s \frac{\pi_{kl}}{\pi_k\pi_l} y_k y_l - \left(\sum_s y_k\right)^2$$

Y un estimador insesgado de la variancia para la media es

$$\hat{V}(\hat{\bar{y}}_{U\pi}) = \frac{1}{N^2}\sum\sum_S \left(\frac{\pi_{kl}}{\pi_k\pi_l} - 1\right) y_k y_l$$

Estimador de Hájek.

Una alternativa al estimador de Horvitz-Thompson para estimar la media poblacional fue presentada por Hájek (1971), que presentó el siguiente estimador

$$\tilde{y}_H = \frac{\hat{t}_{y\pi}}{\hat{N}} = \frac{\sum_S y_k/\pi_k}{\sum_S 1/\pi_k}.$$

donde $\hat{N} = \sum_S(\frac{1}{\pi_k})$ es el π estimador de N.

Se puede visualizar que este estimador es una razón entre dos estimadores poblacionales.

Una aproximación de la variancia está dada por

$$V(\tilde{y}_H) = \frac{1}{N^2}\sum\sum_U \frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\frac{y_k - \bar{y}_U}{\pi_k}\frac{y_l - \bar{y}_U}{\pi_l}$$

Y un estimador de variancia es

$$\hat{V}(\tilde{y}_H) = \frac{1}{\hat{N}^2} \sum\sum_S \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k - \tilde{y}_H}{\pi_k} \frac{y_l - \tilde{y}_H}{\pi_l}$$

Para algunos diseños los estimadores de Horvitz-Thompson y Hajek son idénticos, es decir, ellos producen el mismo valor para todas las muestras, como en los diseños: simple al azar y muestreo estratificado simple al azar. Cuando el tamaño de la población se desconoce no es posible elegir entre los estimadores, con lo cual sólo se puede utilizar el estimador de Hajek. Sin embargo si N es conocido y los estimadores difieren se debe elegir un estimador. Como ya se ha indicado es preferible usar el estimador de Hajek por tres motivos. En primer lugar, se ha visto que por la forma de la variancia de Hajek, éste es preferible cuando los valores $y_k - \tilde{y}_S$ son todos cercanos a la media.

En segundo lugar, tiene un mejor rendimiento en casos donde el diseño no es fijo, o sea, el tamaño de muestra es variable. Si el tamaño muestral pasa a ser mayor que el promedio, la suma del numerador y la suma del denominador tendrán relativamente más términos. Análogamente, si el tamaño de la muestra es pequeño, ambas sumas tendrán pocos términos. La relación conserva de este modo una cierta estabilidad. Al contrario del estimador de Horvitz-Thompson que al tener denominador fijo, carece de esta estabilidad.

Y finalmente en tercer lugar, en casos donde π_k está mínimamente (o negativamente) correlacionado con los valores y_k. Si se supone que la muestra contiene un elemento con un gran valor de y_k pero un pequeño valor de π_k, la suma del numerador será muy grande. Sin embargo, será compensada hasta cierto punto por grandes valores de $1/\pi_k$ en el denominador. En este sentido, el estimador de Hajek es mejor ya que el denominador de N permanece fijo.

ESTIMADORES ROBUSTOS

En muchas encuestas por muestreo, es común la aparición de valores que se alejan del común de los datos, denominados outliers. Los mismos pueden ser observaciones que se corresponden con valores observados y que resultan ser válidos. Estos valores afectan los estimadores tradicionales debido a que los mismos son sensibles ante la aparición de uno o más outliers. Una opción es la de ignorar esos valores. Desechar un valor outliers válido hará que los estimadores clásicos se vuelvan sesgados, pero mantenerlo con una ponderación completa hará al estimador altamente variable porque el valor atípico se presentará sólo en algunas de las muestras posibles.

Por otro lado el outlier puede ser una observación incorrecta, debido a mediciones o codificación errónea o derivada de un elemento fuera de la población objetivo. En este caso manteniendo el outlier con una completa ponderación puede implicar un gran sesgo sumado a una gran variabilidad para los estimadores clásicos. De este modo, descartar los valores atípicos incorrectos reduce tanto el sesgo como la variancia.

Ya que es frecuentemente difícil detectar outliers y decidir si estos son válidos o no, son necesarios los estimadores que se desempeñan bien en términos de sesgo y varianza con independencia de la naturaleza y la detección de posibles valores atípicos.

Se verán seis de los estimadores robustos más difundidos desarrollados hasta la actualidad.

Media Winsorizada.

Una ponderación w_k se adjunta a cada observación de la muestra. La ponderación refleja la probabilidad de inclusión del diseño muestral, corrección por no respuesta y calibración. Se asume que $\sum_{k\epsilon S} w_k = N$.

Sean las observaciones ordenadas $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$ y $w_{[k]}$ la ponderación de $y_{(k)}$. La suma parcial de las ponderaciones de las observaciones ordenadas son definidas como $s_j = \sum_{k=1}^{n} w_k$. Es decir, s_j es la estimación de la función de distribución de y en el punto $y_{(j)}$.

Una versión ponderada de la media winsorizada es definida de la siguiente manera.

Se selecciona un $\alpha \epsilon [0; 0.5)$ y se busca el índice j_l y j_u con

$$j_l = \min\{j/s_j \geq \alpha\}$$

$$j_u = \text{máx}\left\{\{j/s_j < 1 - \alpha\}, j_l\right\}$$

La media winsorizada ponderada es:

$$T_W = \frac{1}{\sum_S w_k}\left(\sum_{j=j_l}^{j_u} w_{[j]} y_{(j)} + \sum_{j=1}^{j_l-1} w_{[j]} y_{(j_l)} + \sum_{j=j_u+1}^{n} w_{[j]} y_{(j_u)}\right)$$

Media truncada.

La media truncada simplemente establece la ponderación de observaciones fuera de j_l y j_u igual a cero, donde j_l y j_u son los mismos índices que para una media winsorizada. Más formalmente

$$u_{[j]} = \begin{cases} 1 \text{ si } \alpha \leq \dfrac{\sum_{i=1}^{j} w_{[i]}}{\sum_{i=1}^{n} w_i} < 1 - \alpha \\ 0 \text{ en otro caso} \end{cases}$$

$$T_T = \frac{\sum_S w_k u_k y_k}{\sum_S w_k u_k}$$

Este estimador puede tener un mayor sesgo que el estimador anterior.

Estimador univariado un paso.

Se necesita un valor inicial para hacer un paso. Sea T0 la primera estimación de la media poblacional, puede ser cualquier estimador mencionado, pero si se quiere robustez, es preferible no utilizar la el estimador de Horvitz-Thompson o de Hajek como valor inicial. Se tomarán como outliers a las observaciones que están alejadas de la estimación inicial. Por lo tanto se define una escala residual según

$$\hat{\sigma} = \text{med}(|y_k - T_0|, w_k)/0{,}67$$

Esta es la desviación media absoluta (MAD) en caso de igual ponderación y siendo T0 la mediana.

Se elige un término constante c > 0 y se definen ponderaciones robustas.

$$u_k = \begin{cases} 1 & \text{si } |y_k - T_0| \le c\hat{\sigma} \\ \dfrac{c\hat{\sigma}}{|y_k - T_0|} & \text{si} |y_k - T_0| > c\hat{\sigma} \end{cases}$$

Finalmente un estimador univariado un paso es definido como la media ponderada

$$T_{0S}(c) = \frac{\sum_S w_k u_k y_k}{\sum_S w_k u_k}$$

Estimador HT robusto un paso.

Se supone que una medida positiva del tamaño x_k es conocida antes de sacar la muestra y tiene correlación positiva con las variables de interés de la encuesta. Se denota con x_{U+} el total poblacional de x_k.

Para una estrategia Horvitz-Thompson la ponderación w_k es la inversa de la probabilidad de inclusión $1/\pi_k$ con $\pi_k = nx_k/x_{U+}$.

Como ya observamos en la sección 2 el estimador Horvitz-Thompson es $T_{HT} = \frac{1}{N}\sum_S w_k y_k$.

El modelo que inspira el estimador Horvit-Thompson es $y_k = \beta x_k + E_k$ con $E(E_k) = 0$ y $Var(E_k) = x_k\sigma^2$. El modelo ayuda a en el desarrollo de la robustificación. Así el residual por la robustificación del HT-estimador es

$$r_k(\beta) = \frac{y_k - \beta x_k}{\sqrt{x_k}} = \frac{y_k - \beta x_{U+}/(nw_k)}{\sqrt{x_{U+}/(nw_k)}}$$

Luego se reemplaza x_i con $x_{U+}/(nw_k)$ y β con NT_0/x_{U+}. Así se utiliza a NT0 como primer estimación del HT robusto un paso. T_0 es un estimador inicial de la media poblacional. Se asume conocido x_{U+} y se calcula el residuo empírico:

$$r_k(T_0) = \frac{y_k - NT_0/(nw_k)}{\sqrt{x_{U+}/(nw_k)}}$$

Para la robustificación, se necesita un estimador de la escala de los residuos, σ o de $\sigma\sqrt{x_{U+}}$. Se usa la mediana del residuo absoluto $\hat{\sigma} = \text{med}(|r_k(T_0)|, w_k)/0{,}67$. Siendo c un término constante, se construye la ponderación robusta.

$$u_k = \begin{cases} 1 \text{ si } |r_k| \le c\hat{\sigma} \\ \dfrac{c\hat{\sigma}}{|r_k|} \text{ en otro caso} \end{cases}$$

Finalmente el estimador HT robusto un paso es

$$T_{HTS} = \frac{1}{N}\frac{\sum_S w_k u_k y_k}{\sum_S u_k/n}$$

Si se reemplaza N por $\sum_S w_i$, así se pasa del estimador HT al estimador de Hajek. El estimador resultante es

$$T_{HS} = \frac{\sum_S w_k u_k y_k}{\sum_S w_k \sum_S u_k/n}$$

Estimador de razón un paso.

Se introdujo información auxiliar determinadas las probabilidades de inclusión en la estrategia Horvitz-Thompson. Se supone una variable auxiliar z_k positiva para toda la población, correlacionada con las variables de interés y que la media poblacional es conocida. En esta situación la estimación clásica para la media poblacional de y es el estimador de razón

$$T_R = \bar{z}_U \frac{\sum_S w_k y_k}{\sum_S w_k z_k}$$

Para una ponderación constante el estimador de razón es el mejor estimador lineal insesgado bajo el modelo $y_k = \beta z_k + E_k$ con $V(E_k) = z_k \sigma^2$.

Se construye un estimador de razón un paso de la siguiente manera: se comienza con un estimador robusto β_0 de la pendiente, también se elige la media ponderada de la pendiente como valor inicial. Luego, se estima la desviación estándar de los residuos según la mediana de los residuos absolutos estandarizados. Se define ponderaciones robustas

$$u_k = \begin{cases} 1 \text{ si } |r_k| \leq c\hat{\sigma} \\ \dfrac{c\hat{\sigma}}{|r_k|} \text{ en otro caso} \end{cases}$$

Los valores extremos pueden tener todavía influencia en la estimación.

Finalmente se calcula un estimador robusto ponderado de la pendiente: $\hat{\beta}_{RS} = \frac{\sum_{k \epsilon S} w_k u_k y_k}{\sum_{k \epsilon S} w_k u_k z_k}$ y el estimador de la media poblacional de y es

$$T_{RS} = \bar{z}_U \hat{\beta}_{RS}$$

Estimadores de variancia.

Para estimar variancias se va a utilizar el método de linearización por series de Taylor. Para el caso de los estimadores truncados y winsorizados el desarrollo es simple y una extensión del estimador de una razón, como lo es el estimador de Hájek. Por ejemplo, para el caso de la media truncada, un estimador aproximado de la variancia viene dado por

$$\hat{V}(T_T) = \frac{1}{\hat{t}_u{}^2} \sum \sum_S \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k u_k - u_k T_T}{\pi_k} \frac{y_l u_l - u_l T_T}{\pi_l}$$

donde $\hat{t}_u = \sum_s w_k u_k$.

Para el caso del estimador de razón un paso, se deriva un estimador de variancia usando la definición implícita del estimador a través de la ecuación estimable

$$\sum_S w_k u_k (r_k(\beta_0)) r_k(\beta) \frac{x_k}{\sqrt{x_k}} = 0$$

La solución para la ecuación con $\beta_0 = \beta$ es el M-estimador con la función ψ de Huber. Para la estimación de la variancia se utiliza el residuo no estandarizado

$$e_k = \frac{y_k - \hat{\beta}_{RS} x_k}{\sqrt{x_k}} \frac{x_k}{\sqrt{x_k}} = y_k - \hat{\beta}_{RS} x_k$$

Se tratará la ponderación robusta $u_k = u_k(r_k(\beta_0))$ como una variable observada. Linearizando la ecuación estimable alrededor de $\hat{\beta}_{RS}$, se obtiene la variancia aproximada

$$Var(\hat{\beta}_{RS}) \approx \frac{1}{(\sum_S w_k u_k x_k)^2} Var\left(\sum_S w_k u_k e_k\right)$$

Para estimar $Var(\sum_S w_k u_k e_k)$ se utilizará la fórmula para estimar variancias en muestreo con reemplazo y se llega al siguiente estimador de variancia de $T_{RS} = \bar{x}_U \hat{\beta}_{RS}$

$$v(T_{RS}) = \frac{(\bar{x}_U)^2 n}{(\sum_S w_k u_k x_k)^2} d^2$$

donde $d^2 = \left(\sum_S \left(w_k u_k e_k - \sum_S \frac{w_k u_k e_k}{n}\right)^2\right)/(n-1)$. Puede incluirse una corrección por población finita $(1 - n/N)$ con el riesgo conocido de subestimación de variancia.

Para el estimador HT robusto un paso el residuo no estandarizado es $e_k = y_k - NT_{HTS}/(nw_k)$. La variancia del estimador resulta igual a

$$v(T_{HTS}) = \frac{1}{N^2} \frac{n}{n-1} \frac{\sum_S (w_k u_k e_k)^2}{(\sum_S u_k/n)^2}$$

Si se conoce las probabilidades de inclusión de segundo orden, se puede utilizar el estimador de variancia de Yates-Grundy-Sen o el estimador de variancia de Horvitz-Thompson.

Para el estimador univariado un paso un paso se tiene

$$v(T_{MS}) = \frac{n}{n-1} \frac{\sum_S (w_k u_k e_k)^2}{(\sum_S w_k u_k)^2}$$

donde $e_k = y_k - T_{0S}$.

ESTIMADORES PRESENTES EN EL PROGRAMA ESTADÍSTICO R

El programa estadístico R constituye uno de los programas más utilizado en la actualidad debido, entre otros aspectos, a que su licencia es libre. Por otro lado, existe un gran desarrollo de códigos con metodologías actuales debido al empleo que tiene el mismo en el ámbito de la investigación. Se presenta a continuación una serie de paquetes del programa R que permiten estimar en el ámbito del muestreo en poblaciones finitas utilizando los estimadores presentados en esta revisión.

El paquete *survey* (Lumley, 2010) permite el cálculo de los estimadores de Horvitz-Thompson y de Hájek y sus correspondientes estimadores insesgados de variancia, para diversos diseños muestrales. Se pueden mencionar las siguientes funciones:

- *svytotal* Calcula el total de la variable.
- *svymean* Calcula la media de la variable.

Para el uso de los métodos Yates y Grundy o Horvitz-Thompson tiene que proporcionarse la matriz cuadrada de probabilidades de inclusión de segundo orden. Además, para el método

de Hajek se puede especificar un vector opcional de probabilidades de inclusión de primer orden. A menos que se utilice un método aproximado de Hajek.

El argumento de variancia opcional del estimador de Horvitz–Thompson (variance="HT") o el estimador de Yates–Grundy (variance="YG"), con el valor predeterminado “HT”.

Con respecto a los estimadores robustos, no existe disponible, al día de la fecha, en la página de descarga del programa R ningún paquete que los considere. El paquete *rhte* (Hulliger, 2011) es uno de los que incluye los mismos y se obtiene a partir del pedido a sus autores Permite el cálculo del estimador de la media winsorizada, de la media truncada, del estimador univariado un paso, del estimador HT robusto un paso, y del estimador de razón un paso junto con el cálculo de sus respectivas variancias. Dado que funciona en forma conjunta con el paquete *survey,* comparte con éste la posibilidad de cálculo para los mismos diseños muestrales. Las funciones que permiten el cálculo de los estimadores robustos son:

- *msvymean* Calcula el estimador robusto de Horvitz-Thompson de las estimaciones de la media ponderada o media para las muestras complejas, utilizando la estimación M.
- *msvyratio* Calcula un estimador de razón robusto para muestas complejas usando el estimador M.
- *msvymean* Calcula la media truncada y winsorizada para muestras complejas.
- *msvytotal* Calcula el estimador M del total de muestras complejas.

CONCLUSIONES

Se presentaron estimadores clásicos y robustos para la estimación de parámetros de poblaciones finitas a partir de muestras seleccionadas en forma probabilística. Los primeros poseen el inconveniente de ser sensibles ante la aparición de valores atípicos. Una solución surge a partir del empleo de estimadores denominados robustos, los cuales son menos sensibles ante la existencia de outliers. Se presenta un conjunto de funciones existentes en el programa R que permite el cálculo de los estimadores y de sus correspondientes estimaciones de variancia.

En estudios futuros se planea la evaluación de los estimadores clásicos y robustos a partir de simulaciones considerando diversos diseños muestrales y datos contaminados con distintos números de observaciones atípicas.

REFERENCIAS BIBLIOGRÁFICAS

Hulliger, B. (1995). *Outlier Robust Horvitz-Thompson Estimators.* Survey Methodology, 21, 79-87.

Hulliger, B. (1999). Simple and Robust Estimators for Sampling. Proceedings of the Survey Research Methods Section, American Statistical Association, 1999, 54-63

Hulliger, B. Alfons, A., Filzmoser, P., Meraner, A., Schoch, T., Templ, M. (2011) *R Programmes for Robust Procedures Including Manual.* AMELI Deliverable D4.1. AMELI Project.

Lohr, S. (1999). *Sampling: Design and Analysis, 2nd Edition.* Cengage Lerning.

Lumley, T. (2010). *Complex Surveys: A guide to Analysis Using R*. New Jersey: Wiley & Sons.

Maronna, R.A., Martin, R.D., Yohai, V.J. (2006). *Robust Statistics.* New York: John Wiley & Sons.

Särdnal, C.E., Swensson, B., Wretman, J. (2003). *Model Assisted Survey Sampling.* Springer & Verlag.