ESTIMACION ROBUSTA EN POBLACIONES FINITAS UTILIZANDO MODELOS DE REGRESION MULTIPLE

Mg. : Máximo Mitacc Meza
Lic.: Félix Vásquez Urbano
Lic.: Carlos Castillo Crespo

1.- INTRODUCCION

El objetivo del presente trabajo es presentar una teoría de estimación robusta para el total poblacional de la variable de interés Y, esto es, de

$$T = \sum_{i=1}^{N} Y_i$$

Semejante a la desarrollada por Royall y Herson (1973a) y Josemar Rodrigues (1984). Esta teoría se desarrolla introduciendo restricciones a la variable auxiliar X y admitiendo que la variable de interés Y es modelada mediante el modelo de regresión múltiple o polinomial. Para tal fin, consideremos la población de interes con N elementos que denotamos con P = { 1, 2, ..., N}. Asociada a la i-ésima unidad poblacional hay un par de números (x_i, y_i), i = 1, 2, ... , N, con x_i conocido e y_i fijo pero desconocido. Por ejemplo, las unidades podrían ser ciudades del Perú con x_i el número de habitantes aptos para el trabajo e y_i el número de habitantes desempleados en la i-ésima ciudad en un mes en particular. Una muestra **s** de tamaño n es seleccionada de la población y los valores de Y asociados con las unidades muestrales son observados. Los números $Y_1, \ldots, Y_N$, cuya suma debemos estimar, son tratados como valores de variables aleatorias independientes $Y_1, \ldots, Y_N$. El valor esperado y la varianza de Y_i depende de x_i y serán denotados por βx_i y $v(x_i)$, respectivamente. Luego, el modelo de regresión simple (ξ- modelo) podemos escribirlo

$$Yi = \beta xi + ei\sqrt{v(xi)} \qquad ,i = 1,2,\ldots N. \qquad (1)$$

donde e_i es una variable aleatoria con $E[e_i] = 0$, $Var[e_i] = \sigma^2$, e_i y e_k son independientes para $i \neq k$, β es un parámetro desconocido y $v(x)$ es una función conocida con $v(x) > 0$, $\forall\ x > 0$.

Si adoptamos el modelo (1), obtendremos una condición tal que el estimador insesgado del total poblacional bajo el modelo anterior se convierta en estimador insesgado bajo el modelo de regresión múltiple.

Utilizamos la notación $\xi(\delta_0, \delta_1, .. \delta_p : v)$, el cual fue introducido por Royall y Herson (1973a), para denotar el modelo de regresión múltiple

$$Y_i = \sum_{j=0}^{p} \delta_j \beta_j x_{ij} + e_i \sqrt{v_i} \quad ,i = 1,2,\ldots,N \tag{2}$$

donde:

i) e_i es una variable aleatoria con $E[e_i] = 0$ y $Var[e_i] = \sigma^2$, e_i y e_k son independientes para $i \neq k$.

ii) $x_{i0} = 1, x_{i1},\ldots,x_{ip}$ son cantidades auxiliares conocidas para $i = 1, 2,\ldots,N$.

iii) a) $v_i = v(x_i)$, si $p = 0$, siendo $v(x)$ una función conocida y

b) $v_i = v(\delta_0, \delta_1 x_{i1},\ldots, \delta_p x_{ip})$, si $p \geq 1$, $i = 1, 2,..,N$.

iv) δ_j toma el valor uno o cero, esto es,

$$\delta_j = \begin{cases} 1, & \text{si } \beta_j x_{ij} \text{ aparece en el modelo} \\ 0, & \text{si } \beta_j x_{ij} \text{ no aparece en el modelo} \end{cases}$$

$j = 0, 1, 2, \ldots,p.$

2. ELECCION DEL ESTIMADOR OPTIMO PARA EL TOTAL POBLACIONAL UTILIZANDO MODELOS DE REGRESION MULTIPLE

DEFINICIÓN 2.1.- Sea **s** una muestra de tamaño n seleccionada de la población finita **P** y ξ el modelo de superpoblación adoptado. Se dice que un estimador $\hat{\theta}$ del parámetro θ es ξ-insesgado si:

$$E_\xi\left[\hat{\theta} - \theta\right] = 0, \forall \xi \tag{3}$$

donde el subíndice indica que la esperanza estimada con respecto a la distribución de probabilidad del modelo ξ.

DEFINICIÓN 2.2.- El error cuadrático medio del estimador $\hat{\theta}$ bajo el modelo ξ es dado por:

$$ECM_\xi\left[\hat{\theta}\right] = E_\xi\left[(\hat{\theta} - \theta)^2\right] = Var_\xi\left[\hat{\theta}\right] + \left\{E_\xi\left[\hat{\theta} - \theta\right]\right\}^2 \tag{4}$$

Si $\hat{\theta}$ es un estimador ξ- insesgado de θ, entonces :

$$ECM_{\xi}\left[\hat{\theta}\right] = Var_{\xi}\left[\hat{\theta}\right]$$

Por el teorema de Gauss - Markov, el estimador lineal ξ-insesgado de θ para el total poblacional **T** bajo el modelo (1) es el siguiente :

$$\hat{T} = \sum_{i=1}^{n} Yi + \left[\frac{\sum_{i=1}^{n} \frac{xiyi}{v(xi)}}{\sum_{i=1}^{n} \frac{xi^2}{v(xi)}}\right] \sum_{i=n+1}^{N} xi \qquad (5)$$

OBSERVACIÓN 2.1.- **i)** Si $v(x) = 1$, entonces el estimador lineal ξ- insesgado para el total poblacional **T** es

$$\hat{T}_0 = \sum_{i=1}^{n} Yi + \left[\frac{\sum_{i=1}^{n} xiyi}{\sum_{i=1}^{n} xi^2}\right] \sum_{i=i+1}^{N} xi \qquad (6)$$

ii) Si $v(x) = x$, de (5) el estimador lineal ξ- insesgado para el total poblacional **T** es el estimador de razón

$$\hat{T}_1 = \hat{T}_R = \left[\frac{\sum_{i=1}^{n} yi}{\sum_{i=1}^{n} xi}\right] \sum_{i=1}^{N} xi \qquad (7)$$

iii) Si $v(x) = x^2$, entonces el estimador lineal ξ-insesgado para el total poblacional **T** es

$$\hat{T}_2 = \sum_{i=1}^{n} Yi = \frac{1}{n}\left[\sum_{i=1}^{n}\left(\frac{Yi}{Xi}\right)\right] \sum_{i=n+1}^{N} xi \qquad (8)$$

iv) Para el modelo (1) con $v(x) = x^2$, Horvitz - Thompson empleando el plan de muestreo p_i, donde p_i es la probabilidad de inclusión definido por

$$\pi_i = \frac{n Xi}{\sum_{j=1}^{N} Xj}, i = 1, 2, \ldots, N.$$

obtiene el estimador lineal ξ-insesgado para el total poblacional **T** dado por

$$\hat{T}_{HT} = \frac{1}{n}\left[\sum_{i=1}^{n}\left(\frac{Y_i}{X_i}\right)\right]\sum_{i=1}^{N} x_i \qquad (9)$$

LEMA 2.1.- Si para cualquier plan de muestreo **p**, los estimadores $\hat{\beta}_1$ y $\hat{\beta}_2$ del parámetro β del modelo ξ = (1) que generan a los estimadores $\hat{T}_1$ y $\hat{T}_2$, respectivamente, satisfacen la condición de que

$$E_\xi\left[\left(\hat{\beta}_1-\beta\right)^2\right] \le E_\xi\left[\left(\hat{\beta}_2-\beta\right)^2\right] \qquad (10)$$

para cada muestra **s** tal que p(s) > 0, entonces

$$ECM_\xi\left[\hat{T}_1\right] \le ECM_\xi\left[\hat{T}_2\right] \qquad (11)$$

DEMOSTRACIÓN

Después de seleccionar la muestra, los los estimadores $\hat{T}_1$ y $\hat{T}_2$ pueden escribirse como

$$\hat{T}_1 = \sum_{i=1}^{n} Y_i + \hat{\beta}_1 \sum_{i=n+1}^{N} x_i$$

$$\hat{T}_2 = \sum_{i=1}^{n} Y_i + \hat{\beta}_2 \sum_{i=n+1}^{N} x_i$$

Es fácil verificar que

$$E_\xi\left[\left(\hat{T}_1 - T\right)^2\right] = \left(\sum_{i=n+1}^{N} x_i\right)^2 E_\xi\left[\left(\hat{\beta}_1 - \beta\right)^2\right] + \sigma^2 \sum_{i=n+1}^{N} v(x_i)$$

Por lo tanto, si (10) es válido, entonces $ECM_\xi\left[\hat{T}_1\right] \le ECM_\xi\left[\hat{T}_2\right]$

También se puede probar que, $\hat{T}$ es un estimador ξ-insesgado para el total poblacional **T**, si y solo si, el estimador generado $\hat{\beta}$ es ξ-insesgado para el parámetro **ß** del modelo (1).

TEOREMA 2.1.- Sean $\hat{T}_0, \hat{T}_1$ y $\hat{T}_2$ estimadores lineales ξ-insesgados para **T** obtenidos en la observación 2.1.

i) Si la variable auxiliar x es tal que $0 < x < 1$, entonces

$$ECM_\xi\left[\hat{T}_2\right] \le ECM_\xi\left[\hat{T}_1\right] \le ECM_\xi\left[\hat{T}_0\right]$$

ii) Si la variable auxiliar x es tal que $x \ge 1$, entonces

$$ECM_\xi\left[\hat{T}_0\right] \le ECM_\xi\left[\hat{T}_1\right] \le ECM_\xi\left[\hat{T}_2\right]$$

DEMOSTRACIÓN

i) Para el supuesto $0 < x < 1$, se prueba facilmente que

$$\mathrm{var}_\xi\left[\hat{T}_2\right] \le \mathrm{var}_\xi\left[\hat{T}_1\right] \le \mathrm{var}_\xi\left[\hat{T}_0\right]$$

Del Lema 1 se sigue que,

$$ECM_\xi\left[\hat{T}_2\right] \le ECM_\xi\left[\hat{T}_1\right] \le ECM_\xi\left[\hat{T}_0\right]$$

ii) Si $x \ge 1$, entonces los estimadores generadores $\hat{\beta}_0, \hat{\beta}_1$ y $\hat{\beta}_2$ satisfacen $\mathrm{var}_\xi\left[\hat{\beta}_0\right] \le \mathrm{var}_\xi\left[\hat{\beta}_1\right] \le \mathrm{var}_\xi\left[\hat{\beta}_2\right]$

Luego por el Lema 2.1 se sigue que

$$ECM_\xi\left[\hat{T}_0\right] \le ECM_\xi\left[\hat{T}_1\right] \le ECM_\xi\left[\hat{T}_2\right]$$

TEOREMA 2.2.- Si $\max_{1 \le i \le N}\{nx_i\} \le \sum_{j=1}^{N} x_i$, entonces para cualquier plan de muestreo **p** tal que p(s) > 0, se cumple $ECM_\xi\left[\hat{T}_2\right] \le ECM_\xi\left[\hat{T}_{HT}\right]$

La prueba de este Teorema es similar al Del Teorema 2.1. y Teorema 2.2, es evidente que para $x \ge 1$ se cumple la siguiente desigualdad

$$ECM_\xi\left[\hat{T}_0\right] \le ECM_\xi\left[\hat{T}_1\right] \le ECM_\xi\left[\hat{T}_2\right] \le ECM_\xi\left[\hat{T}_{HT}\right]$$

Por tanto, $\hat{T}_0$ es ξ-óptimo para $x \ge 1$.

3.- ELECCION DEL ESTIMADOR ROBUSTO PARA EL TOTAL DE UNA POBLACION FINITA UTILIZANDO MODELOS DE REGRESION MULTIPLE.

El rol fundamental al seleccionar muestras aleatorias es el obtener muestras que sean representativas de la población en estudio. Cuando aparecen valores extremos o no representativos se origina cierto rechazo de parte del investigador, quien trata de hacer los ajustes necesarios para dividir las unidades de la población en muestrales y no muestrales.

Si no se conoce lo suficiente acerca de la relación entre las variables X e Y que nos permita corregir el no balanceo de la muestra, entonces debemos decidirnos por una muestra que sea representativa con respecto a la variable auxiliar X. Una manera de cumplir con esta restricción es a través de alguna forma de *ALEATORIZACIÓN RESTRINGIDA*.

DEFINICIÓN 3.1 Para cualquier entero positivo p, se dice que la muestra aleatoria **s = s*(p)** de tamaño n seleccionada de una población finita **P** satisface la *CONDICIÓN DE ALEATORIZACIÓN RESTRINGIDA* con respecto a la variable auxiliar x, si

$$\frac{\sum_{i=n+1}^{N} x_{ij}}{N-n} = \frac{\sum_{i=1}^{n} \frac{x_{ij}}{v(x_i)}}{\sum_{i=1}^{n} \frac{1}{v(x_i)}} \qquad j = 0,1,\ldots,p \qquad (12)$$

donde $v(x)$ es una función conocida.

Por el teorema de Gauss-Markov el estimador lineal insesgado uniformemente de mínima varianza (E.I.U.M.V) del total poblacional **T** bajo el modelo $\xi(1: v(x)); (Y_i = \beta_0 + e_i\sqrt{v(x_i)}$, i=1,2,...,N) es el siguiente

$$\hat{T}(1: v(x)) = \sum_{i=1}^{n} Y_i + \frac{(N-n)\sum_{i=1}^{n} \frac{Y_i}{v(x_i)}}{\sum_{i=1}^{n} \frac{1}{v(x_i)}} \qquad (13)$$

TEOREMA 3.1.- Si **s = s*(p)** es una muestra aleatoria de tamaño n seleccionada de una población finita **P**, entonces $\hat{T}(1: v(x))$ es insesgado bajo el modelo de regresión $\xi(\delta_0, \delta_1, \ldots, \delta_p: v)$, para cualquier función v.

DEMOSTRACIÓN

tenemos:

$$E_\xi\left(\hat{T}(1: v(x)) - T\right) = E_\xi\left[\frac{(N-n)\sum_{i=1}^{n} \frac{Y_i}{v(x_i)}}{\sum_{i=1}^{n} \frac{1}{v(x_i)}} - \sum_{i=n+1}^{n} Y_i\right]$$

$$= \frac{(N-n)\sum_{i=1}^{n}\sum_{j=0}^{p} \frac{\delta_j\beta_j x_{ij}}{v(x_i)}}{\sum_{i=1}^{n} \frac{1}{v(x_i)}} - \sum_{i=n+1}^{N}\sum_{j=0}^{p} \delta_j\beta_j x_{ij} = \sum_{j=0}^{p} \delta_j\beta_j \left[\frac{(N-n)\sum_{i=1}^{n} \frac{x_{ij}}{v(x_i)}}{\sum_{i=1}^{n} \frac{1}{v(x_i)}} - \sum_{i=n+1}^{n} x_{ij}\right] = 0$$

si s = s*(p).

Así, si seleccionamos una muestra aleatoria $\mathbf{s} = \mathbf{s}^*(\mathbf{p})$ de tamaño n, el estimador $\hat{T}(1: v(x))$ es robusto en el sentido de que es insesgado bajo cualquier modelo de regresión múltiple.

OBSERVACIÓN 3.1.- Si **s** es una muestra aleatoria de tamaño n seleccionada de una población finita **P**; entonces el estimador lineal insesgado uniformemente de mínima varianza del total poblacional **T** bajo el modelo de superpoblación ξ (0, 0,.., 0, δ_j = 1,..0 .., 0: $v(x)x_{ij}$) es

$$\hat{T}_j = \hat{T}_j\left(0,\ldots,\delta_j = 1,\ldots,0: v(x)x_{ij}\right) = \sum_{i=1}^{n} Y_i + \left(\sum_{i=n+1}^{N} x_{ij}\right)\left[\frac{\sum_{i=1}^{n}\frac{Y_i}{v(x_i)}}{\sum_{i=1}^{n}\frac{x_{ij}}{v(x_i)}}\right]$$

$j = 0,1,2,\ldots,p.$

TEOREMA 3.2.- Si $\mathbf{s} = \mathbf{s}^*(\mathbf{p})$ es una muestra aleatoria de tamaño n seleccionada de una población finita **P**, entonces $\hat{T}(1: v(x))$ es un E.I.U.M.V. de T bajo el modelo de regresión múltiple $\xi\left(\delta_0, \delta_1, \ldots, \delta_p : v^*(\underline{x})\right)$, donde

$$v^*(\underline{x}) = v(\underline{x})\sum_{j=0}^{p} a_j \delta_j x_{ij} \qquad i = 1,2,\ldots,N$$

$$x = \left(x, \delta_0, \delta_1 x_{i1}, \ldots, \delta_p x_{ip}\right) \quad \text{y} \quad a_j > 0 \quad j = 0,1,\ldots,p$$

DEMOSTRACIÓN

Si $\mathbf{s} = \mathbf{s}^*(\mathbf{p})$, entonces de la observación 3.1 obtenemos:

$$\hat{T}_j\left(0,\ldots 0, \delta_j = 1, 0,\ldots, 0: v(x)x_{ij}\right) = \hat{T}(1: v(x)) \qquad \forall j = 0,1,\ldots,p$$

Luego, el estimador es un E.I.U.M.V. de **T** bajo el modelo de regresión múltiple ξ (0,..,0, δ_j = 1,0,...0: $v(x)x_{ij}$), para j = 0, 1, 2, ..., p. Ahora, considerando el modelo $\xi_j\left(\delta_0, \delta_1, \ldots, \delta_j = 1, \ldots, \delta_p : v(x)x_{ij}\right)$ se tiene

$$E_{\xi_j}\left[\hat{T}(1: v(x)) - T\right]^2 = \frac{\sigma^2(N-n)^2}{\left[\sum_{i=1}^{n}\frac{1}{v(x_i)}\right]^2}\cdot\sum_{i=1}^{n}\frac{x_{ij}}{v(x_i)} + \sigma^2\sum_{i=n+1}^{N} v(x_i)x_{ij}$$

Así, el error cuádratico medio de $\hat{T}(1: v(x))$ depende únicamente de la función $v(x)x_{ij}$ y no de los coeficientes β_0, β_1, ..., β_p. Luego, $\hat{T}(1: v(x))$ es un estimador E.I.U.M.V. de **T** bajo el modelo $\xi_j, \forall j = 0,1,2,\ldots,p.$

De igual forma, el error cuadrático medio del estimador $\hat{T}(1:v(x))$ bajo el modelo $\xi(\delta_0,\delta_1,...,\delta_p:v^*(\underline{x}))$, es

$$E_{\xi_j}\left[\hat{T}(1:v(x))-T\right]^2=\sum_{j=0}^{p}a_j\delta_j E_{\xi_j}\left[\left(\hat{T}(1:v(x))-T\right)^2\right]$$

y $\hat{T}(1:v(x))$ es insesgado bajo el modelo $\xi(\delta_0,\delta_1,...,\delta_p:v^*(\underline{x}))$, entonces, $\hat{T}(1:v(x))$ es un estimador E.I.U.M.V. de **T** bajo el modelo ξ.

4.- ELECCION DEL ESTIMADOR ROBUSTO PARA EL TOTAL DE UNA POBLACION FINITA UTILIZANDO MODELOS DE REGRESION POLINOMIAL

Si reemplazamos $x_{ij}=x_i^j$, i = 1,-2,..., N y j = 0, 1, ...,p con el modelo (2) entonces obtenemos un modelo particular $\xi^{P_0}(\delta_0,\delta_1,...,\delta_p:v(x))$, conocido como modelo de regresión polinomial. Este modelo lo podemos escribir así:

$$Y_i=\sum_{j=0}^{p}\delta_j\beta_j x_i^j+e_i\sqrt{v(x_i)} \qquad i=1,2,...,N \tag{15}$$

donde:

i) e_i es una variable aleatoria con E[e_i] = 0, Var[e_i] = σ^2, e_i y e_k son independientes para $i\neq k$.

ii) $x_i, x_i^2,...,x_i^p$ son cantidades auxiliares conocidas para i = 1, 2,..., N.

iii) $v(x)$ es una función polinomial conocida.

iv) δ_j es un indicador que toma valor cero o uno, es decir

$$\delta_j=\begin{cases}1, & \text{si el término } \beta_j x_i^j \text{ está presente en el modelo}\\ 0, & \text{en el caso contrario}\end{cases}$$

para j = 0, 1, 2, ..., p.

OBSERVACIÓN 4.1.- Si **s** es una muestra aleatoria de tamaño n seleccionada de una población finita **P**; entonces el E.I.U.M.V. del total poblacional **T** bajo el modelo de regresión polinomial $\xi^{P_0}(1: v(x))$ es

$$\hat{T}_{P_0}(1: v(x)) = \sum_{i=1}^{n} Y_i + \frac{(N-n)\sum_{i=1}^{n} \frac{Y_i}{v(x_i)}}{\sum_{i=1}^{n} \frac{1}{v(x_i)}} = \hat{T}_1(1: v(x))$$

Así, el estimador óptimo del total poblacional bajo el modelo $\xi^{P_0}(1: v(x))$ es el estimador de razón $\hat{T}_1$.

Bajo el modelo $\xi^{P_0}(0,.,0,\delta_j = 1,0..0: v(x))$ el sesgo del estimador $\hat{T}_{P_0}(0,1:x)$ para cualquier muestra s de tamaño n seleccionada de la población finita es

$$Sesg\left[\hat{T}_{P_0}(0,1:x)\right] = N\bar{x}\sum_{j=0}^{p} \delta_j \beta_j \left[\frac{\bar{x}_s^{(j)}}{\bar{x}_s} - \frac{\bar{x}^{(j)}}{\bar{x}}\right] \qquad (16)$$

donde

$$\bar{x}_s^{(j)} = \frac{\sum_{i=1}^{n} x_i^j}{n} \quad , \quad \bar{x}^{(j)} = \frac{\sum_{i=1}^{N} x_i^j}{N} \quad , \quad \bar{x}_s = \bar{x}_s^{(1)} \quad , \quad \bar{x} = \bar{x}^{(1)}$$

DEFINICIÓN 4.1.- Para cualquier entero positivo p, se dice que la muestra aleatoria $\mathbf{s} = s_B^*(p)$ de tamaño n seleccionada de una población finita **P** es *MUESTRA BALANCEADA* con respecto al j-ésimo momento, si satisface la condición

$$\bar{x}_s^{(j)} = \frac{\sum_{i=1}^{n} x_i^j}{n} = \frac{\sum_{i=1}^{N} x_i^j}{N} = \bar{x}^{(j)} \quad , \quad \forall j = 0,1,2,..,p. \qquad (17)$$

OBSERVACIÓN 4.2.-

a) Si la muestra seleccionada de una población finita **P** es $\mathbf{s} = s_B^*(p)$, entonces $\hat{T}_{P_0}(0,1:x)$ es un estimador insesgado de **T** bajo el modelo de regresión polinomial $\xi^{P_0}(\delta_0, \delta_1, ..., \delta_p: v(x))$.

b) Análogamente, si la muestra seleccionada de una población finita **P** es $\mathbf{s} = s_B^*(p)$, entonces

$$\hat{T}_{P_0}(0,1:x) = \hat{T}(1:1) = \hat{T}_E = \frac{N}{n}\sum_{i=1}^{n} Y_i$$

Así, $\hat{T}_{P_0}(0,1:x)$ es E.I.U.M.V. de **T** para una muestra $\mathbf{s} = s_B{}^*(p)$. Una condición de aleatorización restringuida para el modelo de regresión polinomial, similar a lo que se dió en la definición 3.1 es el siguiente.

DEFINICIÓN 4.2.- Para cualquier entero positivo p, se dice que la muestra aleatoria $\mathbf{s} = s_{P_0}{}^*(p)$ de tamaño n seleccionada de una población finita **P** satisface la condición de ALEATORIAZACIÓN RESTRINGIDA con respecto a la variable auxiliar x, si

$$\frac{\sum_{i=n+1}^{N} x_i^j}{N-n} = \frac{\sum_{i=1}^{n} \frac{x_i^j}{v(x_i)}}{\sum_{i=1}^{n} \frac{1}{v(x_i)}}, \quad \forall j = 0,1,2,\ldots,p \tag{18}$$

donde $v(x)$ es una función polinomial conocida.

Bajo el modelo $\xi^{P_0}(\delta_0,\delta_1,\ldots,\delta_p:x)$ el error cuadrático medio del estimador $\hat{T}_{P_0}(1:x)$ para una muestra $\mathbf{s} = s_{P_0}{}^*(p)$ de tamaño n seleccionada de una población finita **P**, es

$$E_{\xi^{P_0}}\left[\hat{T}_{P_0}(1:x) - T\right]^2 = \frac{\sigma^2 N(N-n)\bar{x}_r^-}{n}$$

donde $\bar{x}_r^- = \dfrac{\sum_{i=n+1}^{N} x_i}{N-n}$

Para una fracción pequeña de muestreo, $\bar{x}_r^-$ es aproximadamente igual a $\bar{x} = \dfrac{\sum_{i=1}^{N} x_i}{N}$ .Consecuentemente,

$$E_{\xi^{P_0}}\left[\hat{T}_{P_0}(1:x) - T\right] = \frac{\sigma^2 N(N-n)\bar{x}}{n}$$

Por tanto, podemos decir que la expresión (18) proporciona protección contra el sesgo del estimador $\hat{T}_{P_0}(0,1:x)$ bajo el modelo de regresión polinomial $\xi^{P_0}(\delta_0,\delta_1,\ldots,\delta_p:x)$.

OBSERVACION 4.3.- El estimador $\hat{T}_{P_0}(1:x)$ obtenido con la muestra $\mathbf{s} = s^*{}_{P_0}$ es más eficiente que el estimador de expansión $\hat{T}_E$ obtenido con una

muestra balanceada $\mathbf{s} = s_B^*(p)$, ambos bajo el modelo de regresión polinomial $\xi^{P_0}(1,1:v^*(x))$ donde $v^*(x)=\delta_1^2 x+\delta_2^2 x^2$, es decir

$$E_{\xi^{P_0}}\left[\hat{T}_{p^0}(1:x)-T\right]^2 \le E_{\xi^{P_0}}\left[\hat{T}_E-T\right]^2$$

EJEMPLO 1.- La Tabla 2.1 presenta el Producto Bruto Interno y el Consumo Privado en soles de base 1979 de 20 años. Estos datos fueron obtenidos del compendio estadístico del INEI 1992-1993 y es una muestra seleccionada del conjunto de datos correspondientes al período 1950 - 1992, mediante el muestreo aleatorio simple. Por otro lado se fijó una confiabilidad del 95%, de que el error de estimación sea menor que $\sigma/3$, donde σ^2 es la varianza de la parte aleatoria del modelo. Así, considerando la fórmula

$$n=\frac{NZ_{\alpha/2}{}^2\sigma^2}{(N-1)e^2+Z_{\alpha/2}{}^2\sigma^2}$$

para N = 43, $Z_{.025}$ = 1.96, el tamaño de la muestra resulta

$$n=\frac{43(1.96)^2\sigma^2}{42\left(\frac{\sigma}{3}\right)+(1.96)^2\sigma^2}\cong 20$$

De acuerdo con el marco muestral y usando la tabla de números aleatorios, se obtiene los siguientes componentes de la población(identificados con los números aleatorios): 04, 12, 43, 21, 22, 08, 41, 09, 15, 19, 07, 35, 25, 37, 32, 27, 30, 26, 38, 11. Asociados a cada uno de estos números con los años del período 1950 - 1992 se obtiene la muestra de la Tabla 2.1.

Para estimar el total del consumo privado para dicho período consideremos las variables:

X (VARIABLE AUXILIAR) : PBI anual en soles de base 1979.

Y (VARIABLE PRINCIPAL): Consumo privado anual en soles de base 1979.

Luego utilizando la información de la Tabla 2.1 y el programa 2 del apéndice, los estimadores del total del consumo privado para el período 1950 - 1992 son presentados en la tabla 2.2.

TABLA 2.1

Producto Bruto Interno y Consumo Privado (en soles de base 1979) en el período 1950-1992.

AÑO	PRODUCTO BRUTO INTERNO:X (soles de 1979)	CONSUMO PRIVADO:Y (soles de 1979)
1953	1,615.814	989.402
1957	2,293.028	1,625.757
1958	2,623.875	1,848.779
1960	3,573.298	2,255.686
1961	3,807.715	2,355.772
1964	2,518.595	1,785.390
1965	3,494.779	2,209.014
1968	1,301.269	874.020
1970	3,276.074	2,249.321
1971	3,490.135	1,371.070
1974	3,107.387	2,122.115
1975	3,881.284	2,636.455
1976	4,234.711	2,847.162
1981	3,243.760	2,132.609
1984	3,226.301	2,101.883
1985	1,047.951	705.336
1987	1,504.732	926.610
1988	3,213.039	2,209.671
1990	1,293.882	876.953
1992	1,935.367	1,274.222

Fuente: Datos proporcionados por el INEI (1993)

TABLA 2.2
Estimaciones del total del consumo privado en base a los datos de la Tabla 2.1

ESTIMADOR	ESTIMACIONES	VARIANZAS DE LOS ESTIMADORES DEL PARAMETRO β
$\hat{T}_0$	76,104.038	
$\hat{T}_0$	69,980.281	$\text{var}\left[\hat{\beta}_0\right] = 0.00000000596\sigma^2$
$\hat{T}_1 = \hat{T}_R$	70,150.826	$\text{var}\left[\hat{\beta}_1\right] = 0.0000183\sigma^2$
$\hat{T}_2$	70,360.222	$\text{var}\left[\hat{\beta}_2\right] = 0.05\sigma^2$
$\hat{T}_{HT}$	70,571.850	$\text{var}\left[\hat{\beta}_{HT}\right] = 0.063\sigma^2$

Comparando las varianzas de los estimadores del parámetro β, tenemos

$$\text{var}\left[\hat{\beta}_0\right] \leq \text{var}\left[\hat{\beta}_1\right] \leq \text{var}\left[\hat{\beta}_2\right] \leq \text{var}\left[\hat{\beta}_{HT}\right]$$

Por consiguiente, según el Lema 2.1, Teorema 2.1 y Teorema 2.2, para la restricción de la variable auxiliar X (X > 1), se verifica

$$ECM\left[\hat{T}_0\right] \leq ECM\left[\hat{T}_1\right] \leq ECM\left[\hat{T}_2\right] \leq ECM\left[\hat{T}_{HT}\right]$$

Por lo tanto, para un plan de muestreo aleatorio simple, $\hat{T}_0$ es ξ-óptimo.

EJEMPLO 2. (REGRESIÓN MÚLTIPLE)

El conjunto de datos de la tabla 2.3, representa a las importaciones anuales correspondientes al período 1950-1992 (N=43). También se registra, los valores de dos variables auxiliares que se utilizan para estimar el total de importaciones en el período indicado y son:

Variables auxiliares:

X1: PM (índice de precios de las importaciones en base a 1979)
X2: PBI79 (Producto bruto interno en soles de base 1979)

Variable principal

Y : M79 (Importaciones en soles de base 1979).

Fijando una confiabilidad del 95%, de que el error de estimación sea menor que $\sigma/3$, el tamaño de la muestra resulta:

$$n = \frac{NZ^2_{\alpha/2}\sigma^2}{(N-1)e^2 + Z^2_{\alpha/2}\sigma^2} = \frac{43(1.96)^2\sigma^2}{42\left(\frac{\sigma}{3}\right)^2 + (1,96)^2\sigma^2} \cong 20$$

Luego utilizando la tabla de números aleatorios, se obtuvo los siguientes componentes de la población(identificados con los números): 06, 18, 42, 03, 22, 08, 41, 09, 15, 19, 07, 35, 24, 28, 32, 27, 40, 26, 38, 10, que da lugar a la muestra que se presenta en la Tabla 2.3

TABLA 2.3.

Datos de importaciones y variables auxiliares

AÑO	M79(Y)	PM(X_1)	PBI(X_2)
1952	182.268	2.9917	994.860
1955	222.254	3.5459	1168.836
1957	286.889	3.7655	1301.269
1958	253.206	4.6966	1293.882
1959	211.869	5.6473	1341.448
1961	308.143	5.6236	1615.814
1963	384.380	5.4420	1815.554
1966	603.191	6.2679	2284.921
1967	608.549	7.7902	2623.875
1971	681.572	9.5755	2844.345
1973	841.905	17.3443	3213.039
1975	738.450	36.6938	3289.336
1977	591.830	1823.899	3494.779
1985	540.598	5605.679	3573.928
1986	650.715	8226.970	3904.219
1987	747.357	11939.60	4234.711
1989	508.361	2774932	3428.614
1990	628.012	635000000	3323.106
1991	677.343	1070000000	3226.301
1992	543.974	5.7800	2201.563

Fuente: Datos proporcionados por el INEI (1993)

Según (8), el valor del estimador del total poblacional Resulta

$$\hat{T}_2 = \hat{T}_2(0, \delta_2 = 1 : x_{i2}) = \sum_{i=1}^{n} Y_i + \left(\frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_{i2}} \right) \sum_{i=n+1}^{n} x_{i2}$$

$$= 10,210.87 + \left(\frac{10210.87}{51174.4} \right) 57,198.095$$

$$= 21,623.638$$

Por lo tanto, la importación total en el período 1950-1992 fue de 21,623.638 nuevos soles.

CONCLUSIONES

Las principales conclusiones que se deducen del presente trabajo son :

1.- El estimador $\hat{T}_2$ es óptimo cuando la restricción de la variable auxiliar x es $0 < x < 1$ bajo el muestreo aleatorio simple sin reposición .

2.- El estimador $\hat{T}_0$ es óptimo cuando la restricción de la variable auxiliar x es $x \geq 1$, tanto para el muestreo aleatorio simple con reemplazamiento como para el muestreo de probabilidad de inclusión definido por Horvitz-Thompson.

3.- El estimador $\hat{T}(1 : v(x))$ es robusto bajo el modelo de regresión múltiple $(\delta_0, \ldots, \delta_p : v(x^*(x)))$ empleando el muestreo que satisface la condición de aleatorización restringida con respecto a la variable auxiliar .

4.- El estimador $\hat{T}_E$ es robusto bajo el muestreo balanceado con respecto al j-ésimo momento, para estimar el total de una población finita que se ajusta a un modelo de regresión polinomial de grado p y función de varianza proporcional a xj para j = 0,1,...,p .

BIBLIOGRAFIA

1.- Mitacc Meza, Máximo y Castillo Crespo, Carlos y Vásquez Urbano, Félix.- Estimación robusta en poblaciones finitas utilizando modelos de regresión múltiple; Investigación realizada con el CIPI. Facultad de Ingenieria Industrial. Universidad de Lima.

2.- Pereyra, C.A. y Rodrigues, J. 1983. Robust Linear Prediction in Finite Populations. International Statistics Review, 51, Págs: 293-300.

3.- Rodrigues, J., 1984.- Robust Estimation and Finite Population. Probability and Mathematical Statistics, Vol. 4, Fasc. págs 197-207.

4.- Royall, R. M. y Herson, J. , 1973.- Robust Estimation in Finite Populations I. J.A.S.A., Dec 1973, Vol. 68, Nº 344, págs. 880-889.

5.- Scott, A. J., Brewer, K.R.W. y Ho, E.W.H., 1979.- Finite Population Sampling and Robust Estimation. J.A.S.A. June 1978, Vol. 73, Nº 362. págs. 359-361.