Tesis de Maestría

Estimación de varianza para estimadores por calibración bajo diseños complejos bietápicos por conglomerados

Molinari, Claudia P.

2009

Tesis presentada para obtener el grado de Magister de la Universidad de Buenos Aires en Estadística Matemática de la Universidad de Buenos Aires

Estimación de Varianza para Estimadores por Calibración bajo Diseños Complejos Bietápicos por Conglomerados

Claudia P. Molinari

Tesis para optar al título de Magister en Estadística Matemática

Instituto del Cálculo-Facultad de Ciencias Exactas y Naturales
UNIVERSIDAD DE BUENOS AIRES

Director: Lic. Gerardo Mitas

Co-directora: Dra. Elena Martínez

Diciembre 2009

"En los momentos en que el entusiasmo y la fuerza nos abandonan están los que realmente nos aman para impulsarnos de nuevo"

Gracias a todos los que me alentaron a llegar al final de este camino,

en especial a mi esposo y a mis hijos

ÍNDICE

INTRODUCCIÓN

Los estimadores de calibración son utilizados actualmente en la mayoría de las grandes encuestas para la estimación de totales de población. Sin embargo, el cálculo de su varianza no es inmediato y se utilizan aproximaciones.

El estimador más extensamente utilizado es el *estimador de Horvitz Thompson*, siempre que no exista información auxiliar, ya que resulta *admisible* entre los estimadores insesgados. Cuando existe información externa disponible, se consideran estimadores específicos que la utilizan y que mejoran la precisión, aún cuando se resigna la presencia de sesgo. Entendemos por información externa a una o varias variables relacionadas con la variable de estudio, cuyos totales marginales son conocidos a nivel poblacional.

Un enfoque posible es el que lleva a emplear al *Estimador de Regresión Generalizada* (GREG) en el cual el peso inicial de cada unidad, derivado del diseño muestral, es ajustado asumiendo un modelo de regresión lineal empleando la información auxiliar disponible. Sin embargo algunos de esos factores de ajuste pueden ser negativos y producir pesos negativos, lo cual resulta de difícil interpretación.

Deville y Särndal (1992) continuaron la idea de corregir los pesos iniciales proponiendo los *estimadores de calibración* que incluyen al estimador GREG como un caso particular. Estos estimadores surgen de minimizar una distancia (a través de funciones apropiadas) entre los pesos iniciales y los nuevos, llamados calibrados, sujeta a ciertas restricciones originadas por la información auxiliar disponible.

En 1992 Deville y Särndal, demostraron que *cualquier miembro de la clase de los estimadores de calibración es asintóticamente equivalente al estimador GREG* y como consecuencia *su varianza asintótica es la correspondiente a dicho estimador de regresión.* Sin embargo, cuando se utiliza un muestreo de dos etapas *sin reposición* se deben calcular las probabilidades de segundo orden y la expresión del estimador de la varianza del estimador GREG puede tornarse muy complicada de implementar, sobre todo cuando se maneja un volumen muy grande de datos.

En este trabajo se considera el problema de estimar la varianza asintótica del estimador de raking, un estimador que surge por calibración de los pesos originales, cuando se emplea un muestreo de conglomerados en dos etapas, sin reposición.

Para la primera etapa se aplica un muestreo de conglomerados proporcional al tamaño siguiendo el método de Sampford, y un muestreo simple al azar sin reposición, para la segunda etapa.

Para el cálculo de la varianza se analizan otros métodos de aproximación a la varianza del estimador GREG basados en replicaciones, como ser el jackknife y el jackknife linealizado. Las fracciones de muestreo que se emplearán en el diseño no serán pequeñas. Luego, no puede asumirse la equivalencia con un muestreo con reposición, como usualmente se asume en la bibliografía para el cálculo de la estimación de la varianza. Por lo tanto se ensayan junto con las aproximaciones mencionadas, diferentes coeficientes de corrección para ajustar el comportamiento asintótico de las varianzas estimadas.

El estudio parte de un conjunto de datos obtenidos sobre una población conocida, de la cual se extrae una muestra estratificada de conglomerados en dos etapas. Las replicaciones consisten en la extracción de varios miles de muestras para calcular la varianza asintótica, jackknife y jackknife linealizada que se comparan con la varianza poblacional.

Este trabajo está organizado en capítulos de la siguiente forma: en el Capítulo I y II se presentan las nociones básicas de muestreo y estimación; en el Capitulo III se hace una breve reseña teórica sobre los estimadores de calibración y en el capitulo IV sobre la estimación de su varianza. En el Capítulo V se describe el trabajo de simulación y en el Capítulo VI se presentan los resultados obtenidos que dieron lugar a las conclusiones que se detallan en el mismo capítulo.

CAPÍTULO I

NOCIONES BÁSICAS DE MUESTREO

1.1 POBLACIÓN, PARÁMETRO Y MUESTRA.

Se considera una población finita de tamaño N cuyos elementos pueden identificarse con subíndices indicados en el conjunto U={1,2,....,N}.

Se quiere estudiar una característica Y que toma valor y_k sobre la k-ésima unidad de dicha población. Por lo general interesa una función de Y sobre U; en este trabajo consideramos la función total es decir,

$$T_y = \sum_{k \in U} y_k$$

Cuando no es posible conocer los valores de Y sobre todos los individuos de U se observa esta variable sobre un subconjunto de U, a fin de *estimar* T_y. Este subconjunto de observaciones conforma la *muestra,* denotada de aquí en más como s. Formalmente se define a s como un subconjunto de U; o sea:

$s=(i_1,i_2,...i_{n(s)})$ donde $1 \leq i_1 \neq i_2 \neq ... \neq i_{n(s)} \leq N$

Se considerarán muestras seleccionadas a partir de algún esquema de selección probabilístico.

1.2 PLANES DE MUESTREO

Un diseño muestral basado en U consiste de un par (S_d, P_d), donde S_d, llamado soporte del diseño d, es el espacio que contiene todas las muestras posibles generadas a partir de un diseño y P_d es una distribución de probabilidad sobre S_d.

Como toda función de probabilidades, P_d verifica que

1. $P_d(S=s)=P_d(s) \geq 0$
2. $\sum_{s \in S_d} P_d(s) = 1$

En particular, si $P_d>0$ para toda muestra s, el diseño se dice semi-medible.

La pertenencia o no de cada elemento de la población U a la muestra s puede indicarse mediante la variable I_k definida como:

$$I_k = \begin{cases} 1 & si\ k \in s \\ 0 & si\ k \notin s \end{cases} \qquad (1.2.1)$$

Se definen las *probabilidades de primer orden* π_k como la suma de las probabilidades de todas las muestras que incluyen a k *y de segundo orden denotadas como* π_{kl} como la suma de las probabilidades de todas las muestras que incluyen a k y l. Si se indica a todas las muestras que incluyen a k con la expresión $s \ni k$, se pueden denotar estas definiciones de la siguiente forma:

$$\pi_k = P(k \in S\ ) = P(I_k = 1) = E(I_k) = \sum_{s \ni k} p(s) \qquad (1.2.2)$$

$$\pi_{kl} = P(kyl \in S\ ) = P(I_k I_l = 1) = E(I_k I_l) = \sum_{s \ni k,l} p(s) \qquad (1.2.3)$$

Cuando k=l , $\pi_{kk} = P(I_k{}^2 = 1) = P(I_k = 1) = \pi_k$ (1.2.4)

Además puede definirse la covarianza entre I_k e I_l de la forma:

$$\Delta_{kl} = Cov(I_k, I_l) = E(I_k I_l) - E(I_k)E(I_l) = \begin{cases} \pi_k(1-\pi_k) & si\ k = l \\ \pi_{kl} - \pi_k \pi_l & si\ k \neq l \end{cases} \qquad (1.2.5)$$

Luego, $\Delta_{kk} = Var(I_k)$ (1.2.6)

Las probabilidades de primer y segundo orden verifican:

- $\sum_k \pi_k = \sum_s n(s) P_d(s)$; n(s) tamaño de la muestra s
- $\sum_k \sum_{l \neq k} \pi_{kl} = \sum_s n(s)(n(s)-1) P_d(s)$

$$\rightarrow \sum_{l \neq k}^{N} \pi_{kl} = \sum_{s \ni k} (n(s) - 1) P_d(s)$$

$$\rightarrow \sum_k \sum_l \pi_{kl} = \sum_{s \in S_d} [n(s)]^2 P_d(s) \qquad (1.2.7)$$

De donde se deduce:

$$\rightarrow E(n(s)) = \sum_{k \in U} \pi_k$$

$$\rightarrow Var(n(s)) = \sum_k \sum_l \pi_{kl} - \left[\sum_k \pi_k \right]^2 \qquad (1.2.8)$$

Un diseño muestral se dice de *tamaño fijo* si cada muestra con probabilidad de selección mayor que cero contiene un número fijo de n elementos de la población.

Para este tipo de diseños se deducen las siguientes propiedades a partir de las anteriores:

$$\rightarrow \sum_U \pi_k = n$$

$$\rightarrow \sum_{k \in U} \sum_{\substack{l \in U \\ k \neq l}} \pi_{kl} = n(n-1)$$

$$\rightarrow \sum_{l \in U} \pi_{kl} = n \pi_k \qquad (1.2.9)$$

Un estimador e=e(s,Y) para el parámetro θ(Y), es insesgado para un diseño muestral d, si la esperanza por diseño del estimador es θ(Y) uniformemente en Y; o sea:

$E(e(s,Y)) = \sum_{s \in S} e(s,Y) P(s) = \theta(Y)$ uniformemente en Y.

Se define el Error Cuadrático Medio (ECM) de un estimador como

$ECM = E(e(s,Y) - \theta(Y))^2$

y un estimador se dice el "mejor estimador" si es de mínima varianza uniforme insesgado, o sea si es insesgado y además el ECM de cualquier otro estimador es mayor. Sin embargo este tipo de estimadores, llamados *emvui* no siempre existen (Basu 1971), y en general no existe para el caso del Total dentro de la teoría del muestreo en poblaciones finitas.

Cuando no existe el emvui puede hallarse un estimador óptimo dentro de los estimadores *admisibles*, que son aquellos para los cuales no existe otro que lo mejore desde el punto de vista del ECM.

1.3 EL ESTIMADOR DE HORVITZ THOMPSON

El estimador de Horvitz Thompson para un total se define como:

$$\hat{T}_\pi = \sum_{k\in s} d_k\, y_k \qquad \text{donde } d_{k=}1/\pi_k \tag{1.3.1}$$

El estimador de Horvitz Thompson es insesgado, ya que :

$$E(\hat{T}_\pi) = E(\sum_{k\in S}\frac{y_k}{\pi_k}) = E(\sum_{k\in U}\frac{y_k}{\pi_k}I_k) = \sum_{k\in U}\frac{y_k}{\pi_k}E(I_k) = \sum_{k\in U}\frac{y_k}{\pi_k}\pi_k = \sum_{k\in U} y_k = T_y$$

La varianza del estimador de Horvitz-Thompson resulta:

$$V\left(\hat{T}_\pi\right) = \sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} \tag{1.3.2}$$

con $\Delta_{kl} = \pi_{kl} - \pi_k\pi_l$

En efecto:

$$\hat{T}_\pi = \sum_s \frac{y_k}{\pi_k} = \sum_U \frac{y_k}{\pi_k} I_k$$

$$V\left(\hat{T}_\pi\right) = \sum_U V(I_k)\left(\frac{y_k}{\pi_k}\right)^2 + \sum_k\sum_{l\neq k} Cov(I_k I_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

Por (1.2.5) y (1.2.6)

$$V\left(\hat{T}_\pi\right) = \sum_U \Delta_{kk}\left(\frac{y_k}{\pi_k}\right)^2 + \sum_k\sum_{l\neq k}\Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} = \sum\sum_U \Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}$$

El estimador definido como

$$\hat{V}\left(\hat{T}\pi\right) = \sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} \tag{1.3.3}$$

resulta insesgado, siempre que $\pi_{kl} > 0$ para todo k y l

Dem:

La misma expresión puede re escribirse en función de las indicadoras I_k

$$\hat{V}\left(\hat{T}\pi\right) = \sum\sum_U I_k I_l \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

y utilizando el hecho que

$$E\left(I_k I_l \frac{\Delta_{kl}}{\pi_{kl}} \right) = \pi_{kl} \frac{\Delta_{kl}}{\pi_{kl}} = \Delta_{kl}$$

resulta que

$$E\left(\hat{V}\left(\hat{T}_\pi\right)\right) = \sum\sum_U \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} = V\left(\hat{T}_\pi\right)$$

La varianza (1.3.2) puede expresarse como

$$V\left(\hat{T}\pi\right) = \sum\sum_U \left(\frac{\pi_{kl}}{\pi_k \pi_l} - 1 \right) y_k y_l = \sum\sum_U \left(\frac{\pi_{kl}}{\pi_k \pi_l} \right) y_k y_l - \left(\sum_U y_k\right)^2 \qquad (1.3.4)$$

Por lo tanto, el estimador (1.3.2) puede escribirse como

$$\hat{V}\left(\hat{T}\pi\right) = \sum_s \sum_s \frac{1}{\pi_{kl}} \left(\frac{\pi_{kl}}{\pi_k \pi_l} - 1 \right) y_k y_l \qquad (1.3.5)$$

En particular, si el diseño es de tamaño fijo la varianza del estimador de Horvitz-Thompson resulta de la siguiente forma:

$$V\left(\hat{T}_\pi\right) = -\frac{1}{2} \sum\sum_U \Delta_{kl} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \qquad (1.3.6)$$

Además, siempre que $\pi_{kl} > 0$ para todo k y l, se obtiene el siguiente estimador insesgado de la varianza, conocido como estimador de Yates, Grundy & Sen:

$$\hat{V}_{YGS}(\hat{T}_\pi) = -\frac{1}{2} \sum_{l \in s} \sum_{k \neq l} \breve{\Delta}_{kl} (\breve{y}_k - \breve{y}_l)^2 \qquad (1.3.7)$$

siendo $\breve{\Delta}_{kl} = \Delta_{kl} / \pi_{kl}$ e $\breve{y}_k = y_k / \pi_k$ (1.3.8)

Dem:

Utilizando la notación introducida en (1.3.8), la expresión (1.3.6) resulta:

$$V\left(\hat{T}_\pi\right) = -\frac{1}{2}\sum\sum_U \Delta_{kl}\left(\breve{y}_k - \breve{y}_l\right)^2 \qquad (1.3.9)$$

Desarrollando el cuadrado y la suma se obtiene:

$$V(\hat{T}_\pi) = -\frac{1}{2}\left(\sum\sum_U \Delta_{kl}\breve{y}_k^{\,2} + \sum\sum_U \Delta_{kl}\breve{y}_l^{\,2} - 2\sum\sum_U \Delta_{kl}\breve{y}_k\breve{y}_l\right) = \sum\sum_U \Delta_{kl}\breve{y}_k\breve{y}_l - \sum\sum_U \Delta_{kl}\breve{y}_k^{\,2} \qquad (1.3.10)$$

Considerando que:

$$\sum\sum_U \Delta_{kl}\breve{y}_k^{\,2} = \sum_{k\in U}\breve{y}_k^{\,2}\sum_{l\in U}\Delta_{kl} \quad ,$$

y utilizando (1.2.9) , dado que el diseño es de tamaño fijo,

$$\sum_{l\in U}\Delta_{kl} = \sum_{l\in U}\pi_{kl} - \sum_{l\in U}\pi_k\pi_l = n\pi_k - \pi_k n = 0$$

Con lo cual (1.3.10) resulta igual a (1.3.2).

El hecho de que (1.3.7) es insesgado es inmediato si se reescribe la expresión en términos de la función indicadora:

$$\hat{V}\left(\hat{T}_\pi\right) = -\frac{1}{2}\sum\sum_U I_k I_l \breve{\Delta}_{kl}\left(\breve{y}_k - \breve{y}_l\right)^2$$

Y se utiliza que, siendo $\pi_{kl} > 0$ para todo k,l y $E(I_k I_l \breve{\Delta}_{kl}) = \Delta_{kl}$

El caso de diseño de tamaño fijo más inmediato es aquel en el cual las probabilidades de primer orden son constantes pues el plan de muestreo que se establece asigna igual probabilidad de selección a cada muestra, una vez fijo el tamaño, como es el caso del Muestreo Simple al Azar(MSA).

1.4 MUESTREO SIMPLE AL AZAR

Un muestreo simple al azar es aquél en el cual todas las muestras del mismo tamaño pertenecientes a S tienen igual probabilidad de ser seleccionadas. Por otra parte las probabilidades de primer orden son constantes, es decir

$\pi_k=c \quad k=1,...,N.$

Pueden distinguirse dos tipos de muestreo simple al azar: Sin reemplazo y Con reemplazo.

Muestreo Simple al Azar sin Reemplazo: en este caso un elemento que fue seleccionado no puede serlo nuevamente. Luego, se verifica que:

$$\rightarrow \quad p(s)=\begin{cases}\binom{N}{n}^{-1} & \text{si } n \text{ es el tamaño de muestra a seleccionar}\\ 0 & \text{en caso contrario}\end{cases} \tag{1.4.1}$$

$$\rightarrow \quad \pi_k=\sum_{s\ni k}\binom{N}{n}^{-1}=\binom{N-1}{n-1}\binom{N}{n}^{-1}=\frac{n}{N} \quad k\in U \tag{1.4.2}$$

$$\rightarrow \quad \pi_{kl}=\sum_{s\ni k,l}p(s)=\sum_{s\ni k,l}\binom{N}{n}^{-1}=\binom{N-2}{n-2}\binom{N}{n}^{-1}=\frac{n(n-1)}{N(N-1)} \quad k\neq l\in U \tag{1.4.3}$$

$$\rightarrow \quad \Delta_{kl}=\begin{cases}\pi_{kl}-\pi_k\pi_l=-\dfrac{n(N-n)}{N^2(N-1)} & \text{si } k\neq l\\ \pi_k(1-\pi_k)=\dfrac{n(N-n)}{N^2} & \text{si } k=l\end{cases} \tag{1.4.4}$$

Luego, para el estimador de un total se tiene:

$$\hat{T}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} \frac{y_k}{\frac{n}{N}} = \frac{N}{n} \sum_{k \in s} y_k = N \bar{y}_s \qquad (1.4.5)$$

$$Var(\hat{T}_\pi) = N^2 \frac{N-n}{N} \frac{S^2{}_y}{n} = \frac{N}{n}(N-n) S^2{}_y \qquad (1.4.6)$$

donde S^2_y es la varianza muestral.

Al valor de las probabilidades de inclusión de primer orden f=n/N se las denomina *fracción de muestreo.* Luego, a partir de (1.4.5):

$$\hat{T}_\pi = \frac{1}{f} \sum_{k \in S} y_k \qquad (1.4.7)$$

Una muestra simple al azar tiene la ventaja de que es simple de seleccionar, y es conveniente cuando la población es bastante homogénea. En los casos en que la población es muy grande o es homogénea de a grupos, o bien hay disponible información auxiliar para toda la población, existen otras alternativas mejores desde el punto de vista del error muestral.

1.5 MUESTREO ESTRATIFICADO

En el muestreo estratificado la población U se divide en H subpoblaciones llamadas *estratos* que no se solapan y que en su conjunto conforman la población completa; es decir, si a dichos estratos los llamamos U_h, h=1,...,H, entonces:

$$\bigcup_{h=1}^{H} U_h = U \quad \text{y} \quad U_h \bigcap U_k = \varnothing, \ \forall h \neq k$$

Además, si N_h es el tamaño del estrato U_h se tiene que $\sum_{h=1}^{H} N_h = N$

Se dice que un muestreo es estratificado si se selecciona una muestra probabilística S_h en cada U_h con un diseño de probabilidades p_h y la selección en cada estrato se realiza de manera independiente de las selecciones en los otros estratos. Luego, en este tipo de muestreo:

- Se debe especificar un diseño p_h y un tamaño de muestra por estrato n_h.

- Se debe especificar un estimador en cada estrato. Generalmente la elección es la misma para todos los estratos.
- La muestra aleatoria total es $S= \bigcup_{h=1}^{H} S_h$
- $p(s)= \prod_{h=1}^{H} p_h(s_h)$; donde $p_h(s_h)=P(S_h= s_h)$
- Se suponen conocidos los N_h.
- El tamaño total de la muestra es $n= \sum_{h=1}^{H} n_h$

Estimadores bajo un muestreo estratificado

La expresión del estimador de Horvitz-Thompson para un total bajo este tipo de diseño resulta ser:

$$\hat{T}_{\pi st} = \sum_{h=1}^{H} \hat{t}_{h\pi} \tag{1.5.1}$$

donde $\hat{t}_{h\pi}$ es el estimador de Horvitz-Thompson del total en el estrato h; es decir que en su cálculo se utilizan las probabilidades de primer orden correspondientes al diseño en el estrato.

Por otro lado, dado que los diseños se aplican en forma independiente en cada estrato, la varianza del estimador se calcula como:

$$Var_{st}(\hat{T}_{\pi st}) = \sum_{h=1}^{H} \text{var}(\hat{t}_{h\pi}) \tag{1.5.2}$$

y un estimador insesgado de la misma resulta:

$$\hat{V}_{st}(\hat{T}_{\pi st}) = \sum_{h=1}^{H} \hat{V}(\hat{t}_{h\pi}) \tag{1.5.3}$$

Estimadores para el caso de un muestreo simple sin reemplazo en cada estrato

Para el caso en que dentro de cada estrato se opta por un muestreo simple al azar sin reemplazo, las probabilidades de selección de primer orden son $\pi_k = \frac{n_h}{N_h}$, con k ∈ U_h, siendo U_h el h-ésimo estrato; mientras que las probabilidades de segundo orden y las covarianzas resultan en las siguientes fórmulas:

$$\pi_{kl} = \begin{cases} \dfrac{n_h(n_h-1)}{N_h(N_h-1)} & si\ k\ y\ l \in U_h \\ \dfrac{n_h n_j}{N_h N_j} & si\ k \in U_h; l \in U_j \end{cases} \qquad (1.5.4)$$

$$\Delta_{kl} = \begin{cases} \dfrac{n_h(N_h-n_h)}{N_h^{\,2}} & si\ k = l\ ; k \in U_h \\ -\dfrac{n_h(N_h-n_h)}{N_h^{\,2}(N_h-1)} & si\ k\,y\,l \in U_h, k \neq l \\ 0 & si\ k \in U_h\ y\ l \in U_j, h \neq j \end{cases} \qquad (1.5.5)$$

Utilizando (1.5.1):

$$\hat{T}_{\pi st} = \sum_{h=1}^{H} \hat{t}_{h\pi} = \sum_{h=1}^{H} N_h \bar{y}_h$$

$$\text{donde } \bar{y}_h = \sum_{k \in S_h} Y_k / n_h \qquad (1.5.6)$$

Como las selecciones son independientes entre los estratos y utilizando (1.5.2):

$$Var(\hat{t}_{\pi st}) = \sum_{h=1}^{H} Var(\hat{t}_{h\pi}) = \sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h} s^2_{YU_h} \qquad (1.5.7)$$

La varianza de este estimador puede estimarse en forma insesgada por:

$$\hat{V}(\hat{t}_{\pi st}) = \sum_{h=1}^{H} Var(\hat{t}_{h\pi}) = \sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h} s^2_{YU_h} \qquad (1.5.8)$$

$$donde\ s^2_{YU_h} = \frac{1}{n_h - 1} \sum_{k \in S_h} (y_k - \bar{y}_h)^2$$

1.6 *MUESTREO POR CONGLOMERADOS*

Una situación práctica que se da en muchas circunstancias es que no existe un marco de muestreo que identifique a todos y cada uno de los elementos de la población, pero sin embargo, estos elementos se encuentran agrupados en sub-poblaciones que son llamados *conglomerados*. Luego, puede seleccionarse una muestra a partir de estos agrupamientos. Distinguiremos dos situaciones: el muestreo de conglomerados en una etapa y en dos etapas.

La población se considera particionada en M subconjuntos C_i, i=1,...,M llamados conglomerados. Los tamaños de cada uno de dichos subconjuntos se notan como N_i y por lo tanto:

$$U = \bigcup_{i=1}^{M} C_i \quad \text{y} \quad \sum_{i=1}^{M} N_i = N \tag{1.6.1}$$

$C_i \cap C_j = \phi$

Luego, el total T_y puede escribirse en función de los conglomerados de la siguiente forma:

$$T_y = \sum_{k \varepsilon U} y_k = \sum_{i=1}^{M} \sum_{k \varepsilon C_i} y_k = \sum_{i=1}^{M} t_i$$

donde t_i es el total en el conglomerado i

(1.6.2)

a. Muestreo por Conglomerados en una etapa

El muestreo por conglomerados *en una etapa* se lleva a cabo seleccionando una muestra de conglomerados s_I de tamaño n_I o n_{SL} según el diseño sea de tamaño fijo o no, respectivamente, e *incluyendo luego en la muestra a todas las unidades de los conglomerados seleccionados*.

Por lo tanto, la muestra viene dada por $s = \bigcup_{i \varepsilon s_I} C_i$ y el tamaño de la muestra $n = \sum_{i \varepsilon s_I} n_i$. Consecuentemente en este tipo de muestreo, el número de elementos en la muestra es generalmente aleatorio dado que depende del tamaño de los

conglomerados que son seleccionados, aún cuando sea fijo el tamaño de la muestra de conglomerados.

En un diseño por conglomerados en una etapa la probabilidad de seleccionar el i-ésimo conglomerado es

$$\pi_{Ii} = \sum_{s_I \ni i} p_I(s_I) \text{ , } i = 1,...,M \qquad (1.6.3)$$

donde el subíndice I se utiliza para notar que corresponde a la selección de un conglomerado.

La probabilidad de seleccionar dos conglomerados distintos i y j es:

$$\pi_{Iij} = \sum_{s_I \ni i,j} p_I(s_I) \text{ , } i = 1,...,M \text{ , } j = 1,...,M \text{ con } i \neq j \qquad (1.6.4)$$

Luego, si k es una unidad del conglomerado i se tiene que:

$$\pi_k = P(k \varepsilon s) = P(i \varepsilon s_I) = \pi_{Ii} \qquad (1.6.5)$$

es decir que la probabilidad de inclusión de un elemento k es la probabilidad de inclusión del conglomerado al que pertenece el elemento k.

Para calcular las probabilidades de segundo orden asociadas a dos unidades k y l pertenecientes a la población U, se deben distinguir dos situaciones:

- Si k y l están en el mismo conglomerado i , entonces:

  Π_{kl}=probabilidad de seleccionar el conglomerado i =Π_{Ii}

- Si k y l pertenecen a distintos conglomerados i y j, entonces:

  Π_{kl}=probabilidad de seleccionar los conglomerados i y j =Π_{Iij} con i≠j

El estimador insesgado para el total de la población, bajo este diseño resulta ser:

$$\hat{T}_\pi = \sum_{s_I} \breve{t}_i = \sum_{s_I} \frac{t_i}{\pi_{Ii}} \qquad (1.6.6)$$

pues utilizando (1.6.5):

$$\hat{T}_\pi = \sum_s \breve{y}_k = \sum_{s_I} \sum_{C_i} \frac{y_k}{\pi_k} = \sum_{s_I} \left(\sum_{C_i} y_k \right) / \pi_{Ii} = \sum_{s_I} \frac{t_i}{\pi_{Ii}} = \sum_{s_I} \breve{t}_i$$

La varianza de este estimador viene dada por:

$$V(\hat{T}_\pi) = \sum_{U_I} t_i^2 \left(\frac{1}{\pi_{Ii}} - 1\right) + \sum\sum_{U_I} t_i t_j \left(\frac{\pi_{Iij}}{\pi_{Ii}\pi_{Ij}} - 1\right) \qquad (1.6.7)$$

Que puede ser estimada en forma insesgada, siempre que $\pi_{Iij}>0$ para $i \neq j$ como

$$\hat{V}(\hat{T}_\pi) = \sum_{S_I} \frac{t_i^2}{\pi_{Ii}} \left(\frac{1}{\pi_{Ii}} - 1\right) + \sum\sum_{S_I} \frac{t_i}{\pi_{Ii}} \frac{t_j}{\pi_{Ij}} \left(\frac{\pi_{Iij}}{\pi_{Ii}\pi_{Ij}} - 1\right) \qquad (1.6.8)$$

Si el diseño sobre los conglomerados es de tamaño fijo, se obtiene el estimador de Sen-Grundy& Yates:

$$\hat{V}\left(\hat{T}_\pi\right) = -\frac{1}{2} \sum\sum_{S_I} \breve{\Delta}_{Iij} \left(\breve{t}_i - \breve{t}_j\right)^2 \qquad (1.6.9)$$

A partir de la fórmula anterior se observa que:

- Si los valores de t_i / π_{Ii} son iguales $\forall i$ entonces $\text{var}(\hat{T}_\pi) = 0$ y por lo tanto si se eligen las π_{Ii} proporcionales al tamaño del total t_i en el conglomerado, el muestreo por conglomerados es altamente eficiente pues la varianza será pequeña porque serán pequeñas las diferencias.
- Si los Ni son conocidos, puede elegirse un diseño de manera que los Π_{Ii} resulten proporcionales a los Ni. Como los $t_i = N_i \bar{y}_i = \sum_{C_i} y_k$; el método es bueno si hay poca variación entre las medias de los conglomerados.
- El muestreo por conglomerados con probabilidades iguales es una mala elección si el tamaño de los conglomerados es distinto ya que en ese caso $N_i \bar{y}_i$ variará mucho y la varianza será grande.

Luego, el diseño por conglomerados en una etapa pierde precisión por el *"efecto del conglomerado"* que se refiere a cuán parecidas son las unidades dentro de cada C_i. Sería deseable que en cada conglomerado pudiera reproducirse la variabilidad del universo. A continuación, se presenta un coeficiente que cuantifica el hecho de que esto suceda o no; es decir el *grado de homogeneidad* de los conglomerados.

Coeficiente de homogeneidad y efecto del diseño

Para medir la precisión de un estimador de un total en un diseño muestral de conglomerados en una etapa, es útil trabajar con el *coeficiente de homogeneidad*(δ), definido como:

$$\boxed{\delta = 1 - \frac{S^2_{YW}}{S^2_Y} = 1 - \frac{S^2_D}{S^2_Y}}$$

(1.6.10)

siendo $S^2_{yw} = S^2_D = \frac{1}{N-M}\sum_{U_I}\sum_{Ci}(y_k - \bar{y}_{Ci})^2$ (pooled variance dentro de los clusters) (1.6.11)

$\bar{y}_{Ci}$ es la media dentro del i-ésimo cluster.

Por otro lado, (1.6.11) puede re-escribirse como:

$$S^2_{yw} = \frac{1}{\sum_{U_I}(N_i - 1)}\sum_{U_I}(N_i - 1)S^2_{YC_i} \qquad (1.6.12)$$

siendo S^2_{YCi} la varianza dentro de cada conglomerado.

Luego, δ es positivo o negativo de acuerdo a que la varianza media dentro de los conglomerados (S^2_{YW}) sea menor o mayor que la varianza global S^2_{YU} respectivamente.

- Si δ es pequeño entonces existe diferencia entre los elementos de un mismo conglomerados; es decir que es bajo el grado de homogeneidad entre los clusters.
- Si δ es grande entonces existe similitud entre los elementos de un mismo conglomerados; es decir que el grado de homogeneidad es alto.

La varianza también puede expresarse en términos de la covarianza de las observaciones en el conglomerado:

Si llamamos:

$\overline{N} = \frac{N}{M}$ número promedio de elementos por conglomerados

$K_I = M^2 \frac{(1 - f_I)}{m}$ donde $f_I = m/M$

$$\text{Cov} = Cov(N_i, N_i \overline{Y}_{C_i}{}^2) = \frac{1}{M-1} \sum_{U_I} (N_i - \overline{N}) N_i \overline{Y}_{C_i}{}^2$$

Entonces,

$S^2_{tU} = \overline{N} S^2_{YU} (1 + \frac{N-M}{M-1} \delta) + Cov$, con lo cual para el caso de MSA de clusters resulta la siguiente expresión de la varianza:

$$\boxed{\text{Var}_{MSAC}(\hat{t}) = (1 + \frac{N-M}{M-1} \delta) \overline{N} K_I S^2_{tU} + K_I \, \text{Cov}}$$

(1.6.13)

A partir de ello se obtiene el siguiente efecto de diseño:

$$\boxed{\text{Deff}(MSA, \hat{t}) = \frac{V_{MSAC}}{V_{MSA}} = 1 + \frac{N-M}{M-1} \delta + \frac{\text{Cov}}{\overline{N} S^2_{YU}}}$$

(1.6.14)

Observamos que si los tamaños de los clusters son todos iguales, entonces $N_i = \overline{N}$ y por lo tanto la Cov será nula. Luego, $\text{Deff} \approx 1 + (\overline{N} - 1)\delta$.

Por otra parte si Cov es positiva, el incremento de la varianza debido a la selección de los clusters puede ser peor que en el caso anterior ya que el segundo término puede ser grande; $\text{Deff} = \overline{N} (CV_N / CV_Y)^2$. siendo

$CV_N = \frac{S_{NU_I}}{\overline{N}}$ y $CV_Y = \frac{S_{Y_U}}{\overline{Y}_U}$.

En general, seleccionar clusters por MSA es ineficiente, sobre todo si los clusters son homogéneos y /o de tamaño distinto. La eficiencia puede ser mejorada cuando se dispone de una medida aproximada de cada cluster. Si esa medida es proporcional al total por cluster se puede reducir la varianza del π estimador

utilizando muestreo proporcional al tamaño para la selección de los clusters. Una alternativa es utilizar muestreo de clusters estratificado con los estratos tales que la variación del tamaño de los estratos sea pequeña.

b. Muestreo por Conglomerados en dos etapas

En este caso se seleccionan m conglomerados y luego se seleccionan n_i sub unidades dentro de ellos.

Sea U ={1,...,k,...,N} la población de estudio compuesta por M conglomerados C_i i=1,...,M que llamaremos *Unidades Primarias (PSU)*. A su vez, cada una de dichas M unidades primarias están compuestas por N_i unidades que llamamos *Unidades Secundarias (SSU), de manera tal que* $\sum_{i=1}^{M} N_i = N$.

El muestreo se lleva a cabo en dos pasos:

- una muestra S_I de unidades primarias es seleccionada bajo un diseño $p_I(s_I)$. El tamaño de la muestra seleccionada es m.
- Para cada i $\in$ S_I se extrae una muestra s_i de tamaño n_i según un diseño $p_i(*/s_I)$.

Este tipo de muestreo es ventajoso cuando dentro de un mismo conglomerado es esperable observar cierta similitud entre los elementos, o sea cierta homogeneidad y por lo tanto podría resultar poco económico medir a todas las unidades dentro del conglomerado. El muestreo en dos etapas se plantea entonces como una posible solución para evitar el efecto de la correlación. En este caso es necesario estimar los totales t_i dentro de cada conglomerado utilizando la sub muestra. Si la variación dentro de los conglomerados es pequeña, los estimadores de dichos totales tendrán un varianza chica aún para una sub muestra de tamaño moderado, lo cual hace que este tipo de muestreo sea utilizado con bastante frecuencia.

En un muestreo de dos etapas hay dos hipótesis inherentes: Invarianza e Independencia:

- Invarianza: el diseño de selección de las unidades secundarias es el mismo para cada conglomerado; es decir que los planes $p_I(*/s_I)$ de la segunda etapa no dependen de lo que pasó en la primera etapa en el sentido en que se emplearán los diseños en la segunda etapa cualquiera sea la muestra de la primera.

  Como consecuencia: $P(S_I=s_I)= P(S_I=s_I/S_I)$

- Independencia: las selecciones realizadas en la segunda etapa dentro de cada conglomerado son independientes entre sí. Por lo tanto:

$$P(\bigcup_{i\in S_I} s_i / S_I) = \prod_{i\in S_I} P(s_i / S_I).$$

La muestra aleatoria viene dada por $s= \bigcup_{i\in S_I} s_i$ y el tamaño de s es $n=\sum n_i$, el cual en general es aleatorio.

Para obtener los estimadores se deben definir las probabilidades de primer y segundo orden de acuerdo a un plan bietápico:

- π_{Ii} : probablidad de primer orden en Etapa 1; es decir la probabilidad de seleccionar la unidad primaria C_I.
- π_{Iij} : probabilidad de inclusión de segundo orden para las unidades primarias U_i y U_j. A partir de ellas se tiene:

$$\Delta_{Iij} = \begin{cases} \pi_{Iij} - \pi_{Ii}\,\pi_{Ij} & si\ i \neq j \\ \pi_{Ii}(1 - \pi_{Ii}) & si\ i = j \end{cases} \qquad (1.6.15)$$

- $\pi_{k/i}$: probabilidad de seleccionar la unidad k dado que i fue seleccionada en la primera etapa.
- $\pi_{kl/i}$: probabilidad de seleccionar las unidades k y l dado que i fue seleccionada en la primera etapa. A partir de ellas se tiene:

$$\Delta_{kl/i} = \begin{cases} \pi_{kl/i} - \pi_{k/i}\,\pi_{l/i} & si\ k \neq l \\ \pi_{k/i}(1-\pi_{k/i}) & si\ k = l \end{cases}, \text{ i=1,...,m} \qquad (1.6.16)$$

Luego, la probabilidad de que la unidad k sea seleccionada para la muestra es:

$$\pi_k = \pi_{Ii}\,\pi_{k/i} \quad \text{si } k \in C_i \qquad (1.6.17)$$

Las probabilidades de inclusión de segundo orden dependen de que las unidades pertenezcan o no al mismo conglomerado:

$$\pi_{kl} = \begin{cases} \pi_{Ii}\,\pi_{k/i} & si\ k \wedge l \in C_i\ ;\ k = l \\ \pi_{Ii}\,\pi_{kl/i} & si\ k \wedge l \in C_i\ ;\ k \neq l \\ \pi_{Iij}\,\pi_{k/i}\,\pi_{l/j} & si\ k \in C_i \wedge l \in C_j\ i \neq j \end{cases} \qquad (1.6.18)$$

Luego, el estimador de Horvitz-Thompson para un total resulta ser:

$$\hat{T}_\pi = \sum_{i \in S_I} \sum_{k \in s_i} \frac{y_k}{\pi_{Ii}\,\pi_{k/i}} = \sum_{i \in S_I} \frac{\hat{t}_i}{\pi_{Ii}} \qquad (1.6.19)$$

siendo $\hat{t}_i = \sum_{k \in s_i} \frac{y_k}{\pi_{k/i}}$ ya que se debe estimar el total dentro de cada conglomerado pues no se utilizan todos los elementos sino una muestra. Se estima t_i con su correspondiente estimador de Horvitz-Thompson.

En cuanto a la variabilidad, existirán dos fuentes de variación: la debida a la selección de los conglomerados y la debida al sub-muestreo dentro de los conglomerados. Por lo tanto, en este tipo de plan la varianza puede descomponerse en dos sumandos: la componente de la varianza debida a la primera etapa (V_{PSU}) y la debida a la segunda etapa (V_{SSU})

$Var_2(\hat{T}_\pi) = V_{PSU} + V_{SSU}$

con $V_{PSU} = \sum \sum_{U_I} \Delta_{Iij} \breve{t}_i \breve{t}_j$

y $V_{SSU}=\sum_{i=1}^{M}\frac{V(\hat{t}_i)}{\pi_{Ii}}$ siendo $V(\hat{t}_i)=\sum_{k\in C_i}\sum_{l\in C_i}\Delta_{kl/i}\frac{y_k}{\pi_{k/i}}\frac{y_l}{\pi_{l/i}}$ i=1,...,M

(1.6.20)

Demostración:

$Var(\hat{T}_\pi)=Var(E(\hat{T}_\pi / S_I)) + E(Var(\hat{T}_\pi / S_I))$

Para el primer término vale que:

$Var(E(\hat{T}_\pi / S_I))=Var(E(\sum_{i\in S_I}\hat{t}_i / S_I))$

Por la propiedad de invarianza

$E(\sum_{i\in S_I}\hat{t}_i / S_I)=\sum_{i\in S_I}E(\hat{t}_i / S_I)=\sum_{i\in S_I}E(\hat{t}_i)=\sum_{i\in S_I}\frac{t_i}{\pi_{Ii}}$ $\quad si\ t_i=\sum_{k\in S_i}y_k$

Luego:

$Var(E(\hat{T}_\pi / S_I))=Var(\sum_{i\in S_I}\frac{t_i}{\pi_{Ii}})=\sum_{i=1}^{M}\sum_{j=1}^{M}\frac{t_i}{\pi_{Ii}}\frac{t_j}{\pi_{Ij}}\Delta_{Iij}$

Para el segundo término se puede escribir:

$E(Var(\hat{T}_\pi / S_I))=E(Var(\sum_{i\in S_I}\frac{\hat{t}_i}{\pi_{iI}} / S_I)$

Por las propiedades de invarianza e independencia resulta

$Var(\sum_{i\in S_I}\frac{\hat{t}_i}{\pi_{Ii}} / S_I)=\sum_{i\in S_I}Var(\frac{\hat{t}_i}{\pi_{Ii}} / S_I)=\sum_{i\in S_I}\frac{Var(\hat{t}_i)}{\pi^2{}_{Ii}}$

Luego

$E(Var(\hat{T}_\pi / S_I))=E(\sum_{i\in S_I}\frac{Var(\hat{t}_i)}{\pi^2{}_{Ii}})=\sum_{i=1}^{M}\frac{Var(\hat{t}_i)}{\pi_{Ii}}$

donde $Var(\hat{t}_i)$ está dado por la fórmula (1.6.20)

Los correspondientes estimadores resultan:

$\hat{V}_{PSU} = \sum_{i \in S_I} \sum_{j \in S_j} \frac{\hat{t}_i}{\pi_{Ii}} \frac{\hat{t}_j}{\pi_{Ij}} \frac{\Delta_{Iij}}{\pi_{Iij}} - \sum_{i \in S_I} \frac{1}{\pi_{Ii}} \left(\frac{1}{\pi_{Ii}} - 1\right) \hat{V}(\hat{t}_i)$ (el segundo término se debe restar para que el estimador resulte insesgado).

$$\hat{V}_{SSU} = \sum_{i \in S_I} \frac{\hat{V}(\hat{t}_i)}{\pi^2_{Ii}}$$

Luego, al sumar ambas expresiones se obtiene:

$$\hat{V}_{2etapas}(\hat{T}_\pi) = \sum_{i \in S_I} \sum_{j \in S_j} \frac{\hat{t}_i}{\pi_{Ii}} \frac{\hat{t}_j}{\pi_{Ij}} \frac{\Delta_{Iij}}{\pi_{Iij}} + \sum_{i \in S_I} \frac{\hat{V}(\hat{t}_i)}{\pi_{Ii}} \quad ; \quad \text{con } \pi_{Iii} = \pi_{Ii} \qquad (1.6.21)$$

$$\text{siendo } \hat{V}(\hat{t}_i) = \sum_{k \in S_I} \sum_{l \in S_i} \frac{y_k y_l}{\pi_{k/i} \pi_{l/i}} \frac{\Delta_{kl/i}}{\pi_{kl/i}} \quad ; \quad \text{con } \pi_{kk/i} = \pi_{k/i} \qquad (1.6.22)$$

Sin embargo, (1.6.15) puede ser bastante complicado de aplicar, sobre todo porque las $\hat{V}(\hat{t}_i)$ deben calcularse para cada i-ésima unidad seleccionada en la muestra de primera etapa. Luego, sería interesante contar con un estimador alternativo que sea más sencillo desde el punto de vista práctico.

Sârndall demostró que si se considera como estimador sólo el primer término de (1.6.16), éste resulta un estimador con cierto sesgo:

Si $\hat{V}^* = \sum_{i \in S_I} \sum_{j \in S_j} \frac{\hat{t}_i}{\pi_{Ii}} \frac{\hat{t}_j}{\pi_{Ij}} \frac{\Delta_{Iij}}{\pi_{Iij}}$ entonces: $E(\hat{V}^*) = V_{2etapas}(\hat{T}_\pi) - \sum_{U_I} V_i$ y por lo tanto el sesgo de $\hat{V}^*$ resulta: $B(\hat{V}^*) = -\sum_{U_I} V_i$, lo cual indica que $\hat{V}^*$ subestima la varianza de $\hat{T}_\pi$.

Al analizar el sesgo relativo:

$$\frac{B(\hat{V}^*)}{V_{2etapas}(\hat{T}_\pi)} = -\frac{\sum_{U_I} V_i}{\sum\sum_{U_I} \Delta_{ij} \check{t}\check{t}_j + \sum_{U_I} \frac{V_i}{\pi_{Ii}}}$$, se observa que la subestimación no produce, en general, un sesgo importante si las π_{Ii} son pequeñas.

En los capítulos siguientes consideraremos la estimación de la varianza a través de $\hat{V}^*$.

CAPÍTULO II

MUESTREO PROPORCIONAL AL TAMAÑO

2.1 MUESTREO CON PROBABILIDADES DESIGUALES

En las aplicaciones es muy frecuente que la población pueda dividirse en conglomerados que, por lo general, no tienen el mismo tamaño. Por ejemplo, si se desea estimar el total de desocupados urbanos en una provincia y se consideran las ciudades como conglomerados, es claro que no todas tendrán la misma cantidad de viviendas ni de individuos. En un muestreo simple al azar todas las ciudades tendrían igual probabilidad de participar en la muestra ya que el tamaño de la misma no es tomado en cuenta. Sin embargo, desde el punto de vista del muestreo sería más representativo dar mayor probabilidad a las más grandes ya que el total a estimar, particularmente el total de la población, estará fuertemente influenciado por las ciudades grandes.

Si recordamos la fórmula de varianza de Yates Grundy & Sen para el estimador de Horvitz Thompson del total de una población para el caso de un diseño de tamaño fijo, observamos que dicha varianza puede ser nula si las π_i son proporcionales a las Y_i puesto que se anularían los términos elevados al cuadrado:

$$Var_{YGS}(\hat{T}_\pi) = \sum_{k \in U} \sum_{k<l} (\pi_k \pi_l - \pi_{kl})[\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}]^2 \qquad (2.1.1)$$

En la práctica ello es imposible ya que implicaría conocer todas las y_i. Sin embargo, si existiera una variable auxiliar X conocida, de valores x_k positivos y tal que los x_k fueran proporcionales a y_k, bastaría con elegir un diseño en el que *las probabilidades de primer orden sean proporcionales a X*; o sea,

$$\pi_k = \frac{n x_k}{\sum_{j \in U} x_j} \qquad (2.1.2)$$

pues en ese caso $\frac{y_k}{\pi_k} = \frac{y_k \sum_{j \in U} x_j}{n x_k}$ (2.1.3)

y por lo tanto al ser y_k proporcional a x_k, los términos en la sumatoria (2.1.1) serían 0.

Por otra parte, para un diseño de este tipo el estimador del total será insesgado.

En efecto:

$$\hat{T}_{HT} = \sum_{k\in S}\frac{y_k}{\pi_k} = \left(\sum_{k\in U}x_k\right)\left(\sum_{k\in S}\frac{y_k}{nx_k}\right) = \left(\sum_{k\in U}(1/\alpha)y_k\right)\left(\sum_{k\in S}\frac{\alpha x_k}{nx_k}\right) = \frac{1}{\alpha}T\alpha\sum_{k\in S}\frac{1}{n} = T$$

Este tipo de muestreo se conoce con el nombre de *Muestreo Proporcional al Tamaño o PPT.*

2.2 MUESTREO PROPORCIONAL AL TAMAÑO

Existen diferentes métodos para implementar un muestreo proporcional al tamaño.

En general para que un método resulte adecuado sería deseable que se cumplan las siguientes propiedades:

a. $\pi_i = np_i > 0 \quad \forall i \in U$; donde $p_i = \dfrac{x_i}{\sum_k x_k}$

b. $\pi_{ij} > 0 \quad \forall i, j \in U$

c. $\pi_i \pi_j > \pi_{ij} \quad \forall i, j \in U$

d. que el cálculo de las π_{ij} sea sencillo y no necesite examinar la probabilidad de selección de todas las muestras posibles.

e. que la probabilidad de seleccionar una muestra no dependa del orden de las unidades.

f. que las π_{ij} verifique la condición de Yates-Grundy: $\pi_{kl} \angle \pi_k \pi_l$ para todo $k \neq l$. Esta condición se deduce de (1.3.5) como necesaria para que el estimador de la varianza de SG&Y resulte positivo.

Sin embargo no existe un método que cumpla con las seis propiedades al mismo tiempo.

El método tradicional para un muestreo proporcional al tamaño, corresponde a una selección con reemplazo y es el *Método por Acumulación* en el cual se calcula la distribución acumulada para X, se genera un número aleatorio entre 0 y el total de X y se selecciona el primer elemento de la muestra como el primero que acumula un total mayor a ese valor. Para seleccionar una muestra de n unidades con reemplazo se repite el proceso n veces.

Sin embargo, cuando el tamaño de los valores observados de la variable X es *grande*, la acumulación de los totales puede ser muy tediosa y además el método genera un muestreo con reposición, que no siempre es de interés en la práctica. Existen otras alternativas orientadas más hacia la extracción de muestras *sin reemplazo*. Surgen entonces métodos tales como:

- Madow: es poco eficiente desde el punto de vista de la varianza.
- Sistemático Acumulado: es bastante usado para tamaños de muestra mayores a 2, pero un problema es que por la naturaleza del método puede ocurrir que alguna probabilidad de segundo orden se anule, con lo cual el diseño no generaría un estimador de la varianza insesgado.
- Lahiri-Midzuno: es un mecanismo de selección que tiene la bondad de que las probabilidades de segundo orden sean positivas y por lo tanto genera un estimador de varianza insesgado. Sin embargo, es un método de tipo condicional que requiere que para toda unidad x_k el producto $n\frac{x_k}{\sum_{k\in U} x_k}$ sea mayor a $\frac{n-1}{N-1}$, y en la práctica no es muy frecuente que ésto se cumpla
- Brewer: funciona bien si el tamaño de la muestra es 2.
- Sunter: tiene la desventaja que el algoritmo no genera probabilidades proporcionales al tamaño como se pretende.

Un método que verifica la mayoría de las propiedades enumeradas anteriormente y en particular la referida a la positividad del estimador de varianza de Yates-

Grundy-Sen, es el método de Sampford, que será descripto en detalle en la siguiente sección.

2.3 MÉTODO DE SAMPFORD (1967)

Este método tiene la ventaja de ser de sencilla implementación y que las probabilidades de segundo orden pueden calcularse de una forma relativamente simple.

Definiendo $w_k = \frac{\pi_k}{1-\pi_k}$ kε U (2.3.1), el diseño de Sampford se define a través de la siguiente probabilidad:

$$P_{SAMP}(S) = C_n \sum_{l \in s} \pi_l \prod_{\substack{k \in s \\ k \neq l}} w_k \qquad (2.3.2)$$

con $C_n = \left(\sum_{t=1}^{n} t D_{n-t} \right)^{-1}$ y $D_m = \sum_{s \in S_m} \prod_{k \in s} w_k$ siendo S_m el soporte de las muestras de tamaño m. (2.3.3)

o bien de manera equivalente:

$$P_{SAMP}(S) = C_n \prod_{k \in s} w_k \left(n - \sum_{l \in s} \pi_l \right) \qquad (2.3.4)$$

Sampford demostró que el diseño verifica las propiedades esperadas para lo cual demostró previamente el siguiente lema, cuya demostración se presenta en el Apéndice (1).

Lema 2.3.1

Definiendo g(n,r,k)= $\sum_{s \in S_{n-r}(U \setminus k)} \left(\prod_{l \in s} w_l \right) \left(n - r\pi_k - \sum_{i \in s} \pi_i \right)$ (2.3.5)

donde

$S_{n-r}(U \setminus k) = \{ s \in S_{n-r} \,/\, I_k = 0 \}$ o sea, el conjunto de muestras de tamaño n-r que no incluyen a y_k.

Entonces:

$$g(n,r,k)=(1-\pi_k)\sum_{t=1}^{n} tD_{n-t} \qquad (2.3.6)$$

Se demuestran luego las siguientes propiedades (Sampford, 1967):

i. $\sum_{s\in S_n} p_{SAMP}(s) = 1$

ii. $\sum_{s\in S_n} I_k p_{SAMP}(s) = \pi_k$

iii. $\pi_{kl} = C_n w_k w_l \sum_{t=2}^{n}(t-\pi_k+\pi_l)D_{n-t}(\bar{k},\bar{l})$, siendo $D_z(\bar{k},\bar{l}) = \sum_{s\in S_z(U\setminus\{k,l\})}\left(\prod_{j\in s} w_j\right)$

Dem:

i. Sea z una unidad con probabilidad de inclusión nula, o sea que tanto π_k como w_k son 0; entonces según el Lema:

$$g(n,0,z)=(1-\pi_k)\sum_{t=1}^{n} tD_{n-t} = \sum_{t=1}^{n} tD_{n-t}$$

Por definición $g(n,0,z)=\sum_{s\in S_n(U\setminus k)}\left(\prod_{l\in s} w_l\right)\left(n-\sum_{i\in s}\pi_i\right)$

$$\sum_{s\in S_n} p_{SAMP}(s) = C_n \prod_{k\in s} w_k\left(n-\sum_{l\in s}\pi_l\right) = C_n \sum_{t=1}^{n} tD_{n-t} = 1$$

ii. $\sum_{s\in S_n} I_k p_{SAMP}(s) = C_n w_k\, g(n,1,k) = \pi_k C_n \sum_{t=1}^{n} tD_{n-t} = \pi_k$

iii. Se puede escribir $\pi_{kl} = C_n w_k \psi_{kl}$

siendo $\psi_{kl} = \sum_{s\in S_z(U\setminus\{k,l\})}\left(\prod_{j\in s} w_j\right)\left(n-\pi_k-\pi_l-\sum_{j\in s}\pi_j\right)$

(2.3.7)

Supongamos ahora que las unidades k y l se agregan en una nueva unidad α, o sea que el tamaño de la población (U') ahora es *N-1*, $\pi_\alpha = \pi_k + \pi_l$ y . Llamando g' y D' a las correspondientes funciones g y D definidas sobre la población U', entonces:

$$\psi_{kl} = \sum_{s \in S_z(U \setminus \{\alpha\})} \left(\prod_{i \in s} w_i \right) \left(n - \pi_\alpha - \sum_{j \in s} \pi_j \right) = g'(n,2,\alpha) + \pi_\alpha D'_{n-2}(\bar{\alpha})$$

Como $D'_{n-t} = D'_{n-t}(\bar{\alpha}) + w_\alpha D'_{n-t-1}(\bar{\alpha})$

Para t< n-1, con $D'_0(\bar{\alpha}) = 1$, utilizando el Lema, se obtiene que:

$$\psi_{kl} = (1 - \pi_\alpha) \left[\sum_{t=2}^{n} t D'_{n-t}(\bar{\alpha}) + \sum_{t=2}^{n} t w_\alpha D'_{n-t-1}(\bar{\alpha}) \right] + \pi_\alpha D'_{n-2}(\bar{\alpha}) = \sum_{t=2}^{n} (t - \pi_\alpha) D'_{n-t}(\bar{\alpha})$$

Volviendo a escribir la expresión relativa a la población original se obtiene:

$\psi_{kl} = \sum_{t=2}^{n} (t - \pi_k + \pi_l) D_{n-t}(\bar{k}, \bar{l})$ y reemplazando en (2.3.7) se obtiene la expresión en iii. para π_{kl}

2.4 ***IMPLEMENTACIÓN DEL MÉTODO***

Secuencialmente, el método de Sampford puede aplicarse siguiendo los siguientes pasos:

Siendo

→ X_k: tamaño unidad k (medido de acuerdo a la información auxiliar provista por la variable X)

→ $\sum_{i \in U} X_i$: tamaño total de la población

→ n: tamaño de la muestra a seleccionar,

Se define el tamaño relativo como $Z_k = \frac{X_k}{\sum_i X_i}$

Luego:

➔ Se selecciona la primera unidad con el método acumulativo total proporcional a los Z_k .

➔ Se calculan para cada unidad los siguientes tamaños ajustados:

$$Z^*_k = \frac{X_k}{\sum_i X_i - nX_k}, \text{ o bien en forma equivalente: } \quad Z^*_k = \frac{Z_k}{1 - nZ_k} \qquad (2.4.1)$$

- Se selecciona una muestra de tamaño n-1 con probabilidad proporcional a los tamaños ajustados, con reemplazo.

- Si la muestra contiene algún elemento repetido entonces se rechaza la muestra y se selecciona una nueva. La muestra se acepta sólo si contiene n unidades diferentes.

Por la metodología de construcción este método se incluye dentro de los métodos de "tipo rechazo", ya que va generando una familia de muestras hasta que encuentra la que cumple las condiciones buscadas.

Este método verifica la mayoría de las propiedades enumeradas pero tiene la desventaja de que pueden llegar a rechazarse muchas muestras, por eso los softwares generalmente lo aplican si la probabilidad de rechazo de una muestra es baja.

Las probabilidades de primer orden en este caso, resultan ser:

$$\pi_k = nZ_k \quad , \tag{2.4.2}$$

Y las probabilidades de segundo orden pueden escribirse según (2.3.7) utilizando la siguiente expresión equivalente:

$$\pi_{kl} = K\lambda_k\lambda_l \sum_{r=2}^{n} \frac{[r - n(\pi_k + \pi_l)L_{n-t}(k,l)]}{n^{r-2}} \tag{2.4.3}$$

siendo:

$$\lambda_k = \frac{Z_k}{1 - nZ_k}$$

$L_m = \sum_{s(m)} \lambda_{i1}\lambda_{i2}\ldots\lambda_{im}$ con s(m): todas las muestras posibles de tamaño m, para $m=1,2,\ldots,N.$

$L_m(k,l)$ se define en forma similar a L_m pero la suma se realiza sobre todas las posibles muestras de tamaño m que no incluyan a las unidades k y l.

$$K=\left(\sum_{r=1}^{n} \frac{r\,L_{n-r}}{n^r}\right)^{-1}$$

CAPITULO III

ESTIMADORES CON EL USO DE INFORMACIÓN AUXILIAR EXTERNA

El estimador de Horvitz Thompson si bien es insesgado, no es eficiente cuando se cuenta con información externa a la muestra a nivel poblacional. En esos casos existen otros estimadores alternativos ya que el uso de la información auxiliar permite, la mayoría de las veces, disminuir el ECM a costa de generar sesgo, que será pequeño bajo ciertos supuestos.

La existencia de información auxiliar motiva la definición de los *Estimadores de Regresión.*

3.1 *Estimadores de Regresión*

Siendo $U=\{1,...,k,...,N\}$ una población finita de la cual se extrae una muestra $s=(i_1,i_2,...i_{n(s)})$ donde $1\leq i_1\neq i_2\neq...\neq i_{n(s)} \leq N$ según un diseño $p(.)$ con probabilidades de primer orden de inclusión π_k y de segundo orden $\pi_{kl}>0$, se asume que, para cada unidad observada, se tiene información referida a una serie de variables relevadas para cada k sobre las cuales se tienen datos marginales a nivel *poblacional*; es decir, para cada $k\varepsilon s$ se observan los vectores $\mathbf{x}_k \in R^p$ cuyos totales poblacionales se conocen.

Luego, el problema que se plantea es estimar $T_y = \sum_U y_k$ habiendo observado $(y_k, \mathbf{x}_k)$, $k\varepsilon s$ y conociendo el vector $\mathbf{t_x}$ de totales de los $\mathbf{x}_k$ $\forall k\varepsilon U$ ($\boldsymbol{t_X} = \sum_U \boldsymbol{x_k}$). Los estimadores de regresión se basan en aplicar un modelo de regresión lineal para el vector **Y** basado en las variables auxiliares.

3.1.1 El estimador de Razón

Supongamos que la información auxiliar viene dada por $X\in R$ pero de manera tal que existe una relación de proporcionalidad directa aproximada entre esta variable y la variable de interés Y. Es decir que, para cada x_k, los cocientes y_k/x_k son aproximadamente constantes. Se supondrá además que la varianza de los y_k es proporcional a x_k, es decir

$$\begin{cases} E(y_k)=\beta x_k \\ V(y_k)=\sigma^2 x_k \end{cases} \qquad (3.1.1.1)$$

Una forma de aproximar β es utilizando la teoría de regresión de Y sobre X, con lo cual:

$$\beta = \frac{\sum_U \frac{x_k y_k}{\sigma^2 x_k}}{\sum_U \frac{x^2 k}{\sigma^2 x_k}} = \frac{\sum_U \frac{y_k}{\sigma^2}}{\sum_U \frac{x_k}{\sigma^2}} = \frac{\sum_U y_k}{\sum_U x_k} = \frac{T_y}{t_x} \qquad (3.1.1.2)$$

Podemos estimar β a partir de la muestra, utilizando el π-estimador de t_y y t_x:

$$\hat{T}_y = \sum_s \frac{y_k}{\pi_k} = \sum_s \breve{y}_k \quad ; \quad \hat{t}_x = \sum_s \frac{x_k}{\pi_k} = \sum_s \breve{x}_k \qquad (3.1.1.3)$$

y por lo tanto

$$\hat{\beta} = \frac{\sum_s \breve{y}_k}{\sum_s \breve{x}_k} \qquad (3.1.1.4)$$

$$\hat{T}_{yr} = \sum_U x_k \frac{\sum_s \breve{y}_k}{\sum_s \breve{x}_k} \qquad (3.1.1.5)$$

3.1.2 El Estimador de Regresión Generalizada

Asumamos que la información auxiliar viene dada por p variables conocidas para cada una de las unidades u_k. Se define a la matriz Nxp $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)'$ donde cada elemento $\boldsymbol{x}_k = (\boldsymbol{x}_{k1}, \boldsymbol{x}_{k2}, \ldots\ldots, \boldsymbol{x}_{kp})'$ es el vector px1 de valores de las p variables para el k-ésimo elemento de la muestra observada. Generalizando el proceso descripto en el párrafo anterior, se realiza la regresión de **Y** sobre **X** de manera de minimizar la suma de cuadrados de los residuos, es decir que el problema se plantea como:

hallar $\boldsymbol{\beta} / (\boldsymbol{Y}-\boldsymbol{X\beta})' (\boldsymbol{Y}-\boldsymbol{X\beta}) = \boldsymbol{E}'\boldsymbol{E} = \sum_{i \in U} E_i^2 = \sum_{i \in U} (y_i - \boldsymbol{x}_i' \boldsymbol{\beta})^2$ *sea mínima*

con $\mathbf{E}=(E_1, E_2, \ldots, E_N)$ vector de residuos del modelo.

Luego, $\beta=(X'X)^{-1}X'Y= (\sum_{i\in U} x_i x_i')^{-1}(\sum_{i\in U} x_i y_i)= T_x^{-1}t$, para el caso que se conocieran los valores de **X** e Y para el total de la población. Sin embargo como en la práctica los que se conocen son los valores para los elementos de una muestra se aplica el estimador de Horvitz Thompson para $\mathbf{T_x}$ y **t**, obteniéndose:

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{T}}_x^{-1}\,\hat{\boldsymbol{t}} = (\sum_{i\in S} \boldsymbol{x}_i \boldsymbol{x}_i'/\pi)^{-1}\sum_{i\in S} \boldsymbol{x}_i y_i/\pi_i = (\boldsymbol{X}'\boldsymbol{\Pi}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Pi}^{-1}\boldsymbol{Y}$$

(3.1.2.1)

donde Π denota la matriz diagonal que contiene las probabilidades de primer orden.

Luego: $\hat{y}_k = \boldsymbol{X}_k'\hat{\boldsymbol{\beta}}$ y una primera expresión del estimador GREG se obtiene como

$T_{yGREG}= \boldsymbol{X}'\hat{\boldsymbol{\beta}}$

Si bien los π-estimadores son insesgados, $\hat{\boldsymbol{\beta}}$ no lo es debido a que la esperanza de la inversa de una matriz no es igual a la inversa de su esperanza; por lo tanto $\hat{T}_{GREG}$ *no es insesgado.*

Por otra parte, se puede reescribir a T_y como:

$$T_y= \sum_{k\varepsilon U} y_k = \sum_{k\varepsilon U} \hat{y}_k + \sum_{k\varepsilon U}(y_k - \hat{y}_k)$$

donde los valores de $y_k - \hat{y}_k = e_k$ son los residuos del ajuste. La primera sumatoria sería $(\sum_{k\varepsilon U} \boldsymbol{x}_k)'\hat{\boldsymbol{\beta}}$, mientras que la segunda sumatoria solo puede calcularse para la muestra, por lo tanto se puede estimar utilizando el estimador de Horvitz Thompson, obteniéndose:

$$\hat{T}_{YGREG} = (\sum_{k\varepsilon U} \boldsymbol{x}_k)'\hat{\boldsymbol{\beta}} + \sum_{k\varepsilon S}(y_k - \hat{y}_k)/\pi_k = (\sum_{k\varepsilon U} \boldsymbol{x}_k)'\hat{\boldsymbol{\beta}} + \sum_{k\varepsilon S} e_k/\pi_k = (\sum_{k\varepsilon U} \boldsymbol{x}_k)'\hat{\boldsymbol{\beta}} + \sum_{k\varepsilon S} \breve{e}_k$$

(3.1.2.2)

Sin embargo, no es necesario conocer los valores de $\mathbf{x}_k$ para todos los valores de la población ya que si se conocen los totales marginales poblacionales de las variables auxiliares, $\mathbf{t_x}=(t_{x1}, t_{x2},\ldots, t_{xp})'$ y su estimador de Horvitz-Thompson, $\hat{\boldsymbol{t}}_{x\Pi} = (\hat{t}_{x1\Pi}, \hat{t}_{x2\Pi},\ldots,\hat{t}_{xp\Pi})'$, al redistribuir en 3.1.2.2 se tiene:

$$\hat{T}_{yGREG} = (\sum_{k\varepsilon U} \boldsymbol{x}_k)' \hat{\boldsymbol{\beta}} + \sum_{k\varepsilon S}(y_k / \pi_k) - \sum_{k\varepsilon S}(\hat{y}_k / \pi_k) = \boldsymbol{t}_x' \hat{\boldsymbol{\beta}} + \hat{T}_{y\pi} - \sum_{k\in S} \boldsymbol{x}_k' \hat{\boldsymbol{\beta}} / \pi_k$$

que constituye una expresión del estimador de regresión en función de los marginales de **X** y de los valores de Y sobre la muestra.

Luego, puede expresarse al estimador GREG como:

$$\hat{T}_{yGREG} = \hat{T}_{y\pi} + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi})' \hat{\boldsymbol{\beta}} \qquad (3.1.2.3)$$

De (3.1.2.3) se observa que el estimador GREG puede pensarse como el estimador de Horvitz-Thompson más un término de ajuste.

Reemplazando $\hat{\boldsymbol{\beta}}$ se obtiene una expresión alternativa:

$$\hat{T}_y = \hat{T}_{y\pi} + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi})' \hat{\boldsymbol{\beta}} = \sum_s \pi^{-1} y_i + \sum_s (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi})' (\sum_s \boldsymbol{x}_i \pi^{-1} \boldsymbol{x}_i')^{-1} \boldsymbol{x}_i \pi^{-1} y_i$$

$$= \sum_s [1 + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi})' (\sum_s \boldsymbol{x}_i \pi^{-1} \boldsymbol{x}_i')^{-1} \boldsymbol{x}_i] \pi^{-1} y_i$$

Si llamamos $g_{i,s} = 1 + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi})' (\sum_s \boldsymbol{x}_i \pi^{-1} \boldsymbol{x}_i')^{-1} \boldsymbol{x}_i$, entonces se define:

$$\hat{T}_{YGREG} = \sum_s g_{i,s} y_i / \pi_i \qquad (3.1.2.4)$$

Dado que $y_i = E_i + \boldsymbol{x}_i' \boldsymbol{\beta}$; al reemplazar y_i en (3.1.2.4), queda que

$$\hat{T}_{YGREG} = \sum_s g_{i,s} \boldsymbol{x}_i' \boldsymbol{\beta} / \pi_i + \sum_s g_{i,s} E_i / \pi_i$$

En el capítulo siguiente se verá una deducción de la varianza aproximada del estimador de Regresión.

3.2 Estimadores de Calibración

La definición de *estimador de calibración* se basa en el concepto de considerar pesos iniciales d_k, como por ejemplo los pesos de diseño, y modificarlos utilizando la información auxiliar disponible pero de manera que guarden la mínima distancia posible a los pesos originales d_k. Ésto se motiva por el hecho que los d_k generan un estimador insesgado y por lo tanto se buscará generar nuevos pesos que no se alejen demasiado de ellos. Se obtienen así nuevos pesos w_k para realizar la estimación.

Más precisamente, los nuevos pesos w_k se obtienen mediante la minimización de una función distancia adecuada entre los d_k y los w_k; sujeto a la restricción $\sum_s w_k \boldsymbol{x}_k = \sum_U \boldsymbol{x}_k$.

No necesariamente los nuevos pesos generarán un estimador insesgado pero sí es esperable que lo sea en forma aproximada.

Una distancia natural que puede considerarse en primer lugar es la asociada al estadístico χ^2 de Pearson: $E_p\left\{\sum_s \frac{(w_k - d_k)^2}{d_k}\right\}$. (3.2.1) ; donde E_p indica la esperanza respecto a la probabilidad asociada al diseño.

A pesar de que la distancia definida en (3.2.1) que da origen al estimador GREG, resulte familiar también es posible considerar otras funciones de distancia alternativas, las cuales tendrán una serie de propiedades en común. Definiremos entonces una función distancia de manera más general.

3.2.1 Deducción del estimador de Calibración

Sea G(u) una función distancia con las siguientes propiedades:

1. G(u) es positiva, diferenciable y estrictamente convexa; o sea G(u)>0 y G''(u)> 0.
2. G(1)=G'(1)=0
3. G''(1)=1

G(u) está definida sobre un intervalo D(u) que contiene a 1 .

$G(w_k/d_k)$ mide la distancia entre los pesos de diseño d_k y los nuevos pesos w_k. Luego puede considerarse como distancia promedio a $E_p\left\{\sum_S G_k(w_k / d_k)\right\}$, la cual estimamos a partir de la muestra dada con $\sum_s d_k G(w_k / d_k)$.

Por lo tanto el problema de hallar los pesos w_k se plantea como el siguiente problema de optimización:

Minimizar $\sum_s d_k G(w_k / d_k) - \boldsymbol{\lambda}' \left(\sum_s w_k \boldsymbol{x}_k - \sum_U \boldsymbol{x}_k\right)$,

donde $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_J)'$ es el correspondiente multiplicador de Lagrange. (3.2.1.1)

Al derivar respecto a w_k se obtiene

$g(w_k/d_k) - \boldsymbol{x}'_k \boldsymbol{\lambda} = 0$, donde $g(u)=dG/du$.

Si la solución existe, los supuestos garantizan que es única.

La solución que se obtiene es:

$$w_k = d_k F(\boldsymbol{x}'_k \boldsymbol{\lambda}) \qquad (3.2.1.2)$$

donde $F(u)=g^{-1}(u)$; función inversa de $g(u)$ que refleja la $Im(g)$ sobre $D(u)$ de manera creciente.

La existencia de F está garantizada por la propiedad 1 de G(u) .

Además:

- $F(0)=g^{-1}(0)= (dG(u)/dx)^{-1}(0)=1$ (por la propiedad 2). (3.2.1.3)
- $F'(0)= (g^{-1})'(0)= 1/g'(g^{-1}(0))=1/(d^2G(u)/dx^2)(1)=1$ (por la propiedad 3) (3.2.1.4)

Es necesario entonces hallar el valor de $\boldsymbol{\lambda}$ para determinar los pesos de calibración. Para ello se resuelven las siguientes ecuaciones que son llamadas *ecuaciones de calibración*:

$$\sum_s d_k F(\boldsymbol{x}'_k \boldsymbol{\lambda})\, \boldsymbol{x}_k = \sum_u \boldsymbol{x}_k \qquad (3.2.1.5)$$

donde la incógnita es $\boldsymbol{\lambda}$. Estas ecuaciones pueden resolverse por métodos numéricos.

Una vez hallado $\boldsymbol{\lambda}$ el estimador de calibración queda definido como:

$$\hat{T}_{Ycal} = \sum_s w_k y_k = \sum_{k \in s} d_k y_k F_k(\boldsymbol{x}_k' \boldsymbol{\lambda}) \qquad (3.2.1.6)$$

Si se define la función $\phi_s(\lambda) = \sum_s d_k \{F(\boldsymbol{x}_k' \boldsymbol{\lambda}) - 1\} \boldsymbol{x}_k$ (3.2.1.7)

entonces (3.2.1.1) puede escribirse como

$$\phi_s(\boldsymbol{\lambda}) = \boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi} \qquad (3.2.1.8)$$

con $\boldsymbol{t}_x = \sum_s w_k \boldsymbol{x}_k = \sum_s d_k F_k(\boldsymbol{x}_k' \boldsymbol{\lambda}) \boldsymbol{x}_k$

Es importante notar que si las observaciones y_k son combinación lineal de las $\boldsymbol{x}_k$, es decir, existe un vector β tal que $y_k = \boldsymbol{x}'_k \beta \ \forall k$, entonces $\hat{T}_{Ycal} = T_Y$, cualquiera sea la muestra s y por lo tanto la varianza del estimador será nula.

Luego, los pasos a seguir para obtener el estimador pueden resumirse en dos:

1. Resolver (3.2.1.8) para la función distancia elegida y obtener así $\boldsymbol{\lambda}$. En general se utilizan métodos iterativos para la resolución.
2. Se obtiene el *estimador de calibración* de T_Y como

$$\hat{T}_{Ycal} = \sum_k w_k y_k = \sum_s d_k F_x(\boldsymbol{x}_k' \boldsymbol{\lambda}) y_k$$

Deville y Särndal (1992) proponen diferentes funciones distancia que dan origen a los métodos más utilizados:

3.2.2 Método lineal

En este caso la distancia utilizada corresponde a una función cuadrática:

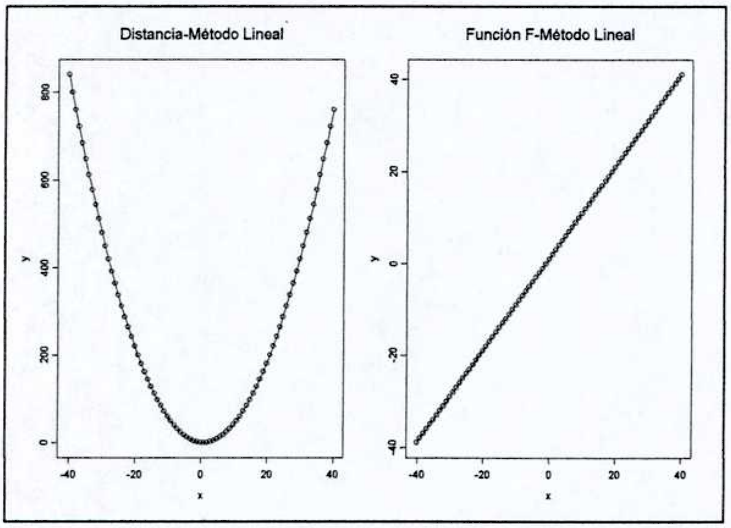
$G(x) = \frac{1}{2}(x-1)^2 \ ; \ x \in R$

la cual tiene asociada $F(u) = \left(\frac{\partial G}{\partial x}\right)^{-1}$ que resulta:

$$\frac{\partial G}{\partial x} = x - 1 \Rightarrow F(u) = 1 + u; \ u \in R$$

En los siguientes gráficos se presentan las funciones G(x) y F(u).

Gráfico1: Funciones asociadas al método lineal



Los pesos calibrados se calculan luego como:

$$\boxed{w_k = d_k(1 + \boldsymbol{x}_k' \boldsymbol{\lambda})} \qquad (3.2.2.1)$$

Al reemplazar en las ecuaciones de calibración, se obtiene:

$$\sum_s d_k(1 + \boldsymbol{x}_k' \boldsymbol{\lambda})\boldsymbol{x}_k = \sum_u \boldsymbol{x}_k \Rightarrow \sum_s d_k \boldsymbol{x}_k + (\sum_s d_k \boldsymbol{x}_k' \boldsymbol{x}_k)\boldsymbol{\lambda} = \sum_u \boldsymbol{x}_k$$

$$\Rightarrow (\sum_s d_k \boldsymbol{x}_k' \boldsymbol{x}_k)\boldsymbol{\lambda} = \boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi} \qquad (3.2.2.2)$$

Luego, el estimador de calibración puede escribirse como:

$$\hat{T}_{YCal} = \sum_{k\in S} w_k y_k = \sum_{k\in S} d_k(1 + \boldsymbol{x}_k'\boldsymbol{\lambda})y_k = \hat{T}_{y\pi} + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi})'(\sum_{k\in S} d_k \boldsymbol{x}_k'\boldsymbol{x}_k)^{-1}\sum_{k\in S} \boldsymbol{x}_k d_k y_k$$

(3.2.2.3)

El estimador de calibración que surge de utilizar esta distancia coincide con el Estimador General de Regresión; es decir que en este caso, *ambos métodos son equivalentes.*

3.2.3 El método Multiplicativo o de Raking

En este caso la distancia utilizada corresponde a la siguiente función:

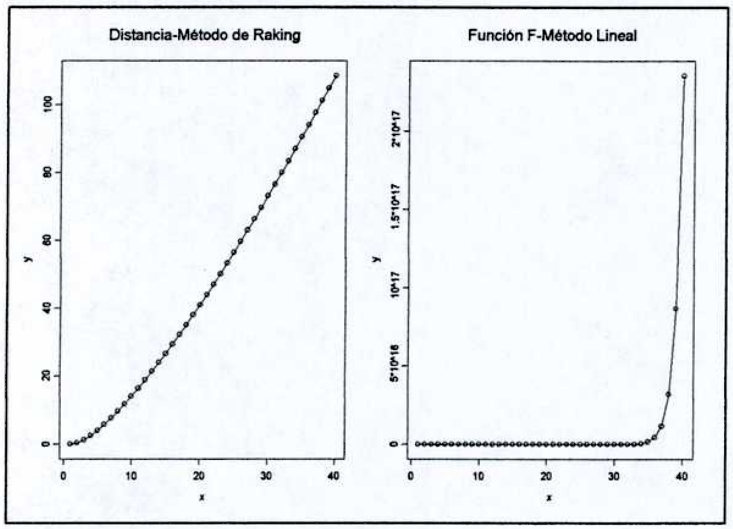
$G(x)=x\log x-x+1;\ \ x>0$ (distancia de tipo "Entropía") (3.2.3.1)

$G'(x)=\log(x)+x\,x^{-1}-1=\log(x) \quad x>0$

$$F(u)=\left(\frac{\partial G}{\partial x}\right)^{-1}=e^{u} \qquad ;\ u \in R$$

A continuación se presentan los gráficos de G(x) y F(u):

Gráfico 2: Funciones asociadas al método multiplicativo



Luego, los pesos de calibración resultan:

$$w_k = \exp\left(\sum_{j=1}^{p} \lambda_j x_{kj}\right) \qquad k=1,\ldots,n \qquad (3.2.3.2)$$

Luego, $w_k>0\ \ \forall k \in s$; es decir que el método de raking siempre genera pesos positivos; sin embargo a veces pueden ser muy grandes.

Muchas veces se desea restringir el rango de los nuevos pesos w_k para evitar la existencia de pesos extremos, pero sin alterar demasiado la estimación puntual; ésto motiva los métodos *Logit y Lineal Truncado.*

3.2.4 El método Logit

Se definen L y U dos constantes tales que $0<L<1<U$ y $A=\frac{U-L}{(1-L)(U-1)}$

Luego se considera la siguiente función de distancia:

$$G(x)=\begin{cases}\frac{1}{A}\left[(x-L)\log\frac{x-L}{1-L}+(U-x)\log\frac{U-x}{U-1}\right] & si\ L<x<U\\ \infty & en\ otro\ caso\end{cases},$$

Para hallar F(u) se debe encontrar la expresión de G´(x):

$$G'(x)=\frac{1}{A}[\log\frac{x-L}{1-L}+(x-L)\frac{1-L}{x-L}\frac{1}{1-L}+(-1)\log\frac{U-x}{U-1}+(U-x)\frac{U-1}{U-x}(\frac{-1}{U-1})$$

$$G'(x)=\frac{1}{A}[\log\frac{x-L}{1-L}-\log\frac{U-x}{U-1}]$$

Para hallar $(G´(x))^{-1}$: $u=G´(x) \Rightarrow Au=\log\frac{\frac{x-L}{1-L}}{\frac{U-x}{U-1}}$

$$\Rightarrow e^{Au}=\frac{(x-L)(U-1)}{(1-L)(U-x)}$$

$$e^{Au}(1-l)(U-x)=(x-L)(U-1)$$

$$e^{Au}(1-L)U-e^{Au}(1-L)x-x(U-1)=-L(U-1)$$

$$e^{Au}(1-L)U-x[e^{Au}(1-L)+(U-1)]=-L(U-1)$$

$$e^{Au}(1-L)U+L(U-1)=x[e^{Au}(1-L)+(U-1)]$$

$$x=\frac{e^{Au}(1-L)U+L(U-1)}{e^{Au}(1-L)+(U-1)}$$

Se obtiene así la función $F(u)=\frac{L(U-1)+U(1-L)e^{Au}}{U-1+(1-L)e^{Au}}$, $L<u<U$

Este método se conoce como *Método logit (L,U)*

Se tiene que $F(-\infty)=L$ y $F(+\infty)=U$ y además los pesos w_k resultantes siempre quedan acotados en el intervalo $[d_kL,d_kU]$.

A continuación se presentan los gráficos de las funciones $G(x)$ y $F(u)$ para diferentes valores de L y U:

Gráfico 3: Funciones asociadas al método Logit con $L=0.3$, $U=2$

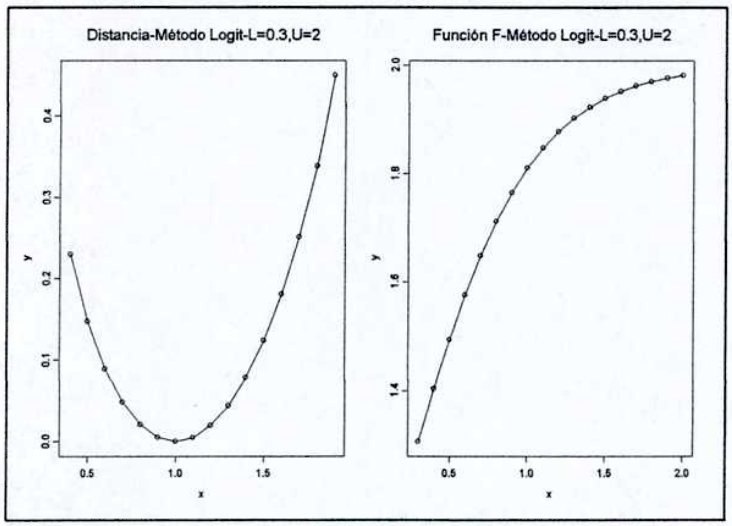


Gráfico 4: Funciones asociadas al método Logit con $L=0.5$, $U=3$

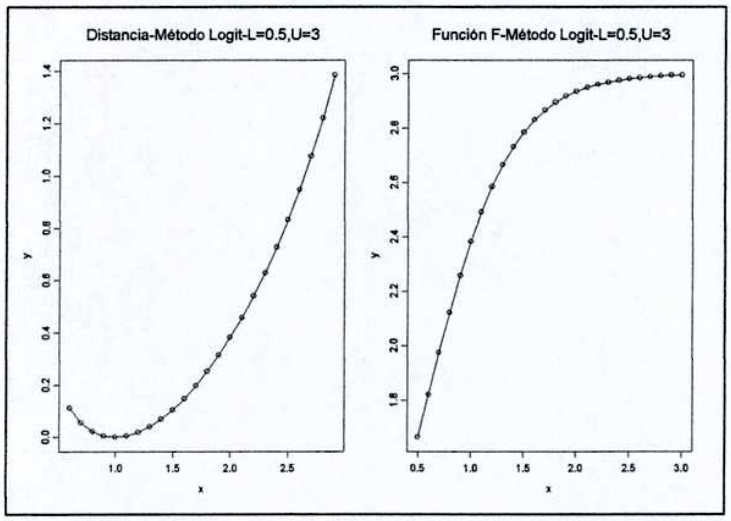
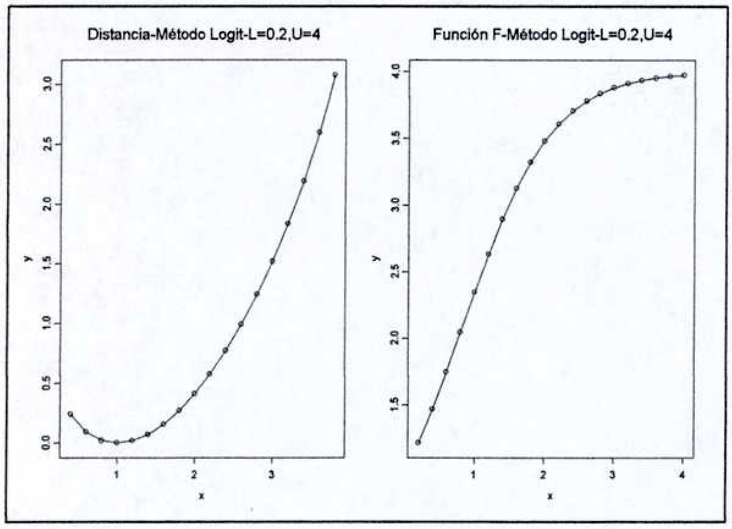


Gráfico 5: Funciones asociadas al método Logit con *L=0.2, U=4*



3.2.5 El método Lineal Truncado

Otra forma de conseguir pesos acotados es utilizar una función lineal acotada, es decir una función definida de la siguiente forma:

Siendo *L* y *U* dos constantes fijas, que verifican $0<L<1<U$, se define:

$$G(x) = \begin{cases} \frac{1}{2}(x-1)^2 & si\ L < x < U \\ \infty & en\ otro\ caso \end{cases}$$

Luego, la función *F* resulta:

$$F(x) = \begin{cases} 1+u & si\ \ L-1 \le u \le U-1 \\ L & si\ u < L-1 \\ U & si\ u > U-1 \end{cases}$$

Y los pesos resultantes w_k quedan acotados en el intervalo $[d_kL, d_kU]$.

A continuación se presentan los gráficos de las funciones G(x) y F(u) del método lineal truncado para diferentes valores de L y U:

Gráfico 6: Funciones asociadas al método Lineal Truncado con *L=0.2;U=3*

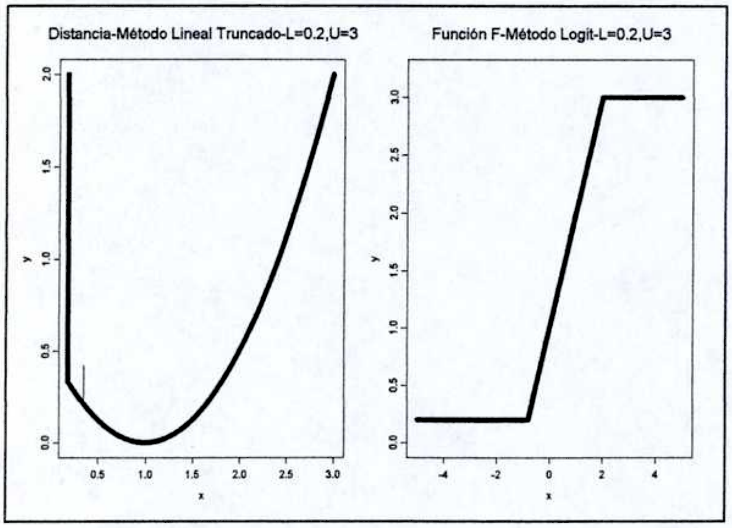
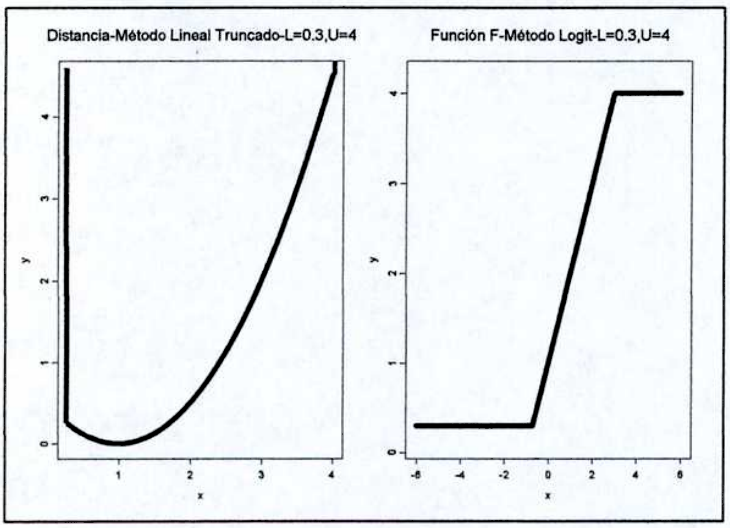


Gráfico 7: Funciones asociadas al método Lineal Truncado con *L=0.3;U=4*



3.2.6 Comparación de los métodos de calibración

El método Lineal es el más sencillo de aplicar ya que sólo requiere una iteración; sin embargo puede generar pesos negativos o bien pesos positivos pero muy grandes (se consideran pesos grandes cuando el cociente entre peso de calibración y peso de diseño es mayor a 3).

El método Multiplicativo o Raking no tiene el problema de generar pesos negativos ya que por su definición son siempre mayores a cero; pero no soluciona el problema de pesos demasiado grandes y por el contrario muchas veces la razón entre peso calibrado y peso de diseño puede ser más grande que si se utiliza el método lineal.

Los métodos Logit y Lineal Truncado solucionan tanto el problema de pesos negativos como el de pesos grandes ya que son siempre positivos y acotados. Sin embargo, no es sencillo elegir los valores de L y U apropiados ya que no pueden seleccionarse en forma arbitraria sino que sus valores máximos y mínimos valores dependen de los datos.

El método a elegir para calibrar en cada problema generalmente resulta del análisis de los pesos resultantes.

3.2.7 Propiedades del estimador de calibración

El estimador de calibración $\hat{T}_{YCal}$ verifica una serie de propiedades asintóticas siempre que puedan asumirse las hipótesis siguientes, en las cuales N indica el tamaño de la población y n el tamaño de la muestra si el diseño es de tamaño fijo o bien la esperanza de dicho tamaño si el diseño es de tamaño aleatorio:

H1: existe $lim\ N^{-1}\boldsymbol{T_x}$

H2: $N^{-1}(\hat{\mathbf{t}}_{xn} - \mathbf{t}_x) \rightarrow 0$ en probabilidad de diseño

H3: $n^{1/2}\ N^{-1}(\hat{\mathbf{t}}_{xn} - \mathbf{t}_x)$ converge en distribución a una Normal Multivariada de esperanza 0 y matriz de covarianza $\boldsymbol{A}$.

Las hipótesis en su conjunto implican que la diferencia entre los totales poblacionales correspondientes a la variable x y sus estimaciones a partir de la

muestra son pequeñas y tienen distribución aproximada Normal multivariada con matriz de covarianza $n^{-1}N^2\boldsymbol{A}$.

También resulta que $nN^{-1}\sum_k\sum_l \Delta_{kl}(\boldsymbol{x_k}/\pi_k)(\boldsymbol{x_l}/\pi_l)' = nN^{-2}V(\hat{\boldsymbol{t}}_{x\pi})$ converge a una matriz fija $\boldsymbol{A}$ y que $N^{-1}(\hat{\mathbf{t}}_{\mathbf{xn}} - \mathbf{t}_{\mathbf{x}}) = O_p(n^{-1/2})$ (3.2.7.1)

La matriz $\boldsymbol{A}$ puede pensarse como una matriz que describe el efecto asintótico asociado al diseño muestral elegido.

A partir de estos supuestos pueden deducirse las siguientes propiedades de la función ϕ:

a. $N^{-1}\phi_s(\boldsymbol{0}) = \boldsymbol{0}$

b. $\phi_s(\boldsymbol{0}) = \boldsymbol{0}$

c. $N^{-1}\phi_s'(\boldsymbol{0}) = N^{-1}T_s$

d. $\phi_s'(\boldsymbol{0}) = \boldsymbol{T} = \lim N^{-1}\sum_U \boldsymbol{x_k}\boldsymbol{x_k}'$

e. Para cada $\boldsymbol{\lambda}, \phi'$ es una matriz definida positiva debido a que $F(u)$ es una función creciente. Luego ϕ es inyectiva y por lo tanto existe ϕ^{-1} definida sobre un conjunto B, y además es continua y diferenciable.

f. Si llamamos $\phi_1 = N^{-1}\phi$, entonces $\left\|\phi_1^{-1}(\boldsymbol{x})\right\| \le \|\boldsymbol{x}\|K(1-\beta)^{-1}$ para $0 \le \beta \le 1/2$ y $K = max_{x\varepsilon B}\left\|(\phi_1^{-1})'(\boldsymbol{x})\right\|$.

Deville y Sändall (1992) demostraron que el estimador de calibración es asintóticamente equivalente al estimador GREG. Para ello, utilizaron una serie de resultados previos que se enuncian a continuación y cuya demostración puede consultarse en el Apéndice (2).

Resultado1: *La ecuación 3.2.1.7 tiene, con probabilidad 1, una única solución perteneciente a un dominio C, cuando* $n \to \infty$ *(C es un convexo abierto que incluye al 0 para todo n)*

Resultado 2

Si $\boldsymbol{\lambda}$*, solución de 3.2.1.7, existe, entonces tiende a 0 en probabilidad de diseño y además* $\boldsymbol{\lambda} = O_p(n^{-1/2})$

Si agregamos los siguientes supuestos:

a) $\max_{n,k} \|\mathbf{x}_k\| = M < \infty$

b) $\max_{n,k} F''(0) = M' < \infty$

pueden entonces demostrarse las siguientes propiedades del estimador de calibración:

Propiedad 1

$$\boldsymbol{\lambda} = \boldsymbol{T}^{-1}(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi}) + O_p(n^{-1})$$

Dem: Sea $\theta_k(u) = F_k(u) - 1 - q_k u$, con $\theta_k(u) = O(u^2)$. Luego, $\theta(u) = \max \theta_k(u) = O(u^2)$. Luego, (3.2.1.5) puede escribirse como

$$\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi} = \sum_s d_k \boldsymbol{x}_k \{\boldsymbol{x}_k' \boldsymbol{\lambda} + \theta_k(\boldsymbol{x}_k' \boldsymbol{\lambda})\}$$

y por lo tanto

$$\boldsymbol{\lambda} - \boldsymbol{T}^{-1}(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi}) = -\boldsymbol{T}^{-1} \sum_s d_k \boldsymbol{x}_k \theta_k(\boldsymbol{x}_k' \boldsymbol{\lambda}).$$

Para λ suficientemente pequeño

$$\left\| \lambda - T^{-1}(t_x - \hat{t}_{x\pi}) \right\| \le \left\| (N^{-1}T)^{-1} \right\| K'' \left\{ N^{-1} \sum_s d_k \|x_k\|^3 \right\} \|\lambda\|^2$$

Como $\left\| (N^{-1}T)^{-1} \right\| = O_p(1)$ y, usando la hipótesis H2 enunciada al principio de esta sección, se tiene que $N^{-1} \sum_s d_k \|\boldsymbol{x}_k\| = O_p(1)$. Por el *Resultado 2* $\|\boldsymbol{\lambda}\|^2 = O_p(n^{-1})$, con lo cual queda demostrado la *Propiedad 1*.

Propiedad 2:

El estimador de calibración definido en 3.2.1.5 es consistente en diseño y además $N^{-1}(\hat{T}_{Ycal} - \hat{T}_{\pi}) = O_p(n^{-1/2})$

Dem: Si llamamos $F'_k(0)=q_k$, entonces, utilizando la hipótesis b)

$$F_k(u)=1+q_k u+\theta_k(u) \qquad (3.2.7.2)$$

Si (3.2.1.7) tiene una solución λ, entonces:

$$(\hat{T}_{Ycal} - \hat{T}_{\pi}) = \sum_s d_k y_k \{q_k \boldsymbol{x}'_k \boldsymbol{\lambda} + \theta_k(\boldsymbol{x}'_k \boldsymbol{\lambda})\}$$

Luego, usando la hipótesis 3 aplicada a la variable $y_k\theta_k(\boldsymbol{x'_k\lambda})$, y dado el hecho que $\max \theta_k(u) = O(u^2)$, se obtiene que

$$N^{-1}\left|\hat{T}_{ycal} - \hat{T}_{\pi}\right| \le N^{-1}\left\{(\sum_s d_k q_k |y_k| \|\boldsymbol{x}_k\|)\|\boldsymbol{\lambda}\|\right\} + O_p(n^{-1})$$

donde

$$N^{-1}\left\{(\sum_s d_k q_k |y_k| \|\boldsymbol{x}_k\|)\|\boldsymbol{\lambda}\|\right\} = O_p(1) \text{ y } \boldsymbol{\lambda}=O_p(n^{-1/2}) \text{ (por el } \textit{Resultado 2}\text{)}$$

Luego, como $\hat{T}_{\pi}$ es consistente por diseño y $N^{-1}(\hat{T}_{Ycal} - T_Y) = O_p(n^{-1/2})$, queda demostrada la propiedad.

Observación:

Como $\hat{T}_{YCal}$ es el estimador "más cercano" a $\hat{T}_{\pi}$, es esperable que verifique algunas de las propiedades de este último. Por ejemplo, como $\hat{T}_{\pi}$ es insesgado por diseño sería esperable que $\hat{T}_{YCal}$ sea al menos asintóticamente insesgado por diseño. Ésto es cierto pero en el caso que 3.2.1.7 admita al menos una solución. Luego, se redefine el estimador como $\hat{T}_{YCal}$ si existe una solución a 3.2.1.7; sino se define como $\hat{T}_{\pi}$ (que equivale a tomar $\boldsymbol{\lambda=0}$)

Finalmente la siguiente propiedad establece la convergencia del estimador de calibración al estimador GREG

Propiedad 3

El estimador de calibración definido por (3.2.1.5) es asintóticamente equivalente al estimador de regresión generalizada $\hat{T}_{YGREG}$ en el sentido que

$N^{-1}(\hat{T}_{Ycal} - \hat{T}_{YGREG}) = O_p(n^{-1})$. *En consecuencia, los dos estimadores comparten la misma varianza asintótica.*

Dem: De (3.2.1.5) y (3.2.7.2) se desprende:

$$N^{-1}\hat{T}_{YCal} = N^{-1}\hat{T}_{Y\pi} + N^{-1}(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi})'T^{-1}\sum_s d_k q_k \boldsymbol{x}_k y_k + O_p(n^{-1}) + N^{-1}\sum_s d_k y_k \theta_k(\boldsymbol{x}_k'\boldsymbol{\lambda})$$

La suma de los dos primeros términos del miembro de la derecha es igual a $N^{-1}\hat{T}_{YGREG}$. El último término se probó en la Propiedad 2 que resulta ser un $O_p(n^{-1})$. Como consecuencia, se tiene que

$n^{1/2}N^{-1}(\hat{T}_{Ycal} - \hat{T}_{YGREG}) = O_p(n^{-1/2})$, con varianza asintótica igual a 0.

3.2.8 Calibración sobre tablas de frecuencia con celdas conocidas

Si los datos se disponen en una tabla de frecuencias de dimensión rxc , siendo U_{ij} la celda ij y N_{ij} la cantidad de individuos de U contenidos en U_{ij}, luego, el tamaño de la población es igual a $N=\sum_i\sum_j N_{ij.}$ y hay dos circunstancias que pueden originarse y que tienen que ver con el tipo de información con que se cuenta a nivel poblacional:

- Calibración con poblaciones conocidas a nivel de celda (Post Estratificación Completa).
- Calibración con poblaciones conocidas a nivel de marginales (Post Estratificación Incompleta).

En cada caso es diferente la forma en que se define el vector de información auxiliar $\mathbf{x_k}$ y la forma de descomponer al vector $\boldsymbol{\lambda}$.

Post Estratificación Completa: en este caso las componentes de los vectores $\boldsymbol{x_k}$ indican si el individuo pertenece o no a la celda ij. La dimensión de $\boldsymbol{x_k}$ es cxr y sus coordenadas son iguales a 0 salvo la que corresponde a la celda a la cual pertenece el k-ésimo individuo, donde vale 1. Por lo tanto, $\boldsymbol{t_x} = \sum_U \boldsymbol{x_k}$ es un

vector de dimensión rc y cada componente es igual a la cantidad de individuos de la población pertenecientes a la celda ij, es decir que es igual a N_{ij}. Por otra parte al realizar el producto $\boldsymbol{x'_k\lambda}$ para cada individuo perteneciente a la celda ij se obtiene una constante que se puede notar λ_{ij}.

Post Estratificación Incompleta: la calibración en este caso se llama *raking generalizado* y se utiliza en lugar de la anterior en los siguientes casos:

- Se conocen los totales poblacionales a nivel marginal pero no a nivel de celdas.
- La muestra en algunas celdas es muy pequeña o sin observaciones. En este caso, si bien la calibración por celdas es posible, produciría estimadores inestables o indefinidos.
- La información auxiliar proviene de una encuesta a partir de la cual y debido a su tamaño muestral se obtienen estimaciones poblacionales más precisas a nivel marginal que a nivel de celdas.

El vector $\boldsymbol{x_k}$ se debe definir de manera que la suma de ellos resuma la información del vector de marginales. Por lo tanto

$$\boldsymbol{x_k}=(\delta_{1.k},\ldots,\delta_{r.k},\ldots,\delta_{.ck})' \qquad (3.2.8.1)$$

donde $\delta_{i.k} = 1$ si el k-ésimo elemento pertenece a la fila i y 0 en caso contrario. De la misma forma $\delta_{j.k} = 1$ si el k-ésimo elemento pertenece a la columna j y 0 en caso contrario.

Resulta entonces que la suma de las $\boldsymbol{x_k}$ contiene los totales poblacionales de las filas (N_{i+}) y de las columnas (N_{+j}):

$$\sum_U \boldsymbol{x_k} = (N_{1+},\ldots,N_{r+},N_{+1},\ldots N_{+c})' \qquad (3.2.8.2)$$

Los estimadores en post estratificación completa e incompleta tienen la misma forma, salvo que en la post estratificación incompleta los totales poblacionales de cada celda resultan estimados a partir de la muestra, mientras que en la post estratificación completa son totales conocidos (Deville, Särndal, Sautory 1993)

3.2.9 Método iterativo para resolver las ecuaciones de calibración

Para resolver las ecuaciones de calibración, y principalmente con fines computacionales, puede utilizarse el Método de Newton para hallar $\boldsymbol{\lambda}$.

Definiendo la función ϕ como en (3.2.1.7): $\phi(\boldsymbol{\lambda}) = \sum_s d_k \{F(\boldsymbol{x}'_k \boldsymbol{\lambda}) - 1\}\boldsymbol{x}_k$, entonces las ecuaciones de calibración dadas por (3.2.1.5) pueden escribirse en términos de ϕ :

$$\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi} - \phi_s(\boldsymbol{\lambda}) = 0 \qquad (3.2.9.1)$$

Iniciando las iteraciones con $\boldsymbol{\lambda}_0 = \boldsymbol{0}$ se van generando los sucesivos valores de $\boldsymbol{\lambda}_m$, m=1,2,... como

$$\boldsymbol{\lambda}_{m+1} = \boldsymbol{\lambda}_m + \left\{\frac{\partial \phi_s(\boldsymbol{\lambda}_m)}{\partial \boldsymbol{\lambda}}\right\}^{-1} \left\{(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi}) - \phi(\boldsymbol{\lambda}_m)\right\} \qquad (3.2.9.2)$$

La expresión de la función ϕ depende de la función de distancia que se utilice; sin embargo para $\boldsymbol{\lambda}_0 = \boldsymbol{0}$ su derivada se calcula como

$$\frac{\partial \phi_s(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = \sum_s d_k \{F'(\boldsymbol{x}'_k \boldsymbol{\lambda})\, \boldsymbol{x}'_k\}\boldsymbol{x}_k \quad \Rightarrow \quad \frac{\partial \phi_s(\boldsymbol{0})}{\partial \boldsymbol{\lambda}} = \sum_s d_k \{F'(0)\, \boldsymbol{x}'_k\}\boldsymbol{x}_k$$

Como se mostró en (3.1.2.4), $F'(0)=1$ y por lo tanto se obtiene el siguiente valor que no depende de la expresión de ϕ

$$\frac{\partial \phi_s}{\partial \boldsymbol{\lambda}}(\boldsymbol{0}) = \sum_s d_k \boldsymbol{x}'_k \boldsymbol{x}_k \qquad (3.2.9.3)$$

Luego, $\boldsymbol{\lambda}_1 = \left(\sum_s d_k \boldsymbol{x}'_k \boldsymbol{x}_k\right)^{-1} \left(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi} - \phi_s(\boldsymbol{0})\right)$.

Como se mostró en (3.2.1.3) , $F(\boldsymbol{0})=1$ y por lo tanto $\phi(\boldsymbol{0}) = 0$, con lo cual

$\boldsymbol{\lambda}_1 = \left(\sum_s d_k \boldsymbol{x}'_k \boldsymbol{x}_k\right)^{-1} \left(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi}\right)$ que, de acuerdo a (3.2.2.1), coincide con la solución de las ecuaciones de calibración cuando se utiliza el método Lineal. Por lo tanto cuando la función de distancia elegida es la que genera el método lineal, el método iterativo finaliza en el paso 1. Para el resto de las funciones se debe continuar hasta la convergencia de $\boldsymbol{\lambda}$.

3.2.10 Uso de la Inversa Generalizada para la obtención de $\boldsymbol{\lambda}$

Uno de los principales problemas en la aplicación de los estimadores de calibración es que en los casos de post estratificación incompleta la matriz de diseño en general no es de rango completo y por lo tanto no admite inversa. Una forma de sortear el inconveniente es reduciendo la dimensión de la matriz mediante la eliminación de las columnas redundantes; sin embargo, hay problemas donde la redundancia no es tan clara de reconocer a priori e incluso en algunos modelos de pesaje el chequeo de columnas redundantes puede resultar impracticable. Una solución al problema es el uso de la *Inversa Generalizada*.

Si $\boldsymbol{A}$ es una matriz pxq de cualquier rango, se define la inversa generalizada de $\boldsymbol{A}$ y se denota $\boldsymbol{A^-}$ como aquella tal que $\boldsymbol{A\,A^-A = A}$.

Si se considera el sistema $\boldsymbol{Ax=y}$ entonces, $\boldsymbol{x=A^-y}$ es una solución del sistema. En efecto: $\boldsymbol{A(A^-y)=A(A^-Ax)=Ax=y}$.

En el caso particular en que se considera la inversa generalizada **G** de la matriz $\boldsymbol{X^t \Lambda X}$, donde $\boldsymbol{\Lambda}$ es una matriz diagonal de orden nxn con elementos diagonales estrictamente positivos y $\boldsymbol{X}$ es la matriz de diseño de un modelo de pesaje, pueden demostrarse las siguientes propiedades:

a) $\boldsymbol{G^t}$ es también inversa generalizada de $\boldsymbol{X^t \Lambda X}$

b) $\boldsymbol{XGX^t \Lambda X = X}$, es decir que $\boldsymbol{GX^t \Lambda}$ es una inversa generalizada de $\boldsymbol{X}$.

c) $\boldsymbol{XGX^t}$ es invariante a la elección de $\boldsymbol{G}$.

d) $\boldsymbol{XGX^t = XG'X^t}$ donde $\boldsymbol{G}$ puede ser o no una matriz simétrica.

Considerando al estimador de regresión generalizada como un caso particular del estimador de calibración podemos reescribir las ecuaciones que lo definen utilizando la inversa generalizada de la siguiente forma:

$$\hat{T}_{YGREG} = \sum_{k \in S} w_k y_k \quad \text{con} \quad w_k = \frac{1}{\pi_k} + \lambda_k \boldsymbol{x^t}_k \boldsymbol{G}\,(\boldsymbol{t_x} - \hat{\boldsymbol{t}}_{\boldsymbol{xn}})$$

donde $\boldsymbol{G}$ denota la inversa generalizada de $\boldsymbol{X^t \Lambda_s X}$ y $\boldsymbol{\Lambda_s} = diag(\frac{1}{\pi_1},\ldots,\frac{1}{\pi_n})$

Propiedad: *si existe un vector de dimensión n,* $\boldsymbol{w}$, *tal que*

$\boldsymbol{X'w = t_x}$, *(3.2.10.1)*

entonces los pesos w_k del estimador de regresión son invariantes a la elección de ***G.***

(Es importante notar que la condición que se requiere equivale a las ecuaciones de calibración)

Dem:

Sea $\boldsymbol{F}$ otra inversa generalizada de $\boldsymbol{X^t \Lambda_s X}$,diferente de $\boldsymbol{G}$, entonces:

Si llamamos v=w-d , con $d=(\frac{1}{\pi_1},\ldots,\frac{1}{\pi_n})^t$, se tiene que:

$$\boldsymbol{XG}(\boldsymbol{t_x} - \hat{\boldsymbol{t}}_{\boldsymbol{xn}}) = \boldsymbol{XGX^tv} \qquad \text{por } (3.2.10.1)$$

$$= \boldsymbol{XFX^tv} \qquad \text{por la propiedad c)}$$

$$= \boldsymbol{XF}(\boldsymbol{t_x} - \hat{\boldsymbol{t}}_{\boldsymbol{xn}}) \quad \text{por } (3.2.10.1)$$

Luego, $\boldsymbol{x^tG}(\boldsymbol{t_x} - \hat{\boldsymbol{t}}_{\boldsymbol{xn}})$ es invariante respecto a $\boldsymbol{G}$ para todo $k \in S$ y por lo tanto los pesos de regresión resultan invariantes respecto a la elección de $\boldsymbol{G}$.

Puede mostrarse también que los pesos utilizando esta inversa reproducen los totales poblacionales de las variables auxiliares:

$$\sum_{k\in s} w_k \boldsymbol{x}_k = \sum_{k\in s} (\pi_k^{-1} + \lambda_k \boldsymbol{x_k}^t \boldsymbol{G}(\boldsymbol{t_x} - \hat{\boldsymbol{t}}_{\boldsymbol{xn}}))\boldsymbol{x_k}$$

$$= \hat{\boldsymbol{t}}_{\boldsymbol{xn}} + \sum_{k\in S} \boldsymbol{x_k}\lambda_k \boldsymbol{x}^t{}_k \boldsymbol{G}(\boldsymbol{t_x} - \hat{\boldsymbol{t}}_{\boldsymbol{xn}})$$

$$= \hat{\boldsymbol{t}}_{\boldsymbol{xn}} + (\boldsymbol{X^t \Lambda X})\boldsymbol{G}(\boldsymbol{t_x} - \hat{\boldsymbol{t}}_{x\pi})$$

$$= \hat{\boldsymbol{t}}_{x\pi} + (\boldsymbol{X^t \Lambda X})\boldsymbol{G}\,\boldsymbol{X^tv} \qquad \text{por } (3.2.10.1)$$

$$= \hat{\boldsymbol{t}}_{\boldsymbol{xn}} + \boldsymbol{X^tv} \qquad \text{por las propiedades b) y d)}$$

$$= \hat{\boldsymbol{t}}_{\boldsymbol{xn}} + (\boldsymbol{t_x} - \hat{\boldsymbol{t}}_{\boldsymbol{xn}}) = \boldsymbol{t_x} \qquad \text{por } (3.2.10.1)$$

CAPÍTULO IV

VARIANZA DE LOS ESTIMADORES

Como se vio en el capítulo anterior, los estimadores generalmente están dados por expresiones complejas. No es sencillo entonces, hallar una expresión de varianza para estos estimadores, que a su vez resulte aceptable desde el punto de vista práctico. Es por ello que se recurre a técnicas de aproximación, siendo una de las más utilizadas, la *linearización de Taylor*, la cual ha sido muy difundida y aplicada en estimaciones a partir de encuestas de gran tamaño. El método consiste en expresar un parámetro no lineal como una función de totales de otras variables, aplicar un desarrollo por series de Taylor y hallar la varianza de esta aproximación.

Más recientemente han surgido como técnicas alternativas algunas tales como el jackknife y el bootstrap que son métodos de remuestreo que van cobrando auge debido al desarrollo de la computación.

4.1 ***Linearización de Taylor***

Se basa en los siguientes principios:

i. Si θ es un parámetro que es función de totales poblacionales t_i, o sea, $\theta=f(t_1,\ldots,t_q)$, entonces un estimador de θ se obtiene aplicando f sobre los π-estimadores de los t_j.

ii. Si θ es un parámetro lineal, $\theta = a_0 + \sum_{j=1}^{q} a_j t_j$, entonces $\hat{\theta} = a_0 + \sum_{j=1}^{q} a_j \hat{t}_{j\pi}$ es un estimador insesgado (Särndall, pag 172).

Una expresión alternativa para $\hat{\theta}$ resulta ser

$$\hat{\theta} = a_0 + \sum_s \breve{u}_k \text{, donde } \breve{u}_k = u_k / \pi_k \text{ y } u_k = \sum_{j=1}^{q} a_j y_{jk} . \qquad (4.1.1)$$

Luego, la varianza resulta: $v(\hat{\theta}) = V(\sum_s \breve{u}_k) = \sum\sum_U \Delta_{kl}\breve{u}_k\breve{u}_l$ que puede estimarse como $\hat{v}(\hat{\theta}) = \sum\sum_s \breve{\Delta}_{kl}\breve{u}_k\breve{u}_l$. (4.1.2)

Cuando θ es una función *no lineal* de los totales, es muy difícil obtener resultados exactos sobre el sesgo y la varianza de $\hat{\theta}$. Se utiliza entonces la técnica de linearizar utilizando un desarrollo de Taylor de primer orden que lleva a una expresión aproximada para $v(\hat{\theta})$ y también para $\hat{v}(\hat{\theta})$.

Luego, *esta técnica aproxima un estimador no lineal* $\hat{\theta}$ *por otro* $\hat{\theta}_0$ *que es una función lineal de los* $\hat{t}_{j\pi}$.

Si la parte del desarrollo de Taylor que se descarta (términos de orden 2 y superior) es pequeña, la aproximación es buena y $\hat{\theta}_0$ tendrá un comportamiento muy similar a $\hat{\theta}$. Por lo tanto, calculando la varianza de $\hat{\theta}_0$, que no es tan complejo de hallar, se obtiene una aproximación de la varianza de $\hat{\theta}$. Del mismo modo hallando un estimador para $v(\hat{\theta}_0)$.

En base a lo expresado se tiene:

$$\hat{\theta} = f(\hat{t}_{1\pi},\ldots,\hat{t}_{q\pi}) \cong f(t_1,\ldots,t_q) + \sum_{j=1}^{q} a_j\left(\hat{t}_{j\pi} - t_j\right)$$

$$\hat{\theta} \cong \hat{\theta}_0 = \theta + \sum_{j=1}^{q} a_j\left(\hat{t}_{j\pi} - t_j\right); \text{ siendo } a_j = \frac{\partial f}{\partial t_j} \quad j=1,\ldots,q \qquad (4.1.3)$$

para muestras grandes $(\hat{t}_{1\pi},\ldots,\hat{t}_{q\pi})$ tomará con alta probabilidad valores muy cercanos a $(t_1,\ldots,t_q)$ y por lo tanto, $\hat{\theta}$ estará muy próximo al estimador linearizado $\hat{\theta}_0$.

Los fundamentos teóricos de esta aproximación via linearización se basan en definiciones y resultados de la teoría de probabilidades que pueden consultarse en el apéndice (3).

El sesgo y la varianza de $\hat{\theta}$ puede aproximarse por el sesgo y la varianza de $\hat{\theta}_0$. Luego, si consideramos la notación introducida en (4.1.1) y utilizamos (4.1.2) llamando *AV* a la varianza aproximada, se tiene:

$$AV(\hat{\theta}) = V(\hat{\theta}_0) = V(\sum_{j=1}^{q} a_j \hat{t}_{j\pi}) = V(\sum_{s} \breve{u}_k) = \sum\sum_{U} \Delta_{kl}\breve{u}_k\breve{u}_l \qquad (4.1.4)$$

Como los a_j son generalmente desconocidos porque dependen de valores a nivel poblacional, para estimarlos se reemplaza cada total del cual depende a_j por su π-estimador, obteniéndose: $\hat{u}_k = \sum_{j=1}^{q} \hat{a}_j y_{jk}$. Al tomar varianza en (4.1.4) y reemplazando los u_k por los $\hat{u}_k$:

$$\hat{V}(\hat{\theta}) \cong \sum\sum_s \breve{\Delta}_{kl} \frac{\hat{u}_k}{\pi_k} \frac{\hat{u}_l}{\pi_l} \qquad (4.1.5)$$

Al ser $\hat{u}_k$ función de los π-estimadores, es consistente para u_k. Luego, $\hat{V}(\hat{\theta})$ es una función de estimadores consistentes de $\hat{u}_k$, y por lo tanto, para muestras grandes se comporta como si estuviera basado en el valor verdadero de u_k. Luego, $\hat{V}(\hat{\theta})$ puede asumirse consistente para $V(\hat{\theta})$.

4.2 Varianza del Estimador de Regresión Generalizada

El estimador de regresión $\hat{T}_{YGREG}$ no es insesgado (Cap III 3.1.2), sin embargo se demostrará a continuación que sí lo es en forma asintótica, y se hallará su varianza aproximada mediante la técnica de linearización.

Propiedad 4.2.1: *el estimador de regresión* $\hat{T}_{yGREG}$ *(dado por 3.1.3)* es asintóticamente insesgado para T con varianza aproximada

$$Var(\hat{T}_{yGREG}) = \sum_U \sum_U \Delta_{kl} \breve{E}_k \breve{E}_l \qquad (4.2.1)$$

donde $\breve{E}_k = \frac{E_k}{\pi_k}$ y $E_k = y_k - \boldsymbol{x}_k' \boldsymbol{\beta}$

Esta varianza puede a su vez estimarse como

$$\hat{V}(\hat{T}_{yGREG}) = \sum_s \sum_s \hat{\Delta}_{kl} \breve{e}_k \breve{e}_l \qquad (4.2.2)$$

siendo $\breve{e}_k = \frac{e_k}{\pi_k}$ $e_k = y_k - \boldsymbol{x}_k' \hat{\boldsymbol{\beta}}$

Dem:

De (3.1.2.3) podemos escribir a $\hat{T}_{YGREG}$ como:

$$\hat{T}_{yGREG} = \hat{T}_{Y\Pi} + (\boldsymbol{t}_X - \hat{\boldsymbol{t}}_{X\Pi})' \hat{\boldsymbol{\beta}} = \hat{T}_{Y\Pi} + (\boldsymbol{t}_X - \hat{\boldsymbol{t}}_{X\Pi})' \hat{T}^{-1}\hat{t} = f(\hat{T}_{Y\Pi}, \hat{\boldsymbol{t}}_{X\Pi}, \hat{T}, \hat{t})$$

(4.2.3)

O sea, que $\hat{T}_{YGREG}$ resulta ser un estimador no lineal que es función del π-estimador $\hat{T}_{y\pi}$, de los π-estimadores $\hat{t}_{x_j\pi}$ que son componentes de $\hat{t}_{x\pi}$; de los $\hat{T}_{jj'\pi}$ componentes de $\hat{T}$ y de los $\hat{t}_{j\pi}$ componentes de $\hat{t}$.

Para utilizar la linearización de Taylor de orden 1 se deben calcular las derivadas *parciales* respecto a cada uno de ellos

$$\frac{\partial f}{\partial \hat{T}_{y\pi}} = 1 \qquad (4.2.4) \qquad\qquad \frac{\partial f}{\partial \hat{\boldsymbol{t}}_{x_j\pi}} = -\hat{\beta}_j \qquad j = 1,\ldots,J \qquad (4.2.5)$$

$$\frac{\partial f}{\partial \hat{T}_{jj'\pi}} = (\boldsymbol{t}_X - \hat{\boldsymbol{t}}_{x\Pi})'(-\hat{T}^{-1}\Lambda_{jj'}\hat{T}^{-1})\hat{t} \qquad j \le j' = 1,\ldots,J \qquad (4.2.6)$$

$$\frac{\partial f}{\partial \hat{t}_{j\pi}} = (\boldsymbol{t}_X - \hat{\boldsymbol{t}}_{x\Pi})'\hat{T}^{-1}\lambda_j \qquad j = 1,\ldots,J \qquad (4.2.7)$$

donde :

λ_j es el j-ésimo vector con j-ésima componente igual a 1 y el resto 0;

$\Lambda_{jj'}$ es la matriz JxJ con 1 en las coordenadas jj' y j'j, y 0 en el resto.

(Graybill, 1983)

Al evaluar las derivadas en el valor esperado de los estimadores:

$\hat{T}_{y\pi} = T_y$, $\hat{\boldsymbol{t}}_{X\Pi} = \boldsymbol{t}_X$, $\hat{T} = T$, y $\hat{\boldsymbol{t}} = \boldsymbol{t}$, (utilizando el hecho que los estimadores de Horvitz-Thompson son insesgados)

se obtienen derivadas nulas para (4.2.6) y (4.2.7). En consecuencia, aplicando (4.1.3) al estimador (4.2.3):

$$\hat{T}_{yGREG} \approx \hat{T}_{yGREG\,0} = T + \mathbf{1}(\hat{T}_\pi - T) - \sum_{j=1}^{J} \beta_j(\hat{t}_{X_j\pi} - t_{X_j}) = \hat{T}_{y\pi} + (\boldsymbol{t}_X - \hat{\boldsymbol{t}}_{X\pi})'\beta =$$

$$= \sum_s \frac{y_k}{\pi_k} + \left(\sum_U \boldsymbol{x}'_k - \sum_s \frac{\boldsymbol{x}'_k}{\pi_k}\right)\boldsymbol{\beta} = \sum_U \boldsymbol{x}'_k\boldsymbol{\beta} + \sum_s \left(\frac{y_k}{\pi_k} - \frac{\boldsymbol{x}'_k}{\pi_k}\boldsymbol{\beta}\right) = \sum_U y^0{}_k + \sum_s \breve{E}_k$$

siendo $y^0_k = \mathbf{x}_k'\boldsymbol{\beta}$ y $E_k = y_k - y^0_k$ (4.2.8)

Calculando la esperanza del estimador se obtiene:

$E(\hat{T}_{yGREG}) \approx E(\hat{T}_{yGREG_0}) = \sum_U y^0_k + E(\sum_s \breve{E}_k) = \sum_U y^0_k + \sum_U (y_k - y^0_k) = \sum_U y_k = T$ por lo tanto $\hat{T}_{yGREG}$ es aproximadamente insesgado.

La varianza aproximada puede calcularse a partir de (4.2.8)

$V(\hat{T}_{yGREG}) = V(\sum_U y^0_k + \sum_S \breve{E}_k) = V(\sum_S \breve{E}_k) = \sum\sum_U \Delta_{kl} \breve{E}_k \breve{E}_l$

Reemplazando los $\breve{E}_k$ por sus estimadores $\breve{e}_k$ se obtiene un estimador de la varianza aproximada:

$$\hat{V}(\hat{T}_{yGREG}) = \sum_s \sum_s \hat{\Delta}_{kl} \breve{e}_k \breve{e}_l \qquad (4.2.9)$$

4.3 Varianza del estimador de calibración

No existe fórmula para la varianza del estimador de calibración ($\hat{T}_{YCAL}$). Sin embargo, en el Capítulo 3 sección 3.2.7 se demostró que los estimadores de calibración convergen al estimador GREG y por lo tanto *comparten la misma varianza aproximada, para muestras grandes.*

Luego, un estimador de varianza aproximada para $\hat{T}_{YCAL}$ sería (4.2.9).

En 1992, Deville y Särndal sugirieron el siguiente estimador para la varianza del estimador GREG, que puede entonces considerarse como un estimador de la varianza asintótica del estimador de calibración, cualquiera sea la función de distancia que se utilice:

$$\hat{V}(\hat{T}_{YGREG}) = \sum_{k\in S}\sum_{l\in S} \left(\frac{\Delta_{kl}}{\pi_{kl}}\right) w_k e_k \; w_l e_l \qquad (4.3.1)$$

con

$\Delta_{kl} = \pi_{kl} - \pi_k \ \pi_l$ $\quad e_k = y_k - \boldsymbol{x}_k' \hat{\boldsymbol{\beta}};$ $\quad \hat{\boldsymbol{\beta}} = (\sum_s d_k \boldsymbol{x}_k \boldsymbol{x}_k')^{-1} \sum_s d_k \boldsymbol{x}_k y_k$, con d_k los pesos de diseño.

Si bien el utilizar en (4.3.1) los pesos d_k del diseño o bien los pesos calibrados w_k producen estimadores consistentes de la varianza, Deville y Särndal sugirieron utilizar los w_k ya que se demuestra que esta opción reduce el sesgo en la varianza cuando se está ante un caso de muestras no muy grandes (Deville-Särndall, 1992)

La estimación de varianza dada por (4.3.1) involucra el cálculo de las probabilidades de segundo orden, π_{kl}, lo cual puede resultar complejo según el diseño que se utilice.

Por otra parte, para obtener la varianza aproximada vía linearización es necesario utilizar las derivadas de primer orden del estimador, lo cual no es posible para estimadores de otros estadísticos diferentes del total, tales como los estadísticos de orden.

Ambas consideraciones hacen recomendable el uso de otras técnicas alternativas para el cálculo de una varianza aproximada. Es por ello que se utilizarán aquellos estimadores aproximadamente insesgados más usuales, entre los que se encuentra la *varianza jackknife y la varianza jackknife linealizada*.

En este trabajo mostraremos para el caso de un total, en el cual pueden calcularse los distintos tipos de varianza mencionados, que los estimadores de varianza jackknife y jackknife linealizada se comportan de manera similar a la varianza asintótica; lo cual sugeriría la conveniencia de su uso de las varianzas jackknife como alternativas a la varianza asintótica en el caso de estadísticos donde esta última no es posible calcularse.

4.4 Estimación de la varianza con técnica Jackknife

El jackknife es un método no paramétrico basado en el remuestreo.

El objetivo de la técnica jackknife en muestreo es producir estimadores de la varianza para un estimador. El jackknife se basa en la partición de una muestra tomada inicialmente en A grupos de igual tamaño mediante un procedimiento aleatorio, que en general responde al mismo diseño que se utilizó para seleccionar la muestra. Nos ocuparemos del caso en que los grupos formados son de tamaño igual a n_I-1 dado que en el diseño que utilizaremos los elementos de la muestra serán conglomerados y por lo tanto, conformar un grupo de tamaño n_I-1 implicará prescindir de un conglomerado completo.

La teoría del método jackknife fue introducida por Quenouille en 1949 quien desarrolló el método con el fin de estimar el sesgo de un estimador, eliminando un dato por vez y recalculando el estimador utilizando los datos restantes. Más tarde, en 1958, Tukey utilizó la idea para construir estimadores de varianza.

Sea $\hat{\theta}$ el estimador de un parámetro θ calculado sobre una muestra de tamaño n. Y sea $\hat{\theta}_{(i)}$ el mismo estimador que $\hat{\theta}$ pero basado en la muestra de tamaño (n-1) que resulta de eliminar la i-ésima observación de la muestra dada. Quenouille definió el estimador jackknife del sesgo de $\hat{\theta}$ como $(n-1)(\bar{\hat{\theta}}_n - \hat{\theta})$, siendo

$\bar{\hat{\theta}}_n = \frac{\sum_{i=1}^{n} \hat{\theta}_{(i)}}{n}$; lo cual lleva a definir el siguiente estimador jackknife para θ:

$$\hat{\theta}_{JK} = \hat{\theta} - sesgo(\hat{\theta}) = n\hat{\theta} - (n-1)\bar{\hat{\theta}}_n \qquad (4.4.1)$$

Dado que $\hat{\theta}_{JK} = n\hat{\theta} - \frac{(n-1)}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)} = \frac{1}{n}\sum_{i=1}^{n}\left(n\hat{\theta} - (n-1)\hat{\theta}_{(i)}\right)$, (4.4.2)

Tukey llamó pseudo-valores a los términos de la sumatoria:

$\tilde{\theta}_{n,i} = n\hat{\theta} - (n-1)\hat{\theta}_{(i)}$ i=1,...,n (4.4.3)

y asumió que:

1. los pseudo-valores $\tilde{\theta}_{n,i}$ pueden considerarse i.i.d
2. $\tilde{\theta}_{n,i}$ tiene aproximadamente la misma varianza que $\sqrt{n}\hat{\theta}$

Luego, puede estimarse la var($\sqrt{n}\hat{\theta}$) con la varianza muestral basada en los pseudo-valores $\tilde{\theta}_{n,1}, \tilde{\theta}_{n,2},\ldots, \tilde{\theta}_{n,n}$, de donde surge luego el siguiente estimador de varianza de $\hat{\theta}$:

$$V_{JK1} = \frac{1}{n(n-1)} \sum_{i=1}^{n}\left(\tilde{\theta}_{n,i} - \frac{1}{n} \sum_{j=1}^{n} \tilde{\theta}_{n,i} \right)^2 \qquad (4.4.4)$$

que equivale a

$$V_{JK1} = \frac{n-1}{n} \sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \bar{\hat{\theta}}_n \right)^2 \quad (4.4.5)$$

ya que:

$$\tilde{\theta}_{n,i} - \frac{1}{n} \sum_{j=1}^{n} \tilde{\theta}_{n,i} = n\hat{\theta} - (n-1)\hat{\theta}_{(i)} - \frac{1}{n} \sum_{i=1}^{n}(n\hat{\theta} - (n-1)\hat{\theta}_{(i)}) = n\hat{\theta} - (n-1)\hat{\theta}_{(i)} - n\hat{\theta} + \frac{n-1}{n} \sum_{i=1}^{n}\hat{\theta}_{(i)}$$

$$= -(n-1)\left(\hat{\theta}_{(i)} - \bar{\hat{\theta}}_n \right)$$

de donde $V_{JK1} = \frac{1}{n(n-1)} \sum_{i=1}^{n}\left(\tilde{\theta}_{n,i} - \frac{1}{n} \sum_{j=1}^{n} \tilde{\theta}_{n,i} \right)^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n}\left((n-1)^2\left(\hat{\theta}_{(i)} - \bar{\hat{\theta}}_n \right)^2 \right)$

$$= \frac{n-1}{n} \sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \bar{\hat{\theta}}_n \right)^2$$

En la práctica también se utiliza el estimador de varianza alternativo:

$$\hat{V}_{JK2} = \frac{1}{n(n-1)} \sum_{i=1}^{n}(\hat{\theta}_{(i)} - \hat{\theta})^2 \qquad (4.4.6)$$

cuya definición se justifica a partir del hecho que $\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2 = \sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \bar{\hat{\theta}}_n \right)^2 + n\left(\bar{\hat{\theta}}_n - \hat{\theta} \right)^2$

y de donde resulta que $\hat{V}_{JK2} \geq \hat{V}_{JK1}$

Puede demostrarse que, en el caso en que θ es el total de la población y $\hat{\theta}$ el π-estimador, el estimador jackknife tiene buenas propiedades en lo que al sesgo se refiere.

4.4.1 Estimador jackknife para el Muestreo Estratificado

Cuando la técnica jackknife se utiliza en un muestreo estratificado, se utilizan otros estimadores basados en (4.4.3) y (4.4.4), pero adaptados a la estratificación.

Si $\hat{\theta}$ es el estimador de θ utilizando la muestra completa, sea $\hat{\theta}_{(hi)}$ el estimador de θ basado en la muestra que queda cuando se omite de la muestra el i-ésimo elemento del estrato h. Luego, un estimador de $\hat{V}(\theta)$ sugerido por Rust(1985), para el caso de un muestreo con reemplazo es:

$$\hat{V}_{JK3} = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{a=1}^{A_h} \left(\hat{\theta}_{(hi)} - \hat{\theta}\right)^2 \quad (4.4.1.1)$$

el cual será insesgado para var $(\hat{\theta})$ si $\hat{\theta}$ es el π-estimador para el total de la población.

En el caso en que el muestreo se realiza sin reemplazo, el estimador sugerido es:

$$\hat{V}_{JK3} = \sum_{h=1}^{H} \frac{(n_h - 1)(1 - f_h)}{n_h} \sum_{a=1}^{A_h} \left(\hat{\theta}_{(hi)} - \hat{\theta}\right)^2 \quad (4.4.1.2)$$

donde f_h es la fracción de muestreo en el estrato h; el coeficiente $(1-f_h)$ aparece, como se vió en el Cap I, como factor de corrección por el no reemplazo.

4.4.2 Estimador jackknife para el Muestreo Estratificado en dos etapas

Para estimar un total bajo un muestro estratificado en dos etapas con reposición, el estimador de varianza para el vendrá dado por:

$$\hat{V}\left(\hat{T}\right) = \sum_{h=1}^{L} \frac{1}{n_h(n_h - 1)} \sum_{i=1}^{n_h} \left(t_{hi} - \bar{t}_h\right)^2 \quad (4.4.1.3)$$

o bien $\hat{V}\left(\hat{T}\right) = \sum_{h=1}^{L} \frac{1}{n_h(n_h - 1)} \sum_{i=1}^{n_h} \left(t_{hi} - \hat{t}_h\right)^2$ (4.4.1.4)

donde L= cantidad de estratos

t_{hi}=total en el i-ésimo cluster= $\sum_k n_h w_{hik} y_{hik}$

$\bar{t}_h$ =promedio de totales entre clusters= $\frac{1}{n_h}\sum_i t_{hi}$

$\hat{t}_h$ = estimador del total en el estrato h

Para el caso sin reposición se ajustará con el coeficiente *(1-f_h).* En este trabajo, como se verá en el Cap V se ensayarán otros coeficientes de ajuste alternativos.

Para estimar la varianza utilizando la técnica jackknife se debe calcular el estimador $\hat{t}_{(gj)}$ para cada g_j ,que se obtiene *utilizando la muestra que queda luego de omitir los datos que provienen del j-ésimo cluster muestreado en el g-ésimo estrato* (Rao 1994). Es decir que se calcula el estimador del total eliminando un cluster por vez, para todos los clusters de todos los estratos.

Luego, para calcular el nuevo estimador se deben recalcular los pesos asignándole 0 a los elementos del cluster removido y modificando los pesos de los clusters que quedaron y que eran del mismo estrato eliminado. Es decir:

$$w_{hik(gj)} = \begin{cases} 0 & si\ h_i = g_j \\ \frac{n_g}{n_g - 1} w_{gik} & si\ h_i = g_j \ \ con \ \ i \neq j \\ w_{hik} & si\ h_i \neq g_j \end{cases} \qquad (4.4.1.5)$$

n_g corresponde a la cantidad de radios muestreados en el g-ésimo estrato.

El estimador $\hat{t}_{(gj)}$ que resulta al omitir el cluster g_j resulta:

$$\hat{t}_{(gj)} = \sum_{h_{ik} \in s} w_{hik(gj)}\, y_{hik} \qquad (4.4.1.6)$$

Una vez obtenido $\hat{t}_{(gj)}$ para todo g_j, es decir, luego de calcular todos los totales eliminando todos los clusters posibles de a uno por vez, se calcula el estimador jackknife de la varianza de $\hat{T}$ a través de la siguiente fórmula:

$$V_J\left(\hat{T}\right) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} \left(\hat{t}_{(gj)} - \hat{T}\right)^2 \qquad (4.4.1.7)$$

Este estimador es aplicable a cualquier estadístico suave, $\hat{\theta} = f(T)$ reemplazando $\hat{t}_{(gj)}$ por $\hat{\theta}_{(gj)} = f(\hat{t}_{(gj)})$.

Canty y Davison (1998) toman la misma expresión de Rao pero incorporan los factores de corrección por población finita:

$$V_J(\hat{T}) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g}(1 - \frac{n_g}{N_g}) \sum_{j=1}^{n_g} \left(\hat{t}_{(gj)} - \hat{T}\right)^2 \qquad (4.4.1.8)$$

De las expresiones anteriores se observa que se obtiene un estimador de la varianza global, y no de las componentes de la varianza.

4.4.2 Estimador jackknife para el estimador de calibración

El estimador de varianza stándard linearizado para el estimador GREG, de acuerdo a lo visto en 4.2 viene dado por (4.4.1.7) intercambiando los t_{hi} por los residuos $\tilde{e}_{hi}$, siendo

$$\tilde{e}_{hi} = \sum_k n_h w_{hik} e_{hik} \quad \text{con} \quad e_{hik} = y_{hik} - \boldsymbol{x}'_{hik}\hat{\boldsymbol{\beta}} \qquad (4.4.2.1)$$

luego, $\hat{V}_L(\hat{T}_{GREG}) = \hat{V}_L(\tilde{e}_{hi})$ (4.4.2.2)

Como muestra Stuckel et al (1996), para la estimación jackknife es necesario recalcular los pesos de calibración w^*_{hik} cada vez que un cluster g_j es removido. Para el estimador GREG estos pesos vienen dados por la siguiente expresión:

$$w^*_{hik}(g_j) = w_{hik}(g_j)\, a_{hik}(g_j) \qquad (4.4.2.3)$$

donde

$$a_{hik}(g_j) = 1 + \boldsymbol{x}'_{hik}\boldsymbol{A}^{-1}(g_j)\left(\boldsymbol{X} - \hat{\boldsymbol{X}}_{(g_j)}\right) \qquad (4.4.2.4)$$

$$\hat{\boldsymbol{A}}_{(g_j)} = \sum_{h_{ik} \in s} w_{hik}(g_j)\ \boldsymbol{x}_{hik}\boldsymbol{x}'_{hik} \qquad (4.4.2.5)$$

$$\hat{\boldsymbol{X}}_{(g_j)} = \sum_{h_{ik} \in s} w_{hik}(g_j)\, \boldsymbol{x}_{hik} \qquad (4.4.2.6)$$

Si notamos $\hat{t}_{GREG}(g_j) = \sum_{hik \in s} w^*_{hik}(g_j)\, y_{hik} = \hat{t}_{(gj)} + \left(\boldsymbol{X} - \hat{\boldsymbol{X}}_{(g_j)}\right)^t \hat{\boldsymbol{\beta}}_{(g_j)}$ (4.4.2.7)

donde $\hat{\boldsymbol{\beta}}_{(g_j)} = \hat{\boldsymbol{A}}_{(gj)}^{-1}\, \hat{b}_{(g_j)}$ con $\hat{b}_{(g_j)} = \sum_{hik \in s} w_{hik}(g_j)\, \boldsymbol{x}_{hik}\, y_{hik}$ (4.4.2.8)

el estimador de varianza jackknife para $\hat{T}_{GREG}$ viene dado por:

$$V_J\left(\hat{T}_{GREG}\right) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} \left(\hat{t}_{GREG}(g_j) - \hat{T}_{GREG}\right)^2 \quad (4.4.2.9)$$

Yung y Rao(1996) demuestran que linealizando la varianza jackknife se obtiene que:

$$V_{JL}\left(\hat{T}_{GREG}\right) = V\left(e_{hi}^*\right) \text{ con } e_{hi}^* = \sum_k n_h w_{hik}^* e_{hik} \quad (4.4.2.10)$$

Además propusieron un estimador de varianza jackknife linearizado que retiene la inversa de la matriz utilizada para la estimación con el estimador GREG, calculada sobre *la muestra total* y establece factores de corrección para los pesos basados en ella, *sin necesidad de recalcularla cada vez que un cluster es eliminado.*

Si $w_{hik(gj)}$ son los pesos muestrales originales modificados por la eliminación del cluster j-ésimo cluster,se definen los siguientes factores de corrección:

$$\tilde{a}_{hik(gj)} = 1 + \left(\frac{w_{hik}}{w_{hik(gj)}}\right) * \boldsymbol{x}'_{hik} \hat{A}^{-1} (\boldsymbol{X} - \hat{\boldsymbol{X}}_{(gj)}) \quad (4.4.2.11)$$

Luego, Yung y Rao proponen los pesos resultantes $\tilde{w}_{hik(gj)}$ definidos como:

$$\tilde{w}_{hik(gj)} = w_{hik(gj)}\, \tilde{a}_{hik(gj)} \quad (4.4.2.12)$$

El estimador utilizando la muestra sin el j-ésimo cluster se calcula luego como

$$\tilde{t}_{GREG(gj)} = \sum_{hik \in s} \tilde{w}_{hik(gj)}\, y_{hik} \quad (4.4.2.13)$$

y la varianza jackknife resultante para un muestreo estratificado:

$$V_{JL}(\hat{T}_{GREG}) = \sum_{h=1}^{L} \frac{n_h - 1}{n_h} \sum_{j=1}^{n_h} (\tilde{t}_{GREG(gj)} - \hat{T}_{GREG})^2 \quad (4.4.2.14)$$

4.5 Consideraciones sobre los métodos

El estimador de varianza por linearización de Taylor es asintóticamente consistente para la varianza de estadísticos suaves, o sea aquellos que pueden expresarse como función de medias o totales. El método implica el cálculo de las derivadas del estimador de interés.

El estimador de varianza jackknife también es asintóticamente consistente par la varianza de estadísticas suaves pero es asintóticamente inconsistente para estadísticas no suaves como la mediana en el caso que se aplique eliminando un dato por vez al conjunto de datos que se tiene (Miller 1974). Sin embargo, empíricamente se ha mostrado que en un muestreo donde se elimina un conglomerado en cada iteración, esto no ocurre y el jackknife es adecuado de utilizar (Rao 1997).

Los últimos estudios parecen indicar al jackknife como el método más operativo por su facilidad de implementación dado los desarrollos computacionales existentes, para distintos tipos de diseños.

CAPÍTULO V
ESTUDIO DE SIMULACIÓN -DISEÑO

El estudio de simulación se llevó a cabo sobre una población real. El objetivo fue evaluar el comportamiento del estimador $\hat{T}_{Ycal}$ y de los estimadores de su varianza originados por la teoría asintótica clásica, por el jackknife y el jackknife linealizado, cuando se realiza un muestreo de conglomerados en dos etapas, sin reposición y donde la fracción de muestreo en la primera etapa no es pequeña, como sí es habitual suponer en las estimaciones de varianza en encuestas de este tipo.

La extracción de muestras, los cálculos de estimadores del total y de varianzas, y los cálculos de medidas resumen, fueron programados en S-plus 2000.

5.1 Descripción de la base de datos

El universo al que pertenecen los datos es una localidad de la provincia de Entre Ríos, donde se aplica la Encuesta Permanente de Hogares (EPH) del INDEC (Instituto Nacional de Estadística y Censos) para medir la desocupación en distintos dominios de interés y en donde se emplea el estimador de Horvitz&Thompson para las estimaciones. La base estuvo formada por 64000 individuos distribuidos en 22400 viviendas dentro de un total de 64 radios. Los radios estaban estratificados en 3 grupos. Los tamaños de los estratos fueron:

- Estrato 1: 26 radios
- Estrato 2: 30 radios
- Estrato 3: 8 radios.

Para cada individuo se contaba con la información de las variables Género, Edad y Estado de Ocupación Laboral.

La variable de interés fue el Total de Desocupados .

5.2 Descripción de la simulación

El trabajo consistió en simular un muestreo por conglomerados en dos etapas según los siguientes métodos:

- Primera Etapa: muestreo estratificado de los radios utilizando muestreo proporcional al tamaño (PPT) siguiendo el método de Sampford.

- Segunda Etapa: selección de hogares (SSU) dentro de los radios (PSU) por Muestreo Simple al Azar (MSA), y utilizando una fracción de muestreo no pequeña.

Descripción de la primera etapa:

Para cada replicación se seleccionaron 6, 9 y 3 radios de los estratos 1,2 y 3 respectivamente. Para ello, de acuerdo al diseño que se mencionó para esta etapa se desarrolló en S-Plus un programa que realizó los siguientes pasos:

1. Verifica que todos los radios cumplan la condición necesaria para aplicar el método: $Z_{hi}<1/n_h$; sino aborta.
2. Selecciona la muestra de tamaño n_h en cada estrato (h=1,2,3) utilizando el método de Sampford

Descripción de la segunda etapa:

Para cada grupo de radios seleccionados, se extrajo una muestra de hogares por MSA según las fracciones de muestreo 0.25; 0.15 y 0.25 para los radios seleccionados en la primera etapa dentro de los estratos 1, 2 y 3 respectivamente. En cada replicación, la muestra quedó finalmente integrada por el total de individuos en los hogares seleccionados.

El número total de replicaciones realizadas fue 10141.

5.3 Cálculo de estimadores

Para cada muestra obtenida se calculó para el Total de Desocupados, el estimador de Horvitz Thompson, el estimador de Regresión Generalizada y el estimador de Calibración mediante el método multiplicativo. Como variables marginales para el raking se consideraron el género y la edad por rangos. Para la estimación por calibración se desarrolló un programa ad-hoc en S -plus. En el cálculo de los pesos de calibración se utilizó la Inversa Generalizada, como se detalló en el Cap III 3.2.10.

Luego, en cada caso, se estimó la varianza del estimador correspondiente a un muestreo estratificado, utilizando:

a. Varianza de Taylor.
b. Varianza jackknife.

c. Varianza jackknife linealizada.

Para los estimadores jackknife y jackknife linealizados se ensayaron distintos coeficientes de ponderación en cada estrato con el objeto de corregir el efecto de la fracción de muestreo, ya que al no ser pequeña, no se asumió que el muestreo podía considerarse con reposición, a efectos de estimar la varianza.

Luego, además de calcular el estimador usual de varianza jackknife para muestreo estratificado 4.4.1.2, se ensayaron los siguientes estimadores:

$$V_{JC}(\hat{T}) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (1 - \pi_i)(\hat{T}_{(gj)} - \hat{T})^2$$

$$V_{JG}(\hat{T}) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} \left(\frac{n_h - \pi_i}{n_h - 1} - \pi_i \right)(\hat{T}_{(gj)} - \hat{T})^2$$

Los estimadores se compararon con los valores verdaderos del Total y de la varianza del estimador.

Como se vió en la sección 1.7.2, para el caso de un muestreo de conglomerados en dos etapas, la varianza del estimador puede descomponerse en dos sumandos, uno correspondiente a la varianza de la primer etapa y el otro correspondiente a la varianza de la segunda etapa. En la estimación de la varianza a partir de la simulación, se trabajó sólo con la primer componente, de acuerdo a lo mencionado hacia el final de la misma sección.

5.4 Cálculo de medidas de resumen

A fin de evaluar la performance de los estimadores se calcularon los siguientes indicadores luego de R=10141 simulaciones:

A. Sesgo Relativo Porcentual del Total de Desocupados estimado por Horvitz Thompson, GREG y raking, respecto del valor poblacional:

$$\frac{E(\hat{T}_Y) - T}{T} * 100 \quad ; \qquad (5.3.1)$$

siendo

$$E(\hat{T}_Y) = \frac{1}{R} \sum_{r=1}^{R} \hat{T}_{Yr}$$

B. Sesgo Relativo Porcentual del estimador de varianza asintótica , del estimador de varianza jackknife y del estimador jackknife linealizado, respecto de la varianza verdadera.

$$\frac{E\left(\hat{V}(\hat{T}_Y) - V_{TRUE}\right) - V_{TRUE}}{V_{TRUE}} * 100 \qquad (5.3.2)$$

siendo $E\left(\hat{V}(\hat{T}_Y)\right) = \frac{1}{R}\sum_{r=1}^{R} \hat{V}_r(\hat{T}_Y)$ y $V_{TRUE} = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{T}_{Yr} - E(\hat{T}_Y)\right)^2$

C. Coeficiente de variación porcentual de la varianza asintótica y de las varianzas jackknife y jackknife linealizada

$$\frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{V}_r(\hat{T}_Y) - V_{TRUE}\right)^2}}{V_{TRUE}} * 100 \qquad (5.3.3)$$

La extracción de muestras, los cálculos de estimadores del total y de varianzas, y los cálculos de medidas resumen, fueron programados en S-plus 2000.

CAPÍTULO VI
RESULTADOS Y CONCLUSIONES

El método de estimación de varianza a partir de la linealización por serie de Taylor ha sido históricamente el método comúnmente utilizado para la varianza de los estimadores de raking; sin embargo, a partir del desarrollo informático de los últimos tiempos, las desventajas que presenta este método han dado lugar a la aplicación y desarrollo de técnicas de replicación.

El método de Taylor, si bien es sencillo en cuanto a su concepto, requiere el desarrollo de una fórmula particular para cada estimador cuya varianza se quiera estimar. Para dicho cálculo son necesarias las derivadas parciales del estimador, lo cual deja afuera la posibilidad de cálculo por ejemplo, para los estadísticos de orden. Por otra parte también se requiere hallar las probabilidades de segundo orden, las cuales dependiendo del diseño que se esté utilizando, pueden resultar complicadas de desarrollar. Estas desventajas del estimador de linealización de Taylor se intensifican cuando el estimador cuya varianza se quiere hallar es un estimador no lineal con una expresión compleja. Por otra parte, en las grandes encuestas generalmente se requiere el cálculo de varios estimadores, no sólo de uno, por lo cual sería ventajoso contar con métodos de estimación de varianza más prácticos, que no requieran demasiado tiempo para su cálculo.

El jacknife y el bootstrap son dos de los métodos de replicación de los cuales más literatura se encuentra en cuanto a su aplicación en la estimación de varianza y su principal ventaja es que no requieren más que la expresión del mismo estimador para los cálculos.

En este trabajo se ha calculado el estimador de varianza del estimador de calibración bajo un diseño complejo como lo es el diseño de conglomerado en dos etapas sin reposición en la primera, lo cual desde el punto de vista del diseño implica una diferencia sobre la mayoría de los trabajos hallados para la estimación de varianza, ya que es común que se utilicen diseños más simples o bien se asuma que el muestreo es con reposición en la primera etapa.
Se ha estimado la varianza utilizando el estimador clásico de Taylor y el método Jacknife con tres variantes:

- jacknife recalibrando los pesos cada vez que un conglomerado se elimina (Rao 1996);
- jacknife linealizando como una alternativa a la recalibración luego de la eliminación de cada conglomerado
- jacknife sin recalibrar, lo cual es cómun que se aplique en la práctica, sin embargo mostraremos que no tiene buena performance;

Luego, los objetivos principales han sido:

I. Evaluar el comportamiento del jacknife en un diseño complejo.

II. Verificar que dicho método de estimación produce resultados similares al método de Taylor y que por lo tanto puede considerarse como un método alternativo cuando Taylor no pueda utilizarse o cuando la expresión del estimador cuya varianza se requiere sea complejo desde el punto de vista de sus derivadas o cuando el método de muestreo dificulte el cálculo de las probabilidades de segundo orden.

III. Comparar la performance de la estimación de varianza jacknife cuando no se vuelven a calibrar los pesos en cada eliminación de conglomerado, con la varianza jacknife en la cual sí se recalibra y mostrar que no genera resultados recomendables.

IV. Evaluar el comportamiento de la varianza jacknife linealizada como alternativa más sencilla o bien, que requiere menos tiempo de procesamiento.

6.1 RESULTADOS DE LAS SIMULACIONES

6.1.1 Estimadores del Total

A partir de las 10141 simulaciones se analizó en primer lugar la performance de los estimadores del Total de Desocupados, cuyo valor real es 3433, calculando el sesgo, el error estándar del sesgo y el sesgo relativo porcentual, para el estimador de Horvitz-Thompson, para el estimador GREG y para el estimador de Calibración:

Tabla I: Sesgo de los estimadores del Total

Estimador	Sesgo Medio	Stand error del sesgo	Sesgo Relat Porcent
HT	-3,71	294,21	-0,11
CAL	-2,70	290,51	-0,08
GREG	-2,68	290,52	-0,08

Los sesgos son similares para los tres estimadores, siendo el sesgo relativo porcentual cercano al 0.1%. Al calcular el valor absoluto del sesgo dividido por el error standard, se obtiene que el sesgo es no significativo en los tres casos, ya que es menor a 1.96. Estos resultados son esperables debido a que el estimador de Horvitz Thompson es insesgado y el estimador de calibración lo es en forma asintótica, como se mostró en el Cap III.

6.1.1 Estimadores de la varianza del estimador de calibración

Las varianzas "verdaderas" (V_{TRUE}, calculadas como la varianza de los 10141 estimadores), obtenidas para el estimador Greg y el estimador de calibración fueron muy similares confirmando la propiedad de que son asintóticamente equivalentes. En la Tabla II se observan los estimadores de varianza medios obtenidos para el estimador de calibración, comparando con su varianza "verdadera":

Tabla II: Media de estimadores de Varianza

Var $_{TRUE}$ para estimador de Calibración = **84388,9**

Estimador de varianza	Media
Taylor	81738,2
Jack recalibrando y sin ajustar	95179,3
Jack recalibrando y ajustando con hi	67148,9
Jack recalibrando y ajustando con li	78877,7
Jack linealizado	94431,8
Jack linealizado y ajustando con hi	66520,4
Jack linealizado y ajustando con li	78137,9
Jack sin recalibrar	120026,8

Tal como fue definido en el Cap V, se calculó el Sesgo Relativo Porcentual y el Coeficiente de variación Porcentual para los siguientes estimadores de varianza, a fin de evaluar su performance:

a. con linealización de Taylor
b. con el método jacknife recalibrando luego de la eliminación de cada conglomerado y sin factores de corrección
c. con el método jacknife recalibrando luego de la eliminación de cada conglomerado y a su vez corrigiendo por los factores definidos en el CapV:

$h_i = (1-\pi_i)$ y $l_i = \left(\frac{n_h - \pi_i}{n_h - 1} - \pi_i \frac{+}{+}\right)$.

d. con método jacknife linealizado
e. con jacknife linealizado y corrigiendo con h_i y l_i.
f. con el método jacknife pero sin recalibrar.

Las varianzas verdaderas obtenidas para el estimador Greg y el estimador de calibración fueron muy similares, lo cual era esperable dado que son asintóticamente equivalentes.

Tabla III: Sesgos y Coeficiente de Variación luego de 10141 simulaciones

Estimador de varianza	Sesgo Porcentual Relativo	Coeficiente de Variación Porcentual
Taylor	-3,14	31,54
Jack recalibrando y sin ajustar	12,79	44,31
Jack recalibrando y ajustando con hi	-20,43	37,97
Jack recalibrando y ajustando con li	-6,53	38,12
Jack linealizado	11,90	43,56
Jack linealizado y ajustando con hi	-21,17	37,94
Jack linealizado y ajustando con li	-7,41	37,66
Jack sin recalibrar	42,23	90,85

El estimador jacknife y jacknife linealizado sobreestimaron la varianza, mientras que al introducir los coeficientes de corrección el signo del sesgo se invierte. Con la varianza jacknife recalibrando y jacknife linealizado se obtienen performances similares en cuanto al sesgo (12.8 y 11.9) y el coeficiente de variación (44.3 y 43.6). Lo mismo se observa al comparar los resultados de utilizar los coeficientes de corrección: con los coeficientes h_i el sesgo relativo porcentual para el jacknife recalibrando resultó -20.4 y para el jacknife linealizado -21.2, mientras que los coeficientes de variación fueron cercanos a 38% en ambos casos. Con los coeficientes l_i el sesgo relativo porcentual fue menor para ambos estimadores jacknife resultando -6.5, -7.4 respectivamente, y los coeficientes de variación rondaron también el 38%. El estimador de varianza con medidas más cercanas más cercano al estimador de Taylor resultó ser el jacknife con los coeficientes de corrección li.

Resulta evidente que si se utiliza el estimador jacknife sin introducir la re-calibración en cada paso de eliminación de un conglomerado, los resultados son desalentadores, por lo cual debe desaconsejarse su uso: en la tabla II el valor medio de la varianza con este estimador es mucho mayor que para los otros; el sesgo relativo porcentual registrado en la Tabla III es de orden mucho mayor que los demás, como así también el coeficiente de variación que llega a ser cercano al 91%.

6.2 CONCLUSIONES

De acuerdo a los objetivos planteados en este trabajo se puede concluir:

I. *El comportamiento del método de estimación jacknife en este diseño complejo resultó aceptable, siempre que se tenga en cuenta la re-calibración o la linealización.*

II. *Los resultados obtenidos con el jacknife utilizando factores de corrección por fracción de muestreo no pequeña fueron cercanos a los correspondientes con el uso del método de Taylor.*

III. *Los resultados del jacknife si no tiene en cuenta la recalibración no son adecuados.*

IV. *La varianza jacknife linealizada resultó una alternativa más sencilla, que requiere menos tiempo de procesamiento.*

Conclusión final

El estimador de varianza de Taylor es considerado tradicionalmente un estimador adecuado, por lo cual es ampliamente utilizado; sin embargo encuentra su limitación en estimadores de expresión compleja, al momento del cálculo de las derivadas del mismo o de las probabilidades de segundo orden; más aún, no resulta aplicable para estimadores no derivables, como por ejemplo los estadísticos de orden (mediana, percentiles). Es necesario entonces, algún estimador más general y que no requiera excesivo tiempo para su cálculo. Surgen así como posibles alternativas los estimadores jacknife y jacknife linealizados.

Este trabajo muestra la validez de la aplicación del método jacknife como alternativa al método tradicional de linealización por serie de Taylor para la estimación de varianza del estimador de calibración bajo un diseño de muestreo complejo, en una encuesta real nacional, para la cual se verifica a través del cálculo de medidas apropiadas, que los estimadores jacknife muestran un buen comportamiento al utilizarse para estimar la varianza del estimador de un Total.

Los resultados sugieren que la varianza estimada a partir del método jacknife linealizado es muy similar a la estimada mediante el método jacknife recalibrando los pesos luego de cada eliminación de conglomerado, por lo cual puede considerarse este método como adecuado para la estimación, dado que a su vez resulta práctico por dos razones: su aplicación se generaliza a todo tipo de estimador, sin importar la complejidad de su expresión, siendo esto común a todo método jacknife; es el que tiene asociado menor costo computacional.

Se ha mostrado aquí a su vez que aplicando coeficientes de corrección por el uso de un muestreo con fracción no pequeña se obtiene un estimador mejorado al comparar con el jacknife habitual en el que se asume una fracción de muestreo despreciable.

Finalmente se ha dejado en evidencia que el sesgo introducido en el estimador de varianza si en el método jacknife se obvia el recalibrado cada vez que se elimina un conglomerado, es grande; por lo tanto se desaconseja su uso.

El cálculo del estimador de varianza a través del método jacknife no ha sido muy difundido en nuestro país. Un objetivo complementario de este trabajo ha sido contribuir a la luz de los resultados, a incentivar su uso en las grandes encuestas de la Argentina, como práctica habitual para la estimación de la varianza, siendo que en otros países del mundo estas técnicas ya están siendo utilizadas.

APÉNDICE

(1) Lema 2.3.1(Sampford, 1967)

Definiendo $g(n,r,k)= \sum_{s \in S_{n-r}(U \setminus k)} \left(\prod_{l \in s} w_l \right) \left(n - r\pi_k - \sum_{i \in s} \pi_i \right)$ (2.3.5)

donde

$S_{n-r}(U \setminus k) = \{ s \in S_{n-r} \,/\, I_k = 0 \}$ o sea, el conjunto de muestras de tamaño $n-r$ que no incluyen a y_k.

Entonces:

$$g(n,r,k)=(1-\pi_k) \sum_{t=1}^{n} t D_{n-t} \qquad (2.3.6)$$

Dem:

Sea $\#(s)=n-r$

$$n - r\pi_k - \sum_{i \in s} \pi_i = n + r - r - r\pi_k - \sum_{i \in s} \pi_i = r(1-\pi_k) + \sum_{i \in s} (1-\pi_i) \qquad (2.3.7)$$

Luego,

$$g(n,r,k)= \sum_{s \in S_{n-r}(U \setminus k)} \left(\prod_{l \in s} w_l \right) \left(n - r\pi_k - \sum_{i \in s} \pi_i \right) =$$

$$= \sum_{s \in S_{n-r}(U \setminus k)} \left(\prod_{l \in s} w_l \right) \left(r(1-\pi_k) + \sum_{i \in s} (1-\pi_i) \right) =$$

$$= r(1-\pi_k) D_{n-r} + \sum_{s \in S_{n-r}(U \setminus k)} \left(\prod_{l \in s} w_l \right) \left(\sum_{i \in s} (1-\pi_i) \right)$$

$$\Rightarrow g(n,r,k) = [r(1-\pi_k)] D_{n-r}(\bar{k}) + h(n,r,k) \qquad (2.3.8)$$

Siendo $h(n,r,k)= \sum_{s \in S_{n-r}(U \setminus k)} \left(\prod_{l \in s} w_l \right) \left(\sum_{i \in s} (1-\pi_i) \right)$ (2.3.9)

y $\qquad D_z(\bar{k}) = \sum_{s \in S_z(U \setminus k)} \left(\prod_{l \in s} w_l \right)$

Se debe notar que $1-\pi_i = \frac{\pi_i}{w_i}$ y por lo tanto, si se reemplaza en (2.3.9) se obtiene:

$$h(n,r,k)= \sum_{s\in S_{n-r}(U\setminus k)} \left(\prod_{l\in s} w_l\right)\left(\sum_{i\in s}\frac{\pi_i}{w_i}\right) = \sum_{s\in S_{n-r}(U\setminus k)} \left(\prod_{l\in s} w_l\right)\left(\sum_{\substack{i=1\\ i\neq k}}^{N}\pi_i(1-I_i)\right) =$$

$$= \sum_{s\in S_{n-r}(U\setminus k)} \left(\prod_{l\in s} w_l\right)\left(n-\pi_k-\sum_{\substack{i\in s\\ i\neq k}}\pi_i\right) = g(n,r+1,k)+r\pi_k D_{n-r-1}(\bar{k}) \qquad (2.3.10)$$

Dado que $D_m(\bar{k}) = D_m - w_k D_{m-1}(\bar{k})$, utilizando (2.3.8) y (2.3.10) se obtiene la relación recursiva:

$$g(n,r,k) = r(1-\pi_k)D_{n-r} + g(n,r+1,k), \qquad (2.3.11)$$

con la condición inicial $g(n,r,k) = (1-\pi_k)$ (c.q.d).

(2) <u>Resultado1</u>: La ecuación 3.2.1.7 tiene, con probabilidad 1, una única solución perteneciente a un dominio C, cuando $n\to\infty$ (C es un convexo abierto que incluye al 0 para todo n)

Dem: sea $z= N^{-1}(\hat{\boldsymbol{t}}_{x\pi} - \boldsymbol{t}_x)$, entonces, con probabilidad 1, z pertenece a una esfera de radio r contenida en C , que puede llamarse B_1. Por otra parte, $N^{-1}\phi_s(\boldsymbol{\lambda})$ tiene función inversa en B_1 con probabilidad tendiendo a 1. Como (3.2.1.7) puede escribirse como $N^{-1}(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{x\pi}) = N^{-1}\phi_s(\boldsymbol{\lambda})$, la ecuación tiene una única solución con probabilidad tendiendo a 1.

<u>*Resultado 2*</u>

Si $\boldsymbol{\lambda}$, solución de 3.2.1.7, existe, entonces tiende a 0 en probabilidad y además $\boldsymbol{\lambda} = O_p(n^{-1/2})$

Dem: Sea $\boldsymbol{\lambda} = (N^{-1}\phi_s')^{-1}(\boldsymbol{z})$ si z pertenece a B_1, sino $\boldsymbol{\lambda}$ es definido arbitrariamente. Como $\phi_s(\boldsymbol{0}) = \boldsymbol{0}$, entonces

$\boldsymbol{\lambda}$ **-0=** $(N^{-1}\phi_s')^{-1}(\boldsymbol{z})-(N^{-1}\phi_s')^{-1}(\boldsymbol{0})$ entonces, usando la propiedad f. se obtiene que $\|\boldsymbol{\lambda}\| \leq \|z\|K(1-\beta)^{-1}$. Como $z = O_p(n^{-1/2})$, entonces existe una constante K' tal que $P(\|z\| \leq K'n^{-1/2}) \to 1$.

Luego, combinando ambas desigualdades,

$P(\|\boldsymbol{\lambda}\| \leq KK'(1-\beta)^{-1}n^{-1/2}) \to 1$, lo cual implica que $\boldsymbol{\lambda} = O_p(n^{-1/2})$.

(3) Definición Sea X_n una sucesión de variables aleatorias, se dice que X_n es de orden inferior en probabilidad a $h_n>0$ si

$$P(\lim_{n->\infty} \frac{X_n}{h_n}) = 0$$

y se indica $X_n = o_p(h_n)$

Definición Sea X_n una sucesión de variables aleatorias, X_n se dice que está acotada por $h_n>0$ si $\forall \varepsilon > 0$, existe un número $M_\varepsilon>0$ tal que

$$P(|X_n| \geq M_\varepsilon h_n) \leq \varepsilon$$

y se anota

$$X_n = O_p(h_n) \qquad \forall\, n=1,2,3...$$

Teorema Si $X_1,...,X_n$ es una sucesión de variables aleatorias tales que $X_k = x_0 + O_p(h_n)$; y sea f(x) una función derivable m veces con derivadas continuas en el punto x_0, y h_n una sucesión de números positivos tales que $\lim_{n->\infty} h_n = 0$, entonces

$$f(X_n) = f(x_0) + \sum_{i=1}^{m-1} (X_n - x_o)^i \frac{f^{(i)}(x_0)}{i!} + O_p(h^m{}_n) \qquad (4.1.1)$$

donde las $f^{(i)}$ representa la i-ésima derivada de la función f

Dem:

Desarrollando f en Serie de Taylor se tiene

$$f(X_n) = f(x_0) + \sum_{i=1}^{m-1} (X_n - x_o)^i \frac{f^{(i)}(x_0)}{i!} + (X_n - x_0)^m \frac{f^{(m)}(c)}{m!}$$

donde c varía entre X_n y x_0. Como $f^{(m)}$es contínua, está acotada en probabilidad, es decir que $f^{(m)}(c) = O_p(1)$ y por lo tanto el último término del desarrollo anterior es una $O_p(h^m_n)$

Teorema *Sean $X_{in},\ldots, X_{jn},\ldots, X_{pn}$ p sucesiones de variables aleatorias tales que $X_{jn}=x_{j0}+O_p(h_n)$, $j=1,\ldots,p, f(x_1,\ldots,x_p)$ una función continua cuyas derivadas parciales existen y son continuas en los puntos x_{j0}, y h_n una sucesión de números positivos tales que $\lim_{n->\infty} h_n = 0$, entonces*

$$f(X_{1n},\ldots,X_{pn}) = f(x_{10},\ldots,x_{p0}) + \sum_{j=1}^{p} (X_{jn} - x_{j0}) \left.\frac{\partial f(x_1,\ldots,x_p)}{\partial x_j}\right|_{x_j = x_{j0}} + O_p(h_n^2)$$

Dem:

Aplicando un desarrollo en Serie de Taylor hasta el segundo orden y de acuerdo a 4.1.1 se obtiene:

$$f(X_{1n},\ldots,X_{pn}) = f(x_{10},\ldots,x_{p0}) + \sum_{j=1}^{p} (X_{jn} - x_{j0}) \left.\frac{\partial f(x_1,\ldots,x_p)}{\partial x_j}\right|_{x_j = x_{j0}} +$$

$$+ \sum_{j=1}^{p}\sum_{i=1}^{p} (X_{jn} - x_{j0})(X_{in} - x_{i0}) \left.\frac{\partial^2 f(x_1,\ldots,x_p)}{\partial x_j \partial x_i}\right|_{x_j = b_j}$$

donde bj son puntos intermedios entre X_{jn} y x_{j0}. Con el mismo argumento utilizado en el teorema se puede asegurar que $(X_{jn} - x_{j0})(X_{in} y x_{i0}) = O_p(h_n^2)$ que multiplica a una cantidad acotada en probabilidad.

BIBLIOGRAFÍA

1. Canty and Davison (1998), *Resampling -based variance estimation for labour force surveys.* The Statistician.
2. Cochran, William(1998) *Técnicas de Muestreo*, CECSA.
3. Deville and Särndal (1992) *Calibration estimators in Survey Sampling*, Journal of the American Statistical Association, Vol 87.
4. Deville, Sändal and Sautory (1993) *Generalized raking procedures in survey sampling*, Journal of the American Satistical Association, Vol 88.
5. Deville, Särndal, Sautory (1993) *Generalized Raking Procedures in Survey Sampling*, American Statistical Association, Journal of the American Statistical Association, Vol 88, Nº 423
6. Efron and Stein (1981)*The jackknife estimate of variance,* The Annals of Statistics, Vol 9.
7. Graybill, F.A (1983). *Matrices with Applications in Statistics*, Belmont
8. Haziza, Mecatti, Rao (2004), *Comparison of variance estimators under Rao-Sampford method, a simulation study*, ASA Survey Research Methods, pag 3638-3643.
9. Miller, R.G (1974), *The jacknife-a review Biometrika* 61 1-15
10. Rao (1994) *Resampling methods for complex surveys*, Journal of the American Satistical Association. Vol 5.
11. Rao , Shao (1997) *Modified balanced repeated replication for complex survey data* . Department of Mathematics and Statistics, Carleton University.
12. Renssen Robbert, Matinus Gerard (2002) *On the Use of Generalized Inverse Matrices in Sampling Theory*, Survey Methodology, Vol 28 Statistics Canada.
13. Sampford, M.R (1967) *On sampling without replacement with unequal probabilities of selection*, Biometrika, Vol 54.
14. Shao and Tu (1995) *The Jackknife and Bootstrap*, Springer.
15. Shao and Wu (1989) *A general theory for jackknife variance estimation,* The Annals of Statistics, Vol 17.

16. Stuckel, Hidiroglou and Särndal (1996) *Variance Estimation for calibration estimators: a comparison of jackknifing versus taylor linearization,* Survey Methodology, Vol 22 Statistics Canada.

*17.*Tillé, Yves (1996) *Sampling Algorithms, Springer.*

18.Tillé, Yves(2006). Curso sobre *Teoría de Muestreo. JIE 2006*

19.Yung and Rao (1996) *Jackknife linearization variance estimators under stratified multi-stage sampling,* Survey Methodology, Vol22 Statistics Canada.

20. Yves Tillé (1996) *Some remarks on unequal probability sampling designs without replacement*, Annales D'Economie et de Statistique, Nº44.

21. Zhang (1998) *Post –Stratification and Calibration*, Statistics Norway nº 216.