METODOS DE ESTIMACIÓN DE LA VARIANZA, CON APLICACIÓN A LA ENCUESTA NACIONAL DE HOGARES AMPLIADA

Fiorella Cavalleri Ferrari

Diciembre de 2008

Índice general

Agradecimientos

A mi compañero de vida, y a mi hija Magdalena por su comprensión durante el tiempo que le dediqué a este trabajo de pasantía.

A mis padres y hermanos por apoyarme siempre

A mi tutor de pasantía, Dr. Juan José Goyeneche por su invalorable contribución a mi formación.

Al Ec. Guillermo Zoppolo, por sus aportes.

A la Lic. Mª Eugenia Riaño, por su colaboración con el uso del software R

Capítulo 1

Introducción

La teoría y las aplicaciones del muestreo han aumentado de forma espectacular en los últimos años. Cientos de estudios se realizan cada año en el sector privado, en la comunidad académica, y en los diversos organismos gubernamentales. Ejemplo de ello son los estudios de mercado y encuestas de opinión pública; estudios relacionados con la investigación académica, grandes encuestas nacionales acerca de la participación de la fuerza laboral, la atención de salud, uso de energía, y la actividad económica. Los estudios a partir de muestras afectan a casi todos los campos de estudio científico, incluidas la agricultura, la demografía, la educación, la energía, el transporte, la atención de salud, la economía, la política, la sociología, etc. A medida que la utilización de las encuestas por muestreo se ha incrementado, también lo ha hecho la necesidad de métodos de análisis e interpretación de los datos resultantes. Una de las principales exigencias de una buena encuesta por muestreo es el cálculo del error de muestreo.

Existen muestreos no probabilísticos que si bien son utilizados en la práctica, carecen de la posibilidad de una estimación válida de los márgenes de error.

Wolter [Wolter, 1985] se centra específicamente en la metodología de es-

timación de la varianza en encuestas con diseños complejos. Wolter dice que si bien la terminología de "Encuestas con diseños complejos" nunca ha sido rigurosamente definida, encuentra lo que considera dimensiones importantes de una encuesta por muestreo con diseño complejo, a saber:

1. El grado de complejidad del diseño para obtener las muestras.
2. El grado de complejidad de los estimadores.
3. La multiplicidad de características o variables de interés.
4. El uso de los datos de la encuesta con fines analíticos y/o descriptivos.
5. La escala o tamaño de la encuesta.

Examinando las primeras dos dimensiones, gran parte de la teoría básica de encuestas por muestreo refiere a diseños simples y estimadores lineales, en tanto que las técnicas modernas utilizan diseños complejos con estimadores lineales, diseños simples con estimadores no lineales y diseños complejos con estimadores no lineales. El estudio implica a menudo diseños tales como estratificación, varias etapas de muestreo, probabilidades de selección diferentes, doble toma de muestras, marcos múltiples, así como ajustes por no respuesta, sobrecobertura o subcobertura, observaciones atípicas, procedimientos de post-estratificación, etc. En cuanto a la dimensión tres, las encuestas por muestreo con diseños complejos implican decenas o cientos de características de interés, esto puede ser contrastado con la teoría básica sobre muestreo que se discute en libros de texto donde sólo una característica o variable se considera de interés. La dimensión cuatro toma la idea de que mientras que las encuestas por muestreo complejo incluyen a la vez objetivos de descripción y análisis, en una encuesta sencilla el objetivo puede equivaler a la descripción de diversas características de la población objetivo como lo puede ser el número de hombres y mujeres que votan por un determinado candidato en una

elección política. Finalmente en lo referente a la dimensión cinco, que alude a la escala de la encuesta, generalmente los estudios complejos involucran gran número de encuestas y por tanto de datos, generándose grandes volúmenes de información que lleva a la necesidad de contar con herramientas informáticas adecuadas para procesamientos de gran magnitud.

En el contexto de las técnicas de muestreo complejas ¿Como se elige un estimador de la varianza? La elección es generalmente difícil. ¿En que priorizar? ¿En la precisión del estimador de la varianza? ¿En la oportunidad, costo, sencillez, o en otras consideraciones administrativas? Si bien cuestiones acerca de la precisión de los estimadores deben tener gran peso sobre la decisión respecto de las distintas opciones, consideraciones de tipo administrativas como los costos, el plazo, etc., también deben contemplarse. En estas circunstancias puede ocurrir que los métodos de estimación que son rentables, sean convenientes, aunque pueda implicar cierta pérdida de precisión. Los métodos de estimación de la varianza deberán ser evaluados a la luz de las consideraciones antes mencionadas. Otro aspecto a considerar es que la mayoría de las encuestas por muestreo complejo, son de carácter polivalente, lo que significa que hay muchas variables y estadísticos de interés, cada una de las cuales requiere una estimación de su correspondiente varianza. La utilización de distintos estimadores puede ser viable en ciertos entornos de estudio cuando el presupuesto, tiempo, recursos profesionales e informáticos son abundantes. En muchos entornos de estudio, sin embargo estos recursos son escasos, y se podrá usar uno o como mucho dos, métodos de estimación de la varianza.

La precisión de un estimador $\widehat{\theta}$, se discute generalmente en términos de la varianza del mismo, es decir $V(\widehat{\theta})$. El valor exacto de la varianza es generalmente desconocido, porque depende de cantidades desconocidas de la población. Después de obtenida la muestra, sin embargo puede ser calculada una estimación de la varianza. Cuando se presentan los resultados es bueno

presentar estimaciones de la varianza, $\widehat{V}(\widehat{\theta})$, de los estimadores. La estimación de la varianza se usa frecuentemente en la construcción de intervalos de confianza, si se asume que la distribución en el muestreo de $\widehat{\theta}$ es aproximadamente normal entonces, el intervalo de confianza de nivel $1-\alpha$ para θ tendrá la siguiente forma:

$$\widehat{\theta} \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\widehat{V}(\widehat{\theta})}$$

El trabajo se estructurará de la siguiente forma: En el capítulo 2 se hará una revisión teórica de los distintas técnicas de estimación de la varianza. Posteriormente el capítulo 3 comenzará exponiendo los principales conceptos y definiciones del censo y la ENHA para luego realizar un ejemplo de simulación Monte Carlo aplicado al censo para el departamento de Colonia , que permitirá comparar las técnicas de estimación de varianza con los resultados poblacionales. Al final del capítulo 3 se realizará una aplicación de distintos estimadores de la varianza sobre la ENHA. Finalmente, en el capitulo 4 se presentarán conclusiones.

Capítulo 2

Marco teórico

2.1. Notación

Una población finita es un conjunto finito de elementos que se denota por $U = \{1, 2, \ldots, N\}$. El número de elementos en la población es N. Una muestra, s, es un subconjunto de la población U: $s \subset U$. El conjunto de todas las muestras posibles se denota por $\mathbf{S} = \{s_1, s_1, \ldots s_M\}$.

Una muestra probabilística es aquella donde se conoce el conjunto de todas las muestras posibles $\mathbf{S}$ y todo $s \in \mathbf{S}$ tiene asociada una probabilidad de selección, $p(s)$, tal que, cada $k \in U$ tiene una probabilidad no nula de inclusión igual a π_k. La probabilidad de inclusión $\pi_k = P(k \in S) > 0 \, \forall k \in U$.

La probabilidad de inclusión de segundo orden para un par de elementos k y $l \in U$, π_{kl}, es

$$\pi_{kl} = P(k \, \mathrm{y} \, l \in S)$$

La variable que interesa medir, y, asume un valor para cada uno de los elementos de la población, siendo y_k la notación genérica, en tanto que:

$$\check{y}_k = \frac{y_k}{\pi_k}$$

$$\Delta_{kl} = Cov(I_k, I_l) = \pi_{kl} - \pi_k \pi_l$$

y

$$\check{\Delta}_{kl} = \frac{\Delta_{kl}}{\pi_{kl}},$$

Siendo I_k e I_l variables indicatrices que valen 1 si el elemento k o l pertenecen a la muestra respectivamente.

El llamado estimador π, $\widehat{t}_\pi$ de un total poblacional fue introducido por Horvitz-Thompson en 1952. Para estimar un total

$$t_y = \sum_U y_k$$

se usa

$$\widehat{t}_\pi = \sum_s \frac{y_k}{\pi_k} = \sum_s \check{y}_k,$$

o sea, para obtener un estimador insesgado de un total poblacional, se pasa de U a s y se "π-expande cada sumando" (multiplicando por el inverso de la probabilidad de inclusión).

Para la construcción de intervalos de confianza se necesita estimar $V(\widehat{\theta})$. Para un estimador Horvitz-Thompson de un total se tiene:

$$V(\widehat{t}_\pi) = \sum\sum_U \Delta_{kl}\check{y}_k\check{y}_l,$$

y la misma se estima insesgadamente por,

$$\widehat{V}(\widehat{t}_\pi) = \sum\sum_s \check{\Delta}_{kl}\check{y}_k\check{y}_l.$$

2.2. Muestra particionada en subgrupos

En varias de las técnicas que se verán a continuación habrá que considerar un número (dígase A) de subgrupos de la muestra original, y un estimador de θ calculado a partir de cada subconjunto.

1. Sea $\widehat{\theta}$ un estimador de θ basado en una muestra probabilística s según un diseño $p(s)$.

2. Sean $s_1, s_2, \ldots, s_a, \ldots, s_A$, A subconjuntos de s. Se consideran:

$$\widehat{\theta}_1, \widehat{\theta}_2 \ldots, \widehat{\theta}_a, \ldots, \widehat{\theta}_A,$$

los A estimadores alternativos de θ basados en los subconjuntos

$$s_1, s_2, \ldots, s_a, \ldots, s_A.$$

Sea

$$\widehat{\theta}^* = \frac{1}{A}\sum_{a=1}^{A}\widehat{\theta}_a$$

el promedio de los A estimadores alternativos de θ.

Para estimar $V(\widehat{\theta}^*)$ se proponen:

$$\widehat{V}_1 = \frac{1}{A(A-1)}\sum_{a=1}^{A}(\widehat{\theta}_a - \widehat{\theta}^*)^2$$

y

$$\widehat{V}_2 = \frac{1}{A(A-1)}\sum_{a=1}^{A}(\widehat{\theta}_a - \widehat{\theta})^2$$

Como

$$\sum_{a=1}^{A}(\widehat{\theta}_a - \widehat{\theta}^*)^2 = \sum_{a=1}^{A}(\widehat{\theta}_a - \widehat{\theta})^2 + A((\widehat{\theta}^* - \widehat{\theta})^2,$$

$$\Rightarrow A(A-1)\widehat{V}_2 = A(A-1)\widehat{V}_1 + A(\widehat{\theta}^* - \widehat{\theta})^2,$$

y por lo tanto, $\widehat{V}_2 \geq \widehat{V}_1$.

De hecho ambas, $\widehat{V}_1$ y $\widehat{V}_2$ se utilizan para estimar la varianza de $\widehat{\theta}$, $V(\widehat{\theta})$, bajo el supuesto de que $V(\widehat{\theta}^*)$ y $V(\widehat{\theta})$ son aproximadamente iguales.

El uso de $\widehat{V}_1$ como un estimador de $V(\widehat{\theta}^*)$ se justifica por el siguiente resultado:

$$E(\widehat{V}_1) = V(\widehat{\theta}^*) - \frac{1}{A(A-1)}\sum_{a=1}^{A}\sum_{b=1}^{A}\mathrm{cov}(\widehat{\theta}_a, \widehat{\theta}_b) + \frac{1}{A(A-1)}\sum_{a=1}^{A}[E(\widehat{\theta}_a) - E(\widehat{\theta}^*)]^2$$

Entonces, si: $\text{cov}(\widehat{\theta}_a, \widehat{\theta}_b) = 0$ para todo $a \neq b$ y $E(\widehat{\theta}_a) = \theta$ para todo $a = 1, 2, \ldots A$, $\widehat{V}_1$ estimará insesgadamente la $V(\widehat{\theta^*})$, esto es, $E(\widehat{V}_1) = V(\widehat{\theta^*})$.

En el caso de que, $\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_A$ estén correlacionados, $\widehat{V}_1$ y $\widehat{V}_2$ estiman sesgadamente a $V(\widehat{\theta^*})$ y $V(\widehat{\theta})$

Los resultados precedentemente planteados fueron presentados por Durbin en 1953 y reexaminados por Wolter en 1985.

2.3. Un estimador simplificado de la varianza

Considérese el siguiente estimador de la varianza:

$$\widehat{V}_0 = \frac{1}{n(n-1)} \sum_s (\frac{y_k}{p_k} - \widehat{t}_\pi)^2$$

con: $p_k = \frac{\pi_k}{n} \, \forall \, k \in s$.

$\widehat{V}_0$ tiene la forma del estimador de la varianza que se aplica para el estimador *pwr* del diseño muestreo aleatorio simple con remplazo y probabilidad proporcional al tamaño. Ver Särndal et al. páginas 97 y 98 [Särndal et al., 1992].

Bajo un diseño cualquiera, con o sin reposición, este estimador de la varianza será un buen estimador en la medida que se trabaje con poblaciones grandes de tal forma que considerar un muestreo sin remplazo sea "similar" a considerar un muestreo con remplazo, es decir que la probabilidad de que una unidad sea seleccionada en más de una oportunidad es pequeña. Otro factor que tendrá incidencia en que $\widehat{V}_0$ sea un buen estimador es la fracción de muestreo, en la medida que esta sea pequeña, $\widehat{V}_0$ será un buen estimador.

2.4. Estimador de Razón

Se planteará el problema que refiere a cuando el parámetro que interesa estudiar es un cociente entre dos totales poblacionales:

$$R = \frac{t_y}{t_z} = \frac{\overline{y}_U}{\overline{z}_U}.$$

Un ejemplo podría ser, como se planteará en el capítulo siguiente, la edad promedio de una determinada población, que es el cociente entre el total de edad y el total de personas de la población.

Un estimador para el parámetro en cuestión viene dado por el cociente de los estimadores de los totales:

$$\widehat{R} = \frac{\widehat{t}_{y\pi}}{\widehat{t}_{z\pi}} = \frac{\overline{y}_s}{\overline{z}_s}$$

De acuerdo al resultado 5.5.1 de [Särndal et al., 1992], puede obtenerse la varianza aproximada de $\widehat{R}$, $AV(\widehat{R})$ y un estimador aproximadamente insesgado de ésta, $\widehat{V}(\widehat{R})$.

Notación:

$$AV(\widehat{R}) = V(\widehat{R}_0) = \frac{1}{t_z^2} \sum \sum_U \Delta_{kl} \frac{E_k}{\pi_k} \frac{E_l}{\pi_l}$$

y

$$\widehat{V}(\widehat{R}) = \frac{1}{t_z^2} \sum \sum_s \check{\Delta}_{kl} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l},$$

donde $E_k = y_k - Rz_k$ y $e_k = y_k - \widehat{R}z_k$.

Expresiones alternativas para $AV(\widehat{R})$ y $\widehat{V}(\widehat{R})$, considerando que,

$$\widehat{R} \doteq \widehat{R}_0 = R + \frac{1}{t_z}(\widehat{t}_{y_\pi} - R\widehat{t}_{z_\pi}),$$

luego:

$$AV(\widehat{R}) = \frac{1}{t_z^2}[V(\widehat{t}_{y_\pi}) + R^2 V(\widehat{t}_{z_\pi}) - 2 * R * \mathrm{cov}(\widehat{t}_{y_\pi}, \widehat{t}_{z_\pi})]$$

$$\widehat{V}(\widehat{R}) = \frac{1}{\widehat{t}_z^2}[\widehat{V}(\widehat{t}_{y_\pi}) + \widehat{R}^2 \widehat{V}(\widehat{t}_{z_\pi}) - 2 * \widehat{R} * \widehat{\mathrm{cov}}(\widehat{t}_{y_\pi}, \widehat{t}_{z_\pi})]$$

2.5. Resultados para varios diseños

2.5.1. Muestreo aleatorio simple

Se verifican las siguientes identidades:

$$\begin{aligned} V(\widehat{t}_\pi) &= N^2(1-f)\frac{S^2_{yU}}{n} \\ \widehat{V}(\widehat{t}_\pi) &= N^2(1-f)\frac{S^2_{ys}}{n} \\ \widehat{V}_0 &= N^2\frac{S^2_{ys}}{n} \end{aligned}$$

Pudiéndose probar que bajo un diseño $p(s)$, arbitrario, de tamaño fijo, se verifica:

$$E(\widehat{V}_0) - V = \frac{n}{n-1}(V_0 - V)$$

Donde:

$$V = -\frac{1}{2}\sum\sum_U \Delta_{kl}(\check{y}_k - \check{y}_l)^2$$

Siendo V_0 la varianza del estimador $(\widehat{t}_{pwr})$ y $p_k = \frac{\pi_k}{n}$, $k = 1, 2, \ldots, N$.

El sesgo de $\widehat{V}_0$ es positivo en los casos en que el muestreo sin reposición (n elementos, probabilidad de inclusión π_k, y estimador π), sea más eficiente que el muestreo con reposición (n elementos, probabilidad de inclusión $p_k = \frac{\pi_k}{n}$ y estimador pwr)

2.5.2. Muestreo estratificado

El estimador simplificado de la varianza:

$$\widehat{V}_0 = \frac{1}{n(n-1)}\sum_s (\frac{y_k}{p_k} - \widehat{t}_\pi)^2$$

deberá aplicarse en cada uno de los estratos por separado.

Así pues, como:

$$\widehat{t}_\pi = \sum_{h=1}^{H} \widehat{t}_{\pi h},$$

se define el estimador simplificado de la varianza de la siguiente forma:

$$\widehat{V}_0 = \sum_{h=1}^{H} \frac{1}{n_h(n_h-1)} \sum_{sh} (\frac{y_k}{p_k} - \widehat{t}_{\pi h})^2$$

siendo $p_k = \frac{\pi_k}{n_h}$

Si se aplica muestreo aleatorio simple, en cada estrato, la forma de $\widehat{V}_0$ será:

$$\widehat{V}_0 = \sum_{h=1}^{H} \frac{N_h^2 S_{ysh}^2}{n_h}$$

2.5.3. Muestreo polietápico

El mecanismo de selección de la muestra es el siguiente: en una primera etapa se seleccionan PSUs (unidades de primera etapa de muestreo), formándose una muestra s_I de n_I $PSUs$ extraída de $U_I = 1, 2, \ldots, N_I$, ya que la población es particionada en N_I subpoblaciones. A posteriori, se extraen submuestras en un determinado número de etapas, teniéndose unidades de segunda, tercera, o más etapas de muestreo.

El total poblacional de y, $t_y = \sum_U y_k$, se podrá estimar con:

$$\widehat{t}_\pi = \sum_{sI} \frac{\widehat{t}_i}{\pi_{Ii}}$$

donde $\widehat{t}_i = \sum_{s_i} \breve{y}_k$ y π_{Ii} es la probabilidad de la PSU i-ésima de ser seleccionada en la primera etapa.

El estimador simplificado de $V(\widehat{t}_\pi)$, es:

$$\widehat{V}_0 = \frac{1}{n_I(n_I-1)} \sum_{s} (\frac{\widehat{t}_i}{p_i} - \widehat{t})^2$$

con: $p_i = \frac{\pi_{Ii}}{n_I}$

Un resultado similar al descripto anteriormente,

$$E(\widehat{V}_0) - V = \frac{n_I}{n_I - 1}(V_0 - V)$$

A continuación, y a propósito del muestreo polietápico se presentarán algunos resultados que serán retomados en el capítulo siguiente, referentes al muestreo en dos etapas [Särndal et al., 1992].

En muestreo en dos etapas, la varianza para el estimador π del total poblacional puede escribirse como la suma de dos componentes:

$$V_{2st}(\widehat{t_\pi}) = V_{PSU} + V_{SSU},$$

con

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

donde $\check{t}_i = \frac{t_i}{\pi_{Ii}}$ y

$$V_{SSU} = \sum_{U_I} \frac{Vi}{\pi_{Ii}},$$

donde

$$V_i = -\frac{1}{2} \sum \sum_{Ui} \Delta_{kl/i} (\check{y}_{k/i} - \check{y}_{l/i})^2$$

Un estimador insesgado para $V_{2st}(\widehat{t_\pi})$ es

$$\widehat{V}_{2st}(\widehat{t_\pi}) = \widehat{V}_{PSU} + \widehat{V}_{SSU} = \sum \sum_{sI} \check{\Delta}_{Iij} \frac{\widehat{t}_{i\pi}}{\pi_{Ii}} \frac{\widehat{t}_{j\pi}}{\pi_{Ij}} + \sum_{sI} \frac{\widehat{V}_i}{\pi_{Ii}}$$

con

$$\widehat{V}_i = -\frac{1}{2} \sum \sum_{si} \check{\Delta}_{kl/i} (\check{y}_{k/i} - \check{y}_{l/i})^2$$

El cálculo de una estimación de la varianza puede llegar a ser engorroso, especialmente sabiendo que una estimación de la varianza $\widehat{V}_i$ debe ser calculada para cada $i \in s_I$. Por tanto, es necesario contar con un estimador de la varianza computacionalmente sencillo.

El siguiente estimador de la varianza

$$\widehat{V}^* = \sum \sum_{sI} \check{\Delta}_{Iij} \frac{\widehat{t}_{i\pi}}{\pi_{Ii}} \frac{\widehat{t}_{j\pi}}{\pi_{Ij}}$$

depende de las estimaciones de los t_i y de las probabilidades de primera etapa.

Las cantidades $\check{\Delta}_{Iij}$, están determinadas por la primera fase del diseño, razón por la cual la única información adicional necesaria para este estimador de la varianza son las estimaciones de los totales en las PSUs.

De los resultados antes mencionados se deduce que:

$$E(\widehat{V}^*) = V_{2st}(\widehat{t_\pi}) - \sum_{U_I} V_i,$$

por lo tanto el sesgo de $\widehat{V}^*$ es

$$B(\widehat{V}^*) = -\sum_{U_I} V_i,$$

y el sesgo relativo del estimador de varianza:

$$\frac{B(\widehat{V}^*)}{V_{2st}(\widehat{t_\pi})} = -\frac{\sum_{U_I} V_i}{\sum\sum_{U_I} \Delta_{Iij}\check{t}_i\check{t}_j + \sum_{U_I} \frac{V_i}{\pi_{Ii}}}$$

La fórmula precedentemente escrita puede revelar que en muchos casos la subestimación de la varianza del estimador $\widehat{V}^*$ carece de importancia. El numerador de la expresión a menudo será pequeño en comparación con el denominador si los π_{Ii} son pequeños, con poca o insignificante subestimación como consecuencia.

2.6. Grupos Aleatorios Independientes (IRG)

La técnica de grupos aleatorios independientes, (IRG, "independent random groups"), conduce a una forma sencilla de estimar la varianza. Tiene su origen en los trabajos de Mahalanobis 1939, 1944 y 1946 y Deming 1956.

La metodología consiste en tomar A muestras independientes y con reposición de tamaño $m = n/A$, es decir, $S_1, S_2, \ldots, S_a, \ldots, S_A$. La muestra completa S está conformada por la unión de las S_a, tal que $S = \bigcup_{a=1}^{A} S_a$.

Considerar:

1. $\widehat{\theta}(s)$ estimador de θ basado en s

2. $\widehat{\theta}_a(s)$ estimador de θ basado en s_a de la misma forma funcional que $\widehat{\theta}(s)$.

   Por la forma de sacar las muestras, $E(\widehat{\theta}_a)$ son todas iguales y los $\widehat{\theta}_a$,$\widehat{\theta}_b$ son incorrelacionados, para lo que $\widehat{V}_1$ es insesgada para $V(\widehat{\theta}*)$

Estimadores aplicados:

$$\widehat{\theta}_{IRG} = \frac{1}{A}\sum_{a=1}^{A}\widehat{\theta}_a$$

$$\widehat{V}_{IRG1} = \frac{1}{A(A-1)}\sum_{a=1}^{A}(\widehat{\theta}_a - \widehat{\theta}_{IRG})^2$$

$$\widehat{V}_{IRG2} = \frac{1}{A(A-1)}\sum A(\widehat{\theta}_a - \widehat{\theta})^2$$

Algunos problemas que se asocian a la técnica de IRG son:

1. La selección y recogida de datos para una serie de muestras independientes puede resultar más costosa y engorrosa que diseñar una gran muestra. Se debe ser muy cuidadoso en no generar una dependencia indeseada entre los $\widehat{\theta}_a$. Esta dependencia podría introducirse a través de los entrevistadores, o en la etapa de la entrada de datos por el personal de procesamiento.

2. Para dar estabilidad a la varianza del estimador, el número de submuestras debe ser grande. En la práctica normalmente no es grande, lo que hace que la varianza del estimador sea inestable. Mahalanobis propone usar sólo cuatro grupos, Deming sugiere diez.

2.7. Grupos Aleatorios Dependientes (DRG)

La técnica de los grupos aleatorios dependientes, en adelante (DRG, "dependent random groups") alude a un intento de adaptar la técnica de grupos aleatorios independientes a muestras que no cumplen con los requisitos de ellos.

Supóngase que primero se extrae una muestra grande de la población en su conjunto por un diseño de muestreo probabilístico; a partir de esta muestra S se ha obtenido un mecanismo aleatorio el cual se utilizará para dividir S en una serie de submuestras disjuntas, las cuales no serán independientes. Estas ideas fueron descriptas por Hansen, Hurwitz y Madow (1953).

Metodología:

Se saca una muestra S de tamaño fijo n, y luego por sencillez se asume que los grupos son de igual tamaño, $m = \frac{n}{A}$. Supóngase que S se divide en grupos por un mecanismo aleatorio de modo que cada grupo tiene el mismo diseño de muestreo que S, [Wolter, 1985].

De U tomo S según $p(s)$. Luego S es particionado en $S_1, S_2, \ldots, S_A$ muestras que se toman de S sin reposición y de forma que cada S_a sigue un diseño similar a $p(s)$. Ahora $S_1, S_2, \ldots, S_A$ son necesariamente disjuntas.

Siendo los estimadores:

$$\widehat{\theta}_{DRG} = \frac{1}{A}\sum_{a=1}^{A}\widehat{\theta}_a = \widehat{\theta}$$

$$\widehat{V}_{DRG1} = \frac{1}{A(A-1)}\sum_{a=1}^{A}(\widehat{\theta}_a - \widehat{\theta}_{DRG})^2$$

$$\widehat{V}_{DRG2} = \frac{1}{A(A-1)}\sum_{a=1}^{A}(\widehat{\theta}_a - \widehat{\theta})^2$$

Nota: Siempre que $\widehat{\theta}$ tenga la forma $\sum_s w_k y_k$ se tiene que $\widehat{\theta}_{DRG} = \widehat{\theta}$ y por tanto:

$$\widehat{V}_{DRG1} = \widehat{V}_{DRG2}$$

2.8. Jackknife

La palabra "Jackknife" alude a las navajas de uso múltiple empleadas para distintas funciones al mismo tiempo, llegando a alcanzar en alguna de ellas, y en determinadas condiciones, la eficacia y utilidad de una herramienta unitaria específica para dicha función. El uso de dicho vocablo referido al método de estimación que se describe, se debe a Tukey (1958), quien refleja con dicho simil semántico, su opinión sobre la utilidad del método como aplicable a una pluralidad de situaciones y problemas diferentes.

Antes de generalizarse el uso del vocablo Jackknife, el método se conocía con el nombre de Método Quenouille de Reducción del Sesgo, pues la idea básica fue concebida por Quenouille (1949) para tratar de disminuir el sesgo en la estimación de la correlación serial en problemas de series cronológicas.

Para poblaciones finitas la técnica de Jackknife es utilizada por primera vez por Durbin (1959), quien discutió el problema de la disminución de la eficiencia, llegando a la conclusión de que en algunos casos el uso de estimadores de Jackknife no sólo reduce el sesgo, sino que puede incluso reducir la varianza.

En muestreo de poblaciones finitas, conforme a la notación: de U se toma S según $p(s)$ (para cualquier diseño excepto el estratificado); S se particiona en $S_1, S_2, \ldots, S_a, \ldots, S_A$. En general S de tamaño n, S_a de tamaño m, y $Am = n$. Entonces $\widehat{\theta}$ estimador de θ basado en S, $\widehat{\theta}_{(a)}$ estimador de θ de la misma forma funcional que $\widehat{\theta}$ pero omitiendo a S_a (es decir basados en $(S - S_a)$).

Pseudovalores:

$$\widehat{\theta}_a = A\widehat{\theta} - (A-1)\widehat{\theta}_{(a)},$$

Promedio de pseudovalores:

$$\widehat{\theta}_{jk} = \frac{1}{A}\sum_{a=1}^{A}\widehat{\theta}_a,$$

Para estimar la varianza se tienen los siguientes estimadores de $V(\widehat{\theta_{jk}})$:

$$\widehat{V_{jk1}} = \frac{1}{A(A-1)} \sum_{a=1}^{A} (\widehat{\theta_a} - \widehat{\theta_{jk}})^2$$

$$\widehat{V_{jk2}} = \frac{1}{A(A-1)} \sum_{a=1}^{A} (\widehat{\theta_a} - \widehat{\theta})^2$$

2.9. Bootstrap

Efron y Tibshirani [Efron, 1993], ambos de la universidad de Stanford, escriben que la palabra bootstrap proviene del relato alemán Aventuras del Barón de Munchausen, donde se relata que estando el barón en el fondo del lago, logro salir del mismo impulsándose tirando de los cordones de sus botas cuando todo parecía perdido.

Fue Efron en el año 1979 quien creó la técnica que venía a mejorar el método Jackknife. Una característica básica del método de bootstrap es el principio de "plug-in" que consiste en la sustitución de la función subyacente de la distribución desconocida F por un estimador de la misma. Se emplea el remuestreo con sustitución para obtener gran número de remuestras sobre las que tener la base de estimación.

Este es uno de los métodos estadísticos denominados de computación intensiva. La metodología Bootstrap permite efectuar inferencias estadísticas sin necesidad de postular previamente que la distribución cumpla ciertas hipótesis que a veces son de difícil justificación.

Implementación para poblaciones finitas:

1. De U se toma S según $p(s)$ con n fijo, sin reposición y probabilidad de inclusión π_k. Luego se genera U^* repitiendo $\frac{1}{\pi_k}$ veces cada y_k con $k \in S$.

2. De U^* se toman A muestras según $p(s)$ y se calculan los $\widehat{\theta}_a$ con $a = 1, 2, \ldots, A$.

3.

$$\widehat{V}_{BS}(\widehat{\theta}) = \frac{1}{A-1}\sum_{a=1}^{A}(\widehat{\theta}_a - \widehat{\theta}^*)^2$$

con

$$\widehat{\theta}^* = \frac{1}{A}\sum_{a=1}^{A}\widehat{\theta}_a$$

.

Ejemplo:

$$\theta = t_y$$

$$\widehat{\theta} = \widehat{t}_{\Pi}$$

En i se toman muestras *pps* de U^* de tamaño n y $p_k = \frac{\pi_k}{n}$
Entonces: $N^* = \sum_{U^*} 1 = \sum_s 1\frac{1}{\pi_k} = \widehat{N}$

$$t_{U^*} = \sum_{U*} y_i^* = \sum_s \frac{y_k}{\Pi_k} = \widehat{t}_{\pi}$$

$$\sum_{U^*} p_k = \sum_s \frac{\pi_k}{n}\frac{1}{\pi_k} = \frac{1}{n}\sum_s = 1$$

$$\widehat{t}_a = \widehat{\theta}_a = \frac{\sum_{j=1}^{n} y_{ij}^*(a)}{p_{ij}^*(a)}$$

para $a = 1, 2, \ldots, A$

El indice i indica que se está en U^* y j indica que salió en la j-ésima extracción.

$$\widehat{V}_{Bt}(\widehat{t}) = \frac{1}{A-1}\sum_{a=1}^{A}(\widehat{t}_a - \widehat{t}^*)^2$$

con:

$$\widehat{t}^* = \frac{1}{A}\sum_{a=1}^{A}\widehat{t}_a$$

$$E(\widehat{V}_{Bs}) = \frac{(n-1)}{n}E(\widehat{V}_0)$$

Capítulo 3

Técnicas de Estimación de la Varianza aplicadas al Censo y a la ENHA

3.1. Definiciones y conceptos referidos al censo y a la ENHA

La encuesta continua de hogares ampliada 2006, fue un relevamiento que realizó el INE, cuya población objetivo son los residentes en viviendas particulares en todo el territorio nacional [Beltrami, 2006].

El marco está basado en los listados por zona censal del Censo 2004 (CF1, Censo fase 1), habiéndose realizado la estratificación sobre este marco.

El contar con los datos del censo fase uno permite acceder al marco muestral de la encuesta, pudiéndose de esta forma no sólo simular el diseño de la ENHA, a partir de la utilización de los datos del censo fase uno correspondientes al departamento de Colonia para estimar parámetros de interés, sino que además permitirá contrastar con los valores poblacionales de los pa-

rámetros así como también comparar las varianzas estimadas con distintas técnicas con las verdaderas varianzas de los estimadores.

La ENHA 2006 cubre a todo el territorio nacional, por lo cual las unidades primarias de muestreo(zonas) se agrupan según el siguiente criterio:

1. Localidades de 5.000 habitantes o más de cada departamento,
2. Localidades urbanas de menos de 5.000 habitantes,
3. Área rural.

En el departamento de Montevideo se cubren todas las zonas censales; la periferia de Montevideo (considerándose conjuntamente con éste como Área Metropolitana) incluye todas las zonas censales de todas las localidades urbanas hasta un límite medio de 30 kilómetros al centro de Montevideo.

En el año 2006, el tamaño de la muestra alcanzó a 87.228 viviendas (7.269 viviendas por mes) distribuido en:

1. 35 % Montevideo,
2. 3 % en la periferia,
3. 31 % en el interior urbano residente en localidades de 5.000 habitantes o más,
4. 12 % en localidades de menos de 5.000
5. 19 % en zonas rurales.

La muestra comprende aproximadamente 259.000 personas.

La técnica de muestreo que se aplica en la $ENHA$ es el muestreo aleatorio estratificado en conglomerados con asignación óptima, en dos o tres etapas de selección.

Selección en dos etapas: en el Montevideo Metropolitano y en las localidades de 5000 habitantes o más del resto del país, se aplican dos etapas para la selección de viviendas. La unidad primaria de muestreo UPM es la manzana (zona), y la unidad secundaria de muestreo USM es la vivienda/hogar. La selección se hace por muestreo sistemático dentro del estrato, con punto de arranque aleatorio e intervalo constante.

Selección en tres etapas: en el resto de los estratos se aplican tres etapas para la selección de viviendas. En este caso, la UPM es la localidad de menos de 5.000 habitantes, la USM es la zona, y la unidad de tercera etapa UTM es la vivienda/hogar.

Tamaño de las UPM: a los efectos de asegurar que todas las zonas del país tengan siempre probabilidad positiva de ser seleccionadas, las UPM son las "zonas censales" a condición de tener un número entre 18 y 160 viviendas particulares ocupadas en el marco. Si esta condición no se cumple, las UPM son agrupamientos o particiones de zonas.

Agrupaciones de localidades y áreas para seleccionar la muestra: con el fin de seleccionar la muestra se definieron 59 estratos de localidades y áreas.

Nota: [Beltrami, 2005b] "Técnicamente dominios (de estudio) pueden o no coincidir con los estratos, pero mientras no se exprese algo en contrario, en este documento serán utilizados como sinónimos". Los dominios de estudio definidos fueron:

1. Todo el país.
2. Montevideo.
3. Resto del país.
4. Departamentos.
5. Área:

- Urbana
- Rural

6. Tamaño de localidades conurbanas:
   - 5.000 habitantes o más.
   - Menos de 5.000 habitantes.

Como se verá posteriormente la agrupación de las localidades de más de 5000 habitantes de cada departamento constituyen un estrato dando lugar de esta forma a la conformación de 18 estratos, en tanto la agrupación del departamento de Montevideo y su periferia (Montevideo metropolitano) forman 5 estratos.

Los estratos son:

1. Montevideo: 4 estratos.
2. Departamentos del resto urbano del país con localidades de 5.000 o más habitantes, incluyendo la periferia metropolitana de Montevideo (19 agrupaciones).
3. Departamentos del resto urbano del país con localidades de menos de 5.000 habitantes (18 agrupaciones).
4. Área rural de cada departamento del resto del país (18 agrupaciones).

Estratos:

Nota: [Beltrami, 2005b] Por "estratificar" se entenderá el conjunto de operaciones tendientes a agrupar las unidades de información, medidas según criterios tales que los grupos resulten internamente homogéneos y externamente heterogéneos".

Los estratos son una partición más fina de los dominios, definidos para lograr mayor homogeneidad entre las unidades y así hacer más eficiente el diseño. La estratificación aplicada es diferente en cada dominio:

1. En Montevideo la estratificación es por nivel socioeconómico, y de acuerdo al documento [Beltrami, 2005b] los pasos de la clasificación fueron:

   - Se clasificaron los segmentos de Montevideo según los cuartiles a los que pertenecían sus promedios de ingreso per capita (en unidades reajustables).
   - Se aplicó la rutina de Cluster generando cuatro grupos: Bajo, Medio Bajo, Medio Alto y Alto.

2. En el resto de localidades mayores del país, los estratos son el conjunto de localidades de más de 5.000 habitantes por departamento, excepto el anillo perteneciente a Canelones y San José desde el límite con Montevideo hasta aproximadamente el kilómetro 30. Este anillo, que incluye todas las localidades conurbanas de cualquier tamaño dentro de su perímetro, junto con los 4 estratos de Montevideo constituyen el Montevideo Metropolitano.

3. El resto de localidades pequeñas del país, forman 18 estratos (uno por cada departamento).

4. En el área rural también se forman 18 estratos.

De las múltiples áreas que releva la E.N.H.A., este trabajo focalizará las aplicaciones en la situación ocupacional, y por tanto se citarán los conceptos que tomó el I.N.E. respecto de las categorías de ocupación:

1. ACTIVO: Comprende a las personas de 14 o más años de edad que tienen al menos una ocupación en la que vierten su esfuerzo productivo a la sociedad, o que, sin tenerla, la buscan activamente durante el período de referencia elegido por la encuesta [Beltrami, 2006].

2. OCUPADO: Personas que trabajaron por lo menos 1 hora durante el período de referencia de la encuesta, o que no trabajaron por estar de vacaciones, por enfermedad, accidente, conflicto de trabajo, etc. Se incluye en esta categoría a los trabajadores familiares no remunerados y a los docentes honorarios.

3. DESOCUPADO: Personas que durante el período de referencia no estaban trabajando por no tener empleo, pero que buscaban un trabajo remunerado o lucrativo, y que se encuentran disponibles para comenzar a trabajar. Esta categoría comprende a las personas que trabajaron antes pero perdieron su empleo (desocupados propiamente dichos), y aquéllos que buscan su primer trabajo. Los desocupados propiamente dichos incluyen a los que reciben un subsidio estatal (seguro de paro) y a los que no lo reciben.

4. INACTIVO: Personas que no aportan su trabajo para producir bienes o servicios económicos. Se clasifican en las siguientes categorías: personas que se ocupan solamente del cuidado de su hogar, estudiantes y personas que sin desarrollar ninguna actividad económica, perciben ingresos.

5. DESEMPLEO ABIERTO: Son las personas pertenecientes a la fuerza de trabajo que estaban sin trabajo durante el período de referencia, que están disponibles para trabajar de inmediato y que habían tomado medidas concretas durante el período de referencia, para buscar un empleo asalariado o un empleo independiente. Se incluyen también en esta categoría a las personas que no buscaron activamente trabajo durante el período de referencia por razones de enfermedad o esperando noticias.

Dentro de la población ocupada se destaca la siguiente subdivisión de

acuerdo a su relación con el trabajo o las características del mismo: subempleo, trabajo no registrado, precariedad e informalidad.

- SUBEMPLEO: El subempleo visible se define como una subcategoría del empleo. Existen tres criterios para identificar entre las personas ocupadas a las visiblemente subempleadas:

  -Trabajan menos de la duración normal.

  -Lo hacen de forma involuntaria.

  -Desean trabajo adicional y están disponibles para el mismo durante el período de referencia.

  Para considerar a una persona en situación de subempleo visible, los tres criterios deberán ser satisfechos simultáneamente.

- PRECARIEDAD o INFORMALIDAD: La concepción de empleo precario será por consiguiente la de incluir los empleos formales en los cuales no se presenta un cumplimiento de las leyes laborales, y en el caso de los empleos considerados tradicionalmente informales, excluir a quienes sí tienen acceso a seguridad social que a pesar de pertenecer a empresas pequeñas o ser independientes se han formalizado. Se considera entonces como trabajo precario o informal:

  -Empleados privados sin cobertura de la seguridad social.

  -Empleados privados con un empleo inestable.

  -Trabajadores no remunerados.

- TRABAJADOR NO REGISTRADO: Operativamente la encuesta considera como trabajador no registrado a todas aquellas personas ocupadas que declaran no realizar aportes a ningún organismo de seguridad social.

Notación:

Tasa Bruta de Participación:

$$TBP = \frac{PEA}{PT} * 100$$

Donde: PEA= Población Económicamente Activa PT = Población Total

Tasa de Actividad, o Tasa Global de Participación:

$$TA = \frac{PEA}{PET} * 100$$

Donde: PET = Población en Edad de Trabajar

Tasa de Desempleo o Desempleo Abierto:

$$TD = TDA = \frac{DA}{PEA} * 100$$

Donde: DA = Desempleo Abierto

Tasa de Subempleo:

$$TS = \frac{S}{PEA} * 100$$

Donde: S= Subempleo

Tasa de Empleo o Tasa de Ocupación (es equivalente al tamaño de la demanda laboral)

$$TE = TO = \frac{O}{PEA} * 100$$

Donde: O = Ocupados

Considerando como se menciona previamente, que el marco muestral de la ENHA es el CF1 (Censo fase 1), en una primera instancia se trabajará con un diseño de muestreo de forma de poder realizar estimaciones que podrán ser contrastadas con los valores poblacionales.

Se tomará la base de datos del CF1 perteneciente al departamento de Colonia, el contenido de la misma se puede consultar en [INE, 2004b].

Previo a la realización de la aplicación se explicitarán las principales definiciones y conceptos [INE, 2004a].

3.1.1. Divisiones Geoestadísticas

1. Departamento Censal: Coincide con los límites político administrativos departamentales.

2. Sección Censal: cada departamento se divide en Secciones Censales. Son porciones importantes de territorio en áreas urbanas y rurales, que guardan cierta relación con los límites de las Secciones Judiciales, aunque no son enteramente coincidentes.

3. Segmento Censal: cada Sección Censal se divide en Segmentos Censales. En áreas urbanas un Segmento Censal es un conjunto de manzanas y en áreas rurales una porción de territorio que agrupa unidades menores con límites físicos reconocibles en el terreno.

4. Zona Censal: son las unidades menores identificables en el terreno por límites naturales o artificiales, de fácil reconocimiento (accidentes geográficos, vías de comunicación, etc.). En el área urbana las zonas coinciden prácticamente con las manzanas y en áreas rurales, con porciones elementales de territorio limitadas por carreteras,caminos vecinales, cursos de agua, etc.

5. Unidad Locativa: las unidades locativas son las viviendas o locales no destinados a viviendas.

6. Vivienda: es toda habitación o conjunto de habitaciones y sus dependencias, que ocupan un edificio o una parte separada y que, por la forma de su construcción, transformación o acondicionamiento, se destina a ser habitada por personas, y que, en el momento de ser censada no se utiliza totalmente para otros fines. A los efectos censales también es considerada como vivienda todo albergue fijo o móvil donde una persona o grupos de personas viven habitualmente.

7. Local no destinado a vivienda: se refiere a las unidades locativas cuya finalidad principal es comercial, industrial, de servicios de salud, educativos, sociales, culturales, deportivos u otros diferentes a vivienda, sin prejuicio de que en los mismos pueda darse el caso de que vivan personas habitualmente.

8. Vivienda particular: es aquella que alberga un hogar particular.

9. Vivienda colectiva: es aquella que alberga un hogar colectivo.

10. Hogar particular: es el conjunto de personas con o sin vínculos de parentesco, que habitan un mismo techo y que al menos para su alimentación dependen de un fondo común o presupuesto para la comida. También se trata de un hogar particular cuando hay una persona que vive sola.

11. Hogar colectivo: es el conjunto de personas que comparten la vivienda por razones de trabajo, atención médica, estudios, militares, de reclusión, religiosas, etc.

3.2. Método de Monte Carlo aplicado al Censo

Se hizo un estudio de Monte Carlo tomando múltiples muestras del censo de población y vivienda para estudiar el comportamiento de los estimadores usados en la ENHA. Esta sección se restringirá a los registros de personas del censo de población y viviendas fase 1 del año 2004 para el departamento de Colonia. Debe destacarse que el archivo no tiene identificadas las zonas censales. En primera instancia se procederá a depurar la base de datos, la cual cuenta con 129.379 registros, eliminándose las viviendas vacías y los hogares colectivos, lo cual corresponde a 10.113 registros de viviendas desocupadas y 1555 registros de personas que viven en hogares colectivos. De esta forma se conforma una base de 117.711 registros de personas en hogares particulares.

Posteriormente se eliminarán los segmentos que contengan menos de 5 hogares (en particular se encontró un sólo segmento el "0512011"que cuenta con 3 personas distribuidas en 2 hogares) con dicha característica. Ulteriormente a la depuración, la base cuenta con 198 segmentos, 40.241 hogares y 117.708 personas.

Por otra parte se generarán nuevas variables: "escolar" definida como las personas cuya edad es mayor o igual a 6 años y menor o igual que 12 años y "población en edad de trabajar" (PET) que se define como las personas de 14 o más años de edad (sin establecerse un límite superior). Contar con los datos censales permitirá calcular valores poblacionales que después serán estimados bajo un determinado diseño de forma de probar el diseño utilizado, ya que se podrán contrastar las estimaciones con los verdaderos valores poblacionales de los parámetros de interés:

1. Total de personas: 117.708.

2. Total de niños en edad escolar: 13.460.

3. Promedio de edad: 36.02.

4. Proporción de personas en edad de trabajar: 78.48 %.

Se utiliza una muestra compleja que trata de replicar el diseño utilizado en la ENHA. Considerándose un diseño por conglomerados en 2 etapas, tal que las unidades de primera etapa de muestreo PSUs son los segmentos censales, los cuales se seleccionarán de forma sistemática con probabilidades de inclusión proporcionales al tamaño del segmento ($SY\pi ps$). Debe destacarse respecto del diseño, que el mismo no es medible ya que no se cumple que $\pi_{kl} > 0 \forall\ k$ y l, razón por la cual no existe un estimador insesgado de la varianza del estimador π. En primera instancia se seleccionaran 50 de los 198 segmentos, lo cual implica una fracción de muestreo en la primera etapa de

aproximadamente 25 %, posteriormente, y en función de los resultados obtenidos se consideró trabajar con una fracción de muestreo considerablemente menor, aproximadamente 5 % para ver como impactaba en los resultados (particularmente en el V_0).

Las probabilidades de inclusión de primer orden π_{I_i}, son proporcionales a la cantidad de hogares del segmento: $\pi_{I_i} = n_I * N_i / \sum_i N_i$, siendo n_I el tamaño de la muestra de primera etapa, y N_i la cantidad de hogares del segmento i. Cuando el tamaño de muestra en la primera etapa es 50 el valor mínimo de π_{I_i} es $0,0087$ mientras que el valor máximo de pi_{I_i} es $0,9692$. El hecho de que no haya ningún segmento con probabilidad de inclusión igual a 1 implica que no habrá ningún segmento de inclusión forzosa. Cuando el tamaño de muestra en la primera etapa es 10, el mínimo π_{I_i} es de 0,0017 en tanto que el valor máximo para π_{I_i} es 0.1938.

Las unidades de segunda etapa de muestreo (SSUs) se seleccionarán mediante un diseño aleatorio simple sin remplazo(SI), tal que se sortearán 5 hogares de aquellos segmentos que fueron seleccionados en la primera etapa. La fracción de muestreo en la segunda etapa es de aproximadamente $0,6$ %. Todos los hogares de los segmentos seleccionados en la primera etapa tienen igual probabilidad de ser seleccionados: ya que $\pi_{I_i} \propto N_i$, y $\pi_{k/i} = 5/N_i$, la probabilidad de los hogares es constante (diseño autoponderado) e igual a $\pi_k = n_I * N_i / \sum_i N_i * 5/N_i = 250/40241$.

Posteriormente se tomarán todas las personas de los hogares seleccionados.

Se presenta la contribución relativa de la primer y segunda etapa a la variabilidad total, el desvío estándar y el sesgo relativo para $nI = 10$, que surgen del cálculo de la verdadera varianza de los parámetros estimados.

TABLA 1: Contribución relativa de la VPSU y la VSSU a la varianza, desvío estándar y sesgo relativo para nI=10

	$\frac{VPSU}{VPSU+VSSU}$	$(VPSU+VSSU)^{0,5}$	Sesgo Relativo
total personas	10.37 %	9477	5.6E-06
promedio edad	13.59 %	2.69	5.3E-06
total escolares	7.78 %	3925	5.7E-06
promedio PET	9.24 %	3.94 %	5.6E-06

TABLA 2: Contribución relativa de la VPSU y la VSSU a la varianza, desvío estándar y sesgo relativo para nI=50

	$\frac{VPSU}{VPSU+VSSU}$	$(VPSU+VSSU)^{0,5}$	Sesgo Relativo
total personas	9.47 %	4217	2.8E-05
promedio edad	8.78 %	1.17	2.8E-05
total escolares	5.16 %	1730	2.9E-05
promedio PET	5.56 %	1.72 %	2.9E-05

Del análisis comparativo de las tablas uno y dos se desprende que para todos los parámetros considerados se tiene que el desvío estándar es menor (aproximadamente 50 %) cuando nI=50, sin embargo la contribución de la primera etapa es mayor cuando nI=10, en tanto que disminuye el sesgo relativo cuando nI=10 en la medida que se tienen π_{Ii} menores.

Se extraen 10.000 muestras a partir de un estudio Monte Carlo, lo cual permitirá tener 10.000 estimaciones de los parámetros mencionados. A partir de las 10.000 simulaciones se obtienen vectores de totales y desvíos. Las varianzas obtenidas a partir de las 10.000 simulaciones se les denominara varianza MC donde MC hace referencia al método de simulación utilizado (Monte Carlo), y a la raíz cuadrada de las varianzas MC se le denominara desvío MC para cada uno de los parámetros de interés. A continuación se presentan los valores comparados de medias y totales poblacionales con estimaciones para $nI = 10$ y $nI = 50$.

TABLA 3: Valores poblacionales y estimaciones Monte Carlo de los parámetros

	Edad promedio	Total escolares	Total personas	promedio PET
Poblacional	36.02	13460	117708	78.48 %
Est.MC $nI = 10$	36.23	13418	117489	78.70 %
Est.MC $nI = 50$	36.05	13460	117693	78.52 %

Donde: Est.MC $nI = 10$ y Est. $nI = 50$ corresponde a los valores estimados para los parámetros de interés por el método Monte Carlo.

De acuerdo a los resultados presentados en la tabla 3, se puede realizar un análisis comparado de los valores poblacionales de los parámetros de interés respecto a las estimaciones realizadas con el paquete survey [Lumley, 2006], el cual estima los totales poblacionales a partir de los totales a nivel de segmento. De los 50 segmentos seleccionados en la primera etapa se sortean 5 hogares en cada uno, se suman las personas de los 5 hogares, se divide entre 5 y multiplica por la cantidad de hogares en la población (que se obtiene sumando los pesos de los 250 hogares de la muestra), lo que es equivalente a utilizar $\pi_{Ii} = n_i * N_i / \sum_i N_i$.

Puede observarse que para los ratios (edad promedio y PET promedio) para nI=10 y para nI=50 los valores simulados están muy próximos a los poblacionales, destacando que cuando nI=50 se aproxima mejor que cuando nI=10. Para los totales (escolares y personas) también los valores simulados son muy cercanos a los poblacionales tanto cuando nI=10 como cuando nI=50, siendo mejores las aproximaciones cuando nI=50. Para el total de escolares cuando nI=50 el valor simulado coincide con el poblacional.

Los siguientes cuadros presentan los desvíos muestrales de los parámetros estimados así como también la raíz cuadrada de los promedios de las varianzas estimadas de los parámetros estimados por $\widehat{V}_0$ y $\widehat{V}_{jk}$ para $nI = 10$

y $nI = 50$ respectivamente.

TABLA 4: Desvío MC de los estimadores y raíz cuadrada del promedio de las varianzas estimadas por V_0 y V_{jk} para nI=10

	promedio edad	total escolares	total personas	promedio PET
$rc(\widehat{V}(\widehat{t}))$	2.72	3868	9370	3.94 %
$rc(media(\widehat{V_0}(\widehat{t}))$	2.85	4024	9832	3.99 %
$rc(media(\widehat{V_{jk}}(\widehat{t}))$	2.88	4024	9832	4.07 %

TABLA 5: Desvío MC de los estimadores y raíz cuadrada del promedio de las varianzas estimadas por V_0 y V_{jk} para nI=50

	promedio edad	total escolares	total personas	promedio PET
$rc(\widehat{V}(\widehat{t}))$	1.16	1746	4231	1.73 %
$rc(media(\widehat{V_0}(\widehat{t}))$	1.27	1806	4403	1.81 %
$rc(media(\widehat{V_{jk}}(\widehat{t}))$	1.27	1806	4403	1.81 %

Notación:

rc: es la raíz cuadrada

t: se le llama genéricamente a cada uno de los parámetros que se quiere estimar ("promedio edad", "total escolares", "total personas", "promedio PET").

TABLA 6: Variación porcentual entre el desvío MC y el desvío poblacional de los estimadores considerando nI=10 y nI=50.

	$[\frac{rc(V_{MC})}{rc(V_{poblacional})} - 1] * 100$ (nI=10)	$[\frac{rc(V_{MC})}{rc(V_{poblacional})} - 1] * 100$ (nI=50)
total personas	-1.13 %	0.32 %
promedio edad	1.25 %	-1.06 %
total escolares	-1.46 %	0.93 %
promedio PET	0.16 %	0.52 %

TABLA 7: Variación porcentual entre la raíz cuadrada del promedio de las varianzas estimadas por V_0 y el desvío MC, y la raíz cuadrada del promedio de las varianzas estimadas por Jackknife y el desvío MC para nI=10.

	$[\frac{rc\overline{V_0}}{rcV_{MC}} - 1] * 100$	$[\frac{rc\overline{V_{jk}}}{rcV_{MC}} - 1] * 100$
total personas	4.93 %	4.93 %
promedio edad	4.64 %	5.93 %
total escolares	4.04 %	4.04 %
promedio PET	1.26 %	3.17 %

TABLA 8: Variación porcentual entre la raíz cuadrada del promedio de las varianzas estimadas por V_0 y el desvío MC, y la raíz cuadrada del promedio de las varianzas estimadas por Jackknife y el desvío MC para nI=50.

	$[\frac{rc\overline{V_0}}{rcV_{MC}} - 1] * 100$	$[\frac{rc\overline{V_{jk}}}{rcV_{MC}} - 1] * 100$
total personas	4.06 %	4.06 %
promedio edad	9.51 %	9.76 %
total escolares	3.44 %	3.44 %
promedio PET	4.39 %	4.78 %

De las tablas cuatro y cinco se desprende que conforme el tamaño de muestra es mayor la varianza es menor, en tanto que si se observan las tablas siete y ocho se aprecia que la reducción la fracción de muestreo no genera diferencias importantes en las variaciones relativas entre la varianza MC y el promedio de las varianzas estimadas por V_0.

Se presenta a continuación los porcentajes de cobertura obtenidos a partir del cálculo de los intervalos de confianza para los parámetros de interés a un 95 % de confianza, tal que para una distribución t-student con nI-1 grados de libertad $(10-1=9)$ y $(50-1=49)$ respectivamente genera valores para el percentil 97,5 de 2.26 y 2.00.

TABLA 9: Porcentaje de cobertura para $nI = 10$:

	promedio edad	total escolares	total personas	promedio PET
IC	96.01 %	94.1 %	95.24 %	94.25 %
ICjk	96.02 %	94.1 %	95.24 %	94.45 %

TABLA 10: Porcentaje de cobertura para $nI = 50$:

	promedio edad	total escolares	total personas	promedio PET
IC	96.49 %	95.33 %	95.72 %	95.68 %
ICjk	96.51 %	95.33 %	95.72 %	95.76 %

En las tablas nueve y diez se puede observar que tanto cuando nI=10 como cuando nI=50 los porcentajes de cobertura son superiores al 95 % (aproximadamente 96 %) lo cual indica que la varianza está siendo sobreestimada, no existiendo diferencias importantes al cambiar el tamaño de muestra en la primera etapa. Debe destacarse que cuando nI=10, para el total de escolares y el promedio de PET, el porcentaje de cobertura es inferior al 95 % (aproximadamente 94 %), habiéndo por tanto una subestimación de la varianza.

3.3. Aplicación de estimación de varianza a la ENHA

Este ejemplo estará restringido al departamento de Montevideo, lo cual lleva a trabajar con una base de 89.919 personas. Considerando que el archivo no tiene identificadas las zonas censales se seleccionará en una primera etapa segmentos censales, y en la segunda etapa se seleccionaran hogares/personas, considerándose a su vez la estratificación previamente definida para el departamento de Montevideo. También debe señalarse que se realizaran estimacio-

nes mensuales para los totales y ratios de interés.

Con la utilización de [Lumley, 2006] se reproducirá un diseño estratificado en dos etapas obteniéndose los siguientes resultados para las estimaciones mensuales de los totales, ratios y varianzas:

TABLA 11: Estimación de totales mensuales por mes

	ptm	petm	peam	pdm
enero	1263565	1026692	649432	84235
febrero	1224179	1015974	608907	72850
marzo	1272275	1027473	623120	82106
abril	1282448	1024518	621039	68069
mayo	1256216	1022387	635268	68527
junio	1221530	989445	597453	61639
julio	1232312	992822	602774	59675
agosto	1259390	1019968	637454	68020
setiembre	1245335	1015717	636271	66327
octubre	1247328	1019519	633692	62953
noviembre	1249099	1003429	628481	63241
diciembre	1250614	1024831	642437	58014

Donde:

ptm: es la población total por mes

petm: es la población en edad de trabajar por mes

peam: es la población económicamente activa por mes

pdm: es la población desocupada por mes

TABLA 12: Varianza muestral de los totales por mes

	ptm	petm	peam	pdm
enero	34060	26209	18474	4626
febrero	32133	25841	17850	4483
marzo	33809	25455	17911	4635
abril	33067	25147	16670	4151
mayo	33151	25875	17467	3996
junio	32382	25156	17665	3903
julio	31821	24096	16170	3936
agosto	36427	27447	18584	4350
setiembre	30740	24372	17738	4232
octubre	34201	25622	18257	3989
noviembre	34499	26492	18356	4285
diciembre	32408	24925	17432	3546

TABLA 13: Estimación de ratios mensuales expresados en porcentajes por mes

	tbpm	tam	tdesocm
enero	51.40 %	63.25 %	12.97 %
febrero	49.74 %	59.93 %	11.96 %
marzo	48.98 %	60.65 %	13.18 %
abril	48.43 %	60.62 %	10.96 %
mayo	50.57 %	62.14 %	10.79 %
junio	48.91 %	60.38 %	10.32 %
julio	48.91 %	60.71 %	9.90 %
agosto	50.62 %	62.50 %	10.67 %
setiembre	51.09 %	62.64 %	10.42 %
octubre	50.80 %	62.16 %	9.93 %
noviembre	50.31 %	62.63 %	10.06 %
diciembre	51.37 %	62.69 %	9.03 %

Donde:

tbpm: es la tasa bruta de participación por mes

tam: es la tasa de actividad por mes

tdesocm: es la tasa de desempleo por mes

TABLA 14: Varianza muestral para los ratios estimados por mes

	tbp	ta	tdesoc
enero	0.66 %	0.76 %	0.60 %
febrero	0.63 %	0.74 %	0.64 %
marzo	0.64 %	0.72 %	0.61 %
abril	0.63 %	0.73 %	0.56 %
mayo	0.61 %	0.73 %	0.57 %
junio	0.70 %	0.82 %	0.54 %
julio	0.70 %	0.79 %	0.55 %
agosto	0.64 %	0.73 %	0.65 %
setiembre	0.65 %	0.76 %	0.58 %
octubre	0.65 %	0.76 %	0.55 %
noviembre	0.69 %	0.77 %	0.59 %
diciembre	0.68 %	0.79 %	0.50 %

Los resultados de las tablas 11 y 13 fueron verificados contra los resultados publicado por el INE, observándose diferencias menores a $0,1\,\%$ para la tasa de actividad por mes y de 0,5 % en el caso de la tasa de desempleo por mes. En el caso de los totales por mes estimados los resultados coincidieron con los publicados.

Las ventajas del método empleado pueden enumerarse en:

- Es relativamente sencillo de implementar en la medida que no es condición necesaria tener las probabilidades de inclusión de segundo orden ni para la primera ni para la segunda etapa de muestreo.
- Ya está incluido dentro del software, haciendo posible su aplicación directa.
- Los resultados que se obtiene son similares a los que llega el INE a pesar

de no haber utilizado un diseño idéntico al de la ENHA por no estar disponibles en las bases que pública dicho instituto las zonas censales.

Capítulo 4

Conclusiones

El uso del estimador de varianza V_0 se justifica en la medida que se cumplen las condiciones de aplicabilidad de dicho estimador. Se utiliza una fracción de muestreo pequeña, que el tamaño de la población es grande, y por lo tanto, el muestreo sin remplazo se comporta como si fuera con remplazo, pues la probabilidad de que una unidad salga seleccionada en más de una oportunidad es pequeña.

Las condiciones teóricas que justifican la aplicación del V_0 han sido verificadas en el ejemplo de simulación Monte Carlo desarrollado para el departamento de Colonia en la medida que las varianzas estimadas por V_0 ó Jaccknife dieron resultados muy próximos a los valores poblacionales, por otra parte los sesgos relativos son muy pequeños lo cual se fundamenta por el hecho de que las π_{Ii} son chicos. En cuanto a los intervalos de confianza, los porcentajes de cobertura que aparecen en las tablas 9 y 10 en todos los casos similares al 95 %.

Si bien no se reprodujo un diseño idéntico al utilizado para la Encuesta Nacional de Hogares Ampliada por no contar con las zonas censales, en la primera etapa se seleccionaron segmentos censales, obteniéndose resultados similares que los obtenidos por el Instituto Nacional de Estadística.

Para los totales estimados las varianzas son relativamente pequeñas en el departamento de Montevideo como se muestra en las tablas 11 y 12 para una estimación de población total de aproximadamente 1.200.000 habitantes la varianza estimada es de aproximadamente 34.000

Otro aspecto a considerar es el referente a los coeficientes de variación los cuales están en el orden de 0.01 % para la población total, población en edad de trabajar y la población económicamente activa en tanto que para el total de desocupados está en el orden de 0.08 %. Para los ratios estimados, tasa bruta de participación, tasa de actividad y tasa de desocupación, los coeficientes de variación son mucho mayores, y están en el orden de 16 %, 14 % y 60 % respectivamente.

Un resultado que llama la atención, y que podría ser abordado en futuras investigaciones, es el de las diferencias tan importantes en los coeficientes de variación de los totales y los de los ratios, destacándose como particularmente elevado el coeficiente de variación de la tasa de desocupación.

Se podría encarar desde otras técnicas de estimación de la varianza como el Bootstrap.

Apéndice A

Apéndice

Se incluye a continuación el código fuente en R, utilizado para la definición del diseño muestral y para el cálculo de las estimadores y sus varianzas. El programa está comentado a los efectos de facilitar su comprensión. Por más detalle, ver la documentación de R asociada al paquete survey: [Lumley, 2006].

```
rm(list=ls())
library(foreign)
library(survey)

#Se utilizó la base de datos de la ENHA 2006
d<-as.data.frame(read.spss(file="enha_m_tesis.sav"))
dim(d)
#head(d)
codsegm<-as.numeric(paste(d$dpto,d$secc,d$segm,sep=""))
dc<-cbind(codsegm,d)
#denha<-svydesign(id=~codsegm+correlativ,strata=~estrato,
weights=~pesomen, data=dc)
```

```
summary(denha)
table(d$estrato)
# La variable de identificación es un código compuesto
# por departamento, sección y segmento.
# Mediante la utilización del paquete survey, se especifico
# el diseño "denha".
# Donde la identificación corresponde a "id=~codsegm+correlativ",
# luego los estratos "strata=~estrato", que son los 4 definidos
# para el departamento de Montevideo (alto, medio alto,
# medio bajo y bajo), para los pesos "weights=~pesomen" se
# utilizo pesomen porque se optó por realizar
# estimaciones mensuales de los parámetros de interés.


pea<-as.numeric(dc$POBPCOAC %in% c("Ocupados",
              "Desocupado, busca por primera vez",
              "Desocupados propiamente dichos",
              "Desocupados en seguro de paro") )
pt<-rep(1,length(dc$POBPCOAC))
pet<-as.numeric(dc$e27>=14)
po<-as.numeric(dc$f62==1 | dc$f63==1 | (dc$f64==1 &
                           dc$f65 %in% c(1,2,3,4)) )
pd<-as.numeric(pea==1 & po==0)
table(pea,pet)
table(pea,po )
table(dc$POBPCOAC, pea)
table(dc$POBPCOAC, pet)
table(dc$POBPCOAC, po )
table(dc$POBPCOAC, dc$f62)
```

```
# corrige inconsistencias
pea[dc$e27<14]<-0
po[dc$e27<14]<-0
po[is.na(po)]<-0

t<-matrix(0,12,5)
v<-matrix(0,12,5)
er<-matrix(0,12,4)
vr<-matrix(0,12,4)

for(m in 1:12)
{
ind<-as.numeric(dc$mes)==m
dcm<-dc[ind,]
ptm <- pt[ind]
petm<-pet[ind]
peam<-pea[ind]
pom <- po[ind]
pdm <- pd[ind]
dcm<-cbind(dcm,ptm,petm,peam,pom,pdm)
denham<-svydesign(id=~codsegm+correlativ,strata=~estrato,
weights=~pesomen, data=dcm)

#Se estiman los totales y sus varianzas con el comando "svytotal"
a     <-as.data.frame(svytotal(~ptm+petm+peam+pom+pdm, denham))
t[m,]<-t(a[,1])
v[m,]<-t(a[,2])
```

```
#Con el comando "svyratio" se estiman los ratios y sus varianzas"
# tbp (t bruta participacion)
b<-svyratio(numerator=~peam, denominator=~ptm,design=denham,
separate=FALSE,se=TRUE)
b<-as.data.frame(predict(b,total=1))
er[m,1]<-b[,1]
vr[m,1]<-b[,2]

# t actividad
b<-svyratio(numerator=~peam, denominator=~petm,design=denham,

separate=FALSE,se=TRUE)
b<-as.data.frame(predict(b,total=1))
er[m,2]<-b[,1]
vr[m,2]<-b[,2]



# t desocupacion
b<-svyratio(numerator=~pdm, denominator=~peam,design=denham,

separate=FALSE,se=TRUE)
b<-as.data.frame(predict(b,total=1))
er[m,4]<-b[,1]
vr[m,4]<-b[,2]

}

colnames(t)<-rownames(a)
colnames(v)<-rownames(a)
```

```
colnames(er)<-c("tbp","ta","tocup","tdesoc")
colnames(vr)<-c("tbp","ta","tocup","tdesoc")

# Donde t corresponde a la matriz que contiene 12 filas una por cada mes,
# y 4 columnas pues se estiman 4 totales,
# v es la matriz de iguales dimensiones que la t y contiene las
# varianzas estimadas de los totales.
# er es una matriz que contiene para cada mes la estimación de los ratios y
# vr es una matriz con la estimación por mes de la varianza de los ratios.
```

Bibliografía

[Beltrami, 2005a] Beltrami, M. (2005a). *Diseño de la muestra para una Encuesta de Hogares Ampliada.* Manual técnico I.N.E.

[Beltrami, 2005b] Beltrami, M. (2005b). *Metodología aplicada para la Estratificación de Montevideo e Interior Urbano.* Manual tecnico I.N.E.

[Beltrami, 2006] Beltrami, M. (2006). *Encuesta Nacional de Hogares Ampliada.* Manual técnico I.N.E.

[Efron, 1993] Efron, B. y Tibshirani, R. (1993). *An introduction to the Bootstrap.*

[INE, 2004a] INE (2004a). *Censo Fase uno definiciones.* Manual técnico I.N.E.

[INE, 2004b] INE (2004b). *Diccionario-Censo-CPV05-Fase uno.* Manual técnico I.N.E.

[Lumley, 2004] Lumley, T. (2004). *Analysis of complex survey samples.* Journal of Statistical Software 9(1): 1-19.

[Lumley, 2006] Lumley, T. (2006). *survey: analysis of complex survey samples.* R package version 3.5-6.

[R, Development Core Team, 2008] R, Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[Särndal et al., 1992] Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York, Springer-Verlag.

[Tillé, 2007] Tillé, Y. y Matei, A. (2007). *sampling: Survey Sampling*. R package version 0.9.

[Wolter, 1985] Wolter, K. M. (1985). *Introduction to Variance Estimation*. New York, Springer-Verlag, Segunda Edición.