Distr.
RESTRINGIDA

LC/DEM/R.131
Serie A, N°227
octubre de 1991

ORIGINAL: ESPAÑOL

C E L A D E

Centro Latinoamericano de Demografía

MARCOS MULTIPLES: SU USO EN LA OBTENCION DE ESTIMACIONES EN AREAS PEQUEÑAS

Este documento fue preparado para ser presentado a la Reunión Técnica sobre "Nuevas Metodologías Estadísticas Aplicadas a la Demografía", organizada por el Centro Latinoamericano de Demografía, con la colaboración de la Universidad Nacional de Rosario, Argentina y del Centro Interamericano de Enseñanza de Estadística (CIENES), y que tuvo lugar los días 2, 3 y 4 de septiembre de 1991. Las opiniones expresadas en este trabajo son de exclusiva responsabilidad de sus autoras y pueden no coincidir con las de las instituciones mencionadas.

NACIONES UNIDAS UNITED NATIONS
CENTRO LATINOAMERICANO DE DEMOGRAFIA

REUNION TECNICA SOBRE "NUEVAS METODOLOGIAS ESTADISTICAS
APLICADAS A LA DEMOGRAFIA"
2,3 y 4 de Setiembre, 1991

MARCOS MULTIPLES: su uso en la obtención de estimaciones en áreas pequeñas

Fabiana del Popolo
Nora Ventroni

Santiago, Chile, 1991

NACIONES UNIDAS
CENTRO LATINOAMERICANO DE DEMOGRAFIA

MARCOS MULTIPLES: su uso en la obtención de estimaciones en áreas pequeñas

Fabiana del Popolo
Nora Ventroni

Santiago, Chile, 1991

TABLA DE CONTENIDO

INTRODUCCION

El proceso de desarrollo que en los últimos años han seguido los países latinoamericanos se ha visto estimulado a considerar la participación de los actores regionales y locales. Debido a ello, se pretende que las políticas planteadas tengan, por un lado la posibilidad de integrar las características reales de los sitios a afectar, y por otro, se asegure la participación de los actores locales desde el inicio de la planificación.

De esta manera, es necesario dotar a los organismos regionales y locales de aquellas herramientas que les permitan insertarse a dicho proceso. Entre las más relevantes se encuentra la información estadístico-demográfica. Esta información debe ser lo suficientemente confiable como para constituir la base sobre la cual se formulen y ejecuten políticas.

Ante esta necesidad de información surgen dos posibilidades, una es la de llevar a cabo una investigación directamente en el área de estudio a través, por ejemplo, de una encuesta por muestreo; otra alternativa es la de utilizar las fuentes de datos existentes utilizando metodologías que arrojen estimaciones confiables y así poder disminuir costos.

Una fuente de datos relevante para ello es el censo de población. En los años en que éste se realiza es factible obtener datos para diversos niveles de desagregación y además existen instrumentos computacionales que permiten el acceso a los niveles deseados (por ejemplo el programa REDATAM). Así, pueden realizarse estimaciones de las características deseadas a nivel local o de divisiones administrativas menores.

Pero no es posible contar con esta información en los períodos intercensales, donde la misma se obtiene a través de encuestas por muestreo arrojando estimaciones confiables sólo a nivel nacional y en algunos casos de divisiones administrativas mayores.

De esta forma nos encontramos ante el problema de obtener estimaciones en "áreas pequeñas". Se entiende por "área pequeña" a la subpoblación (grupos demográficos, subdivisiones geográficas, etc.) de una población que la contiene y de la cual se tiene información estadística razonablemente confiable a través de una muestra. Pero esta muestra resulta pequeña para arrojar estimaciones confiables en la mencionada subpoblación.

Ahora bien, para solucionar este problema existen dos alternativas: una se presenta cuando se cuenta con datos relacionados al área pequeña, y la otra cuando se tiene información directamente del área pequeña pero resulta incompleta (ésto en el sentido de que la información se refiere estrictamente a un subgrupo del área).

Para el caso en que se cuenta con información relacionada al área pequeña se han venido desarrollando metodologías[1/] con el fin de lograr estimaciones confiables. González (1978) describe el método del estimador sintético y dos tipos de modelos de regresión para la estimación de características poblacionales del área pequeña.

En el caso del estimador sintético, se conoce el total o el promedio de una determinada característica para el área dentro de la cual se encuentra el área pequeña. Entonces el estimador se define como una suma ponderado de dicha característica, determinándose los pesos a partir de

---

1/ Ver Bibliografía Complementaria

datos propios del área pequeña. Por ejemplo, supóngase que se desea estimar la tasa de desempleo para una comuna y se cuenta con la tasa de desempleo de una región que contiene a la comuna, el estimador sintético es la suma ponderada de la tasa de desempleo de la región y los pesos serían la distribución de la fuerza de trabajo en la comuna (por ocupación, industria, etc.) obtenida a partir de información censal.

Para utilizar los modelos de regresión debe conocerse información auxiliar correspondiente a las variables independientes involucradas en el modelo. Se propone, además, utilizar al estimador sintético como una de las variables independientes y se prueba el mejoramiento obtenido después de haber excluído los casos de valores extremos.

Entre otros estudios al respecto se encuestran los modelos propuestos por Fay y Herriot (1979), Dempster et al. (1981) y Battese y Fuller (1982). Dempster y Tomberlin (1980) proponen el uso de técnicas empíricas de Bayes para obtener estimaciones en área pequeñas; también MacGibbon y Tomberlin estudiaron estas técnicas para el caso de diseños muestrales multietápicos.

Bajo la otra alternativa, cuando se conoce información directamente del área pero ésta resulta incompleta, puede utilizarse la técnica de "marcos múltiples". La misma consiste en utilizar dos o más marcos muestrales que en su conjunto cubren al total de la población en estudio y de los cuales se extraen muestras en forma independiente.

Ahora, cómo se relaciona ésto con nuestro problema. Por un lado, dado que se trata de un "área pequeña" significa que existe una encuesta por muestreo a un nivel de agregación mayor al de dicha área. En este sentido, se tiene un marco muestral completo pero con un tamaño de muestra escaso. Por otro lado, se posee información directamente del área pero incompleta (por ejemplo, un listado de un subgrupo del área pequeña), la cual estaría suministrando un marco muestral incompleto. Así, se estaría en el caso particular en que un marco contiene al otro.

El presente trabajo evaluará la posibilidad de utilizar esta última técnica para obtener estimaciones en divisiones administrativas menores. Primeramente se describirá en qué consiste la metodología de "marcos múltiples". Luego se probará la misma en una comuna de Santiago, tomando información de la Encuesta Nacional de Empleo y de las fichas CAS.

1. MARCOS MULTIPLES

1.1 Introducción

Cuando se tiene que diseñar una encuesta por muestreo uno se encuentra con varios parámetros por definir: la población en estudio; el método de muestreo; el o los métodos de estimación; las unidades muestrales de observación, de análisis, etc.

Así, una tarea esencial es definir el "marco muestral", intentando que éste cubra al total de unidades de la población en estudio, y a partir del cual se selecciona la muestra.

Muchas veces uno define posibles marcos que cubren aproximadamente todas las unidades poblacionales, pero el diseño muestral a ser aplicado resulta muy costoso, por ejemplo la elaboración de listados de unidades especiales. No obstante, se dispone de otros marcos, cuyo muestreo resulta más económico, los cuales cubren sólo una conocida o aproximadamente conocida fracción de la población.

La técnica de "marcos múltiples" consiste, en el uso combinado de varios marcos, respecto de los cuales se han desarrollado diversos estudios para facilitar su aplicación.

1.2 Antecedentes

En el pasado, ocasionalmente, algunas encuestas utilizaron la técnica de "marcos múltiples" en sus diseños. Tal es el caso de la Encuesta de Agricultura llevada a cabo por el Bureau of the Census de Estados Unidos en 1960. En este caso se consideraron dos marcos: un marco basado en un diseño muestral por área y el otro marco consistía en el listado de establecimientos agropecuarios captados por el último Censo Agrícola (1959).

Anteriormente, "the Statistical Laboratory of the Iowa University" utilizó dos marcos muestrales en un pequeño estudio sobre "El Efecto de la Industrialización en la Agricultura", llevado a cabo por el Departamento de Economía y Sociología de la citada Universidad. Los marcos considerados eran: el marco de área usado habitualmente para muestrear establecimientos agropecuarios; y un listado de los empleados de la Clinton Motor Company, propietarios a su vez de establecimientos agropecuarios.

Ahora bien, dentro de las encuestas que utilizaron "marcos múltiples" en el pasado, merece especial atención The Sample Survey of Retail Stores (Encuesta de Comercios al por Menor) llevada a cabo por el Bureau of de Census en 1949. Podría considerarse a esta encuesta como la "pionera" en el uso de dicha técnica.

Hansen, Hurwitz and Madow (1953) describen las características de la misma, donde se combinan dos marcos muestrales: un listado y un marco de área.

Uno de los principales objetivos de la encuesta era obtener estimaciones del volumen anual de ventas de los negocios minoristas en los Estados Unidos, y la distribución de dichas ventas por tipo de negocio; sean éstos, almacenes de alimentos, mueblerías, farmacia, etc.

Otro propósito importante era medir el porcentaje de cambio en el volumen total de ventas, mes a mes y año a año, para todos los negocios y por categoría de los mismos.

Ciertas características importantes de los comercios influyeron en el diseño muestral, tal como la distribución asintótica de los negocios de acuerdo a la cantidad de ventas, el continuo cambio de comercios minoristas, la no respuesta en cuanto a la declaración de sus ventas, etc.

Del análisis de estas características, además de considerar otros puntos -como las fuentes de información disponibles-, se dedujo el diseño muestral que se detalla a continuación.

Los condados estadounidenses fueron considerados como unidades primarias de muestreo. Estos se agruparon en cuatro grupos de acuerdo al tamaño poblacional. Dentro de cada grupo se realizaron subagrupaciones para homogeneizar la población de acuerdo a criterios geográficos, económicos y/o poblacionales, los que pasan a constituir los estratos.

De cada estrato se seleccionó una unidad primaria con probabilidad proporcional a su tamaño.

Debido al costo de enumerar todos los negocios ubicados en las unidades primarias seleccionadas, se buscó la forma de seleccionar comercios dentro de cada condado, teniendo en cuenta las listas disponibles.

Así, se elaboró un listado de negocios a ser entrevistados mensualmente por correo, a partir de un criterio de asignación óptima 2/. Dado que la fuente de información para elaborar dicho listado no cubría exactamente todos los negocios existentes y que el criterio de tamaño óptimo permitía considerar las firmas más grandes, se realizó una muestra por área para las firmas que no estaban en el listado (generalmente firmas de menor envergadura).

Considerando una estimación a priori del error de muestreo y los fondos disponibles, se determinó el porcentaje de negocios a ser muestreado dentro de los que no estaban en la lista.

2/ Se realizó una aproximación al óptimo teniendo en cuenta que si existe estabilidad en el tamaño de las unidades a ser muestreadas y si la característica a ser estimada está altamente correlacionada con este tamaño, una buena aproximación al óptimo es considerar al desvío standar dentro de cada estrato proporcional al tamaño promedio de las unidades de dicho estrato.

Puesto que los costos de una enumeración completa de las unidades seleccionadas para determinar los comercios fuera de la lista eran demasiado altos y además este listado se volvería obsoleto rápidamente debido a los constantes cambios de negocios, se decidió seleccionar aleatoriamente con probabilidad conocida, segmentos (pequeñas unidades de tierra perfectamente delimitadas) dentro de cada unidad primaria seleccionada, y enumerar todos los comercios que no pertenecían al listado ubicados en estos segmentos.

Así, se tenía por un lado una lista especificada de firmas y por otro una muestra de área multietápica. Entonces un estimador insesgado, por ejemplo del total de ventas era dado por la suma de los totales de ventas obtenidos del listado y del muestreo de área, ponderados por el recíproco de sus respectivas probabilidades de selección.

Sin desmerecer el aporte valioso que las mencionadas encuestas dejaron, es cierto que no existía un desarrollo sistemático de la metodología de marcos múltiples. Podría decirse que es a partir de Hartley (1962) que se vienen desarrollando sucesivos estudios sobre el tema.

En concordancia con lo anterior es que la técnica de marcos múltiples que se abordará en el presente documento parte de la propuesta inicial hecha por Hartley.

1.3 Algunos conceptos

Antes de comenzar con los estimadores propuestos por Hartley, resulta conveniente hacer ciertas definiciones que serán utilizadas a lo largo de todo el documento.

Consideremos primero que se tienen dos marcos A y B, y que se selecciona una muestra independientemente de cada uno de ellos. Los diseños muestrales pueden ser diferentes en cada marco.

Además se supone que cada unidad de la población en estudio pertenece "al menos a uno de los dos marcos" y que es posible identificar para cada unidad en la muestra si pertenece o no al otro marco.

Así, las unidades de la muestra pueden dividirse en tres dominios. Llamaremos dominio "a" al de las unidades que pertenecen sólo al marco A; dominio "b" al de las unidades que pertenecen sólo al marco B; y dominio "ab" al de las unidades que pertenecen a ambos.

Las unidades en la población se dividen conceptualmente en estos tres dominios, pero se contemplan las alternativas de conocer o no el tamaño de cada dominio en la población.

1.4 Notación y simbología

Denotando como N al número de elementos en la población; N_A y N_B al número de elementos del marco A y del marco B respectivamente. Siendo además, N_{ab} el número de elementos incluídos en ambos marcos; N_a el número de elementos incluídos sólo en el marco A; y N_b el número de elementos incluídos sólo en el marco B.

Luego, el total de elementos en la población N se puede expresar de tres formas:

$$\begin{aligned} N &= N_a + N_b + N_{ab} \\ &= N_a + N_B \\ &= N_A + N_b \end{aligned} \qquad (1)$$

Suponiendo un muestreo aleatorio simple, sea n_A el número de elementos en la muestra del marco A y n_B el número de elementos en la muestra del marco B. Teniendo en cuenta los dominios antes definidos, denotaremos con n_a al número de elementos en la muestra del marco A contenidos en el dominio a; n_b el número de elementos en la muestra del marco B contenidos en el dominio b; n'_{ab} el número de elementos en la muestra del marco A contenido en el dominio ab; y n''_{ab} el número de elementos en la muestra del marco B contenido en el dominio ab.

Así:

$$n_A = n_a + n'_{ab} \qquad (2)$$

$$n_B = n_b + n''_{ab} \qquad (3)$$

Denotemos como Y al total poblacional de una determinada característica. Considerando los marcos A y B, Y_A e Y_B serán, respectivamente, los totales poblacionales de una determinada característica en cada marco. Entonces Y_{ab} es el total poblacional considerando el dominio ab; Y_a es el total poblacional en el dominio a; y Y_b es el total poblacional en el dominio b.

En la muestra, sea y_A el total muestral del marco A e y_B el total muestral del marco B. Al igual que con los valores poblacionales, y_a e y'_{ab} son los totales muestrales del marco A para los dominios a y ab respectivamente; y_b e y''_{ab} son los totales muestrales del marco B para los dominios a y ab respectivamente.

La misma simbología y clasificación, en los dominios considerados, se utiliza para las medias poblacionales y muestrales, reemplazando Y/y por $\bar{Y}/\bar{y}$.

1.5 Estimadores propuestos por Hartley

Simplificando el estudio a un muestreo simple al azar, y en el caso en que N_a, N_b y N_{ab} son conocidos, el estimador del total estará dado por:

$$\hat{Y} = N_a \bar{y}_a + N_{ab} (p \bar{y}'_{ab} + q \bar{y}''_{ab}) + N_b \bar{y}_b \qquad (4)$$

donde $(p + q) = 1$.

Una aproximación a la variancia del estimador es:

$$Var(\hat{Y}) \doteq N_A^2 / n_A [\sigma_a^2 (1-\alpha) + p^2 \sigma_{ab}^2 \alpha] + N_B^2 / n_B [\sigma_b^2 (1-ß) + q^2 \sigma_{ab}^2 ß] \qquad (5)$$

Siendo:

$$\alpha = N_{ab} / N_A$$

y

$$ß = N_{ab} / N_B$$

La corrección por finitud fue despreciada y σ_a^2, σ_{ab}^2, σ_b^2 son las variancias poblacionales dentro de cada dominio.

Para la determinación de un "p" óptimo se minimiza (5) en función de "p" y se obtiene que:

$$p_o = \alpha n_A / (\alpha n_A + ß n_B) \qquad (6)$$

En caso de que N_a, N_b y N_{ab} son desconocidos, el estimador para el total poblacional y la variancia de éste son:

$$\hat{Y} = N_A / n_A [y_a + p y'_{ab}] + N_B / n_B [y_b + q y''_{ab}] \qquad (7)$$

$$Var(\hat{Y}) \doteq N_A^2 / n_A [\sigma_a^2(1-\alpha) + p^2 \sigma_{ab}^2 \alpha + \alpha(1-\alpha)(\bar{y}_a - p\bar{y}'_{ab})^2] +$$
$$+ N_B^2 / n_B [\sigma_b^2(1-ß) + q^2 \sigma_{ab}^2 ß + ß(1-ß)(\bar{y}_b - q\bar{y}''_{ab})^2]$$

1.6 Estimadores propuestos por Lund

Lund (1967) sugiere algunas modificaciones a los estimadores propuestos por Hartley en orden de mejorar la eficiencia de los mismos. Aquí también se contemplan las alternativas de conocer o no N_{ab}, cuando sí se conocen N_A y N_B.

Caso de N_a N_{ab} y N_b conocidos

Como explica Lund, el procedimiento de Hartley para obtener el estimador del total no considera la partición aleatoria de los n_A y n_B elementos entre los dominios. Así, se pregunta si se obtendría alguna ganancia haciendo p como función de n'_{ab} y n''_{ab}.

Para obtener la solución, la variancia de (4) se expresa, según conocido teorema, como:

$$Var(\hat{Y}) = E\,[Var(\hat{Y} / n'_{ab}, n''_{ab})] + Var[E\,(\hat{Y} / n'_{ab}, n''_{ab})] \qquad (8)$$

donde la condición $/n'_{ab}, n''_{ab}$ representa la cantidad de elementos pertenecientes al dominio "ab" una vez obtenida la muestra.

El segundo término de (8) es igual a cero, puesto que $E(Y/n'_{ab}, n''_{ab})$ es igual al total poblacional para cualquier valor de "p" y por lo tanto la variancia es cero. Queda por deducir el primer término.

Despreciando la corrección por finitud,

$$Var\,(\hat{Y}/n'_{ab}, n''_{ab}) = N_a^2\,\sigma_a^2 / n_a + p^2\,N_{ab}^2\sigma_{ab}^2/ n'_{ab} + (1-p)^2\,N_{ab}^2\sigma_{ab}^2/ n''_{ab} + N_b^2\,\sigma_b^2 / n_b \qquad (9)$$

Minimizando (9) como función de "p" se obtiene la solución

$$p_o = n'_{ab} / (n'_{ab} + n''_{ab}) \qquad (10)$$

La variancia del estimador con este valor de "p", se obtiene reemplazando (10) en (9) y calculando el valor esperado de este último.

El estimador es entonces:

$$\hat{Y}_L = N_a \bar{y}_a + N_{ab} \bar{y}_{ab} + N_b \bar{y}_b \qquad (11)$$

donde:

$$\bar{y}_{ab} = (n'_{ab} \bar{y}'_{ab} + n''_{ab} \bar{y}''_{ab}) / (n'_{ab} + n''_{ab})$$

y la variancia del estimador es aproximadamente igual a:

$$\begin{aligned} Var(\hat{Y}_L) \doteq \; & N_A^2 (1-\alpha) \sigma_a^2 / n_A + N_A N_B \alpha \beta \sigma_{ab}^2 / (\alpha n_A + \beta n_B) + \\ & + N_B^2 (1-\beta) \sigma_b^2 / n_B \end{aligned} \qquad (12)$$

Para la demostración de la ecuación (12) se usará la aproximación de Taylor.

En la ecuación (a) se explicita otra forma de expresar la variancia del estimador condicionada a n'_{ab} y n''_{ab} , o sea:

$$\begin{aligned} Var\,(\hat{Y}/n'_{ab}, n''_{ab}) = \; & N_a^2 \sigma_a^2 / n_a + N_b^2 \sigma_b^2 / n_b + \\ & + \sigma_{ab}^2 N_{ab}^2 (p^2 / n'_{ab} + q^2 / n''_{ab}) \end{aligned} \qquad (a)$$

donde $(p + q) = 1$

Entonces, si $p_o = n'_{ab} / (n'_{ab} + n''_{ab})$, el tercer término de (a) se puede escribir como sigue:

$$\sigma_{ab}^2 N_{ab}^2 [n'_{ab} / (n'_{ab} + n''_{ab})^2 + n''_{ab} / (n'_{ab} + n''_{ab})^2] = \sigma_{ab}^2 N_{ab}^2 / (n'_{ab} + n''_{ab})$$

Luego:

$$\begin{aligned} Var\,(\hat{Y}/n'_{ab}, n''_{ab}) = \; & N_a^2 \sigma_a^2 / n_a + N_b^2 \sigma_b^2 / n_b + \\ & + \sigma_{ab}^2 N_{ab}^2 / (n'_{ab} + n''_{ab}) \end{aligned}$$

Aproximando Var $(\hat{Y} / n'_{ab}, n''_{ab})$ por Taylor en $E(n'_{ab}) = \alpha n_A$ y $E(n''_{ab}) = \beta n_B$ se obtiene la expresión:

$$\begin{aligned} Var\,(\hat{Y}/n'_{ab}, n''_{ab}) = \; & f(n'_{ab} , n''_{ab}) \\ \doteq \; & f(\alpha n_A , \beta n_B) + (n'_{ab} - \alpha n_A) f'n'_{ab}(\alpha n_A , \beta n_B) \\ & + (n''_{ab} - \beta n_B) f'n''_{ab} (\alpha n_A, \beta n_B) + \end{aligned}$$

$$+ \frac{1}{2} (n'_{ab} - \alpha n_A)^2 f''_{n'n'}(\alpha n_A, \beta n_B) +$$

$$+ \frac{1}{2} (n''_{ab} - \beta n_B)^2 f''_{n'' n''}(\alpha n_A, \beta n_B) +$$

$$+ (n'_{ab} - \alpha n_A)(n''_{ab} - \beta n_B) f''_{n' n''} (\alpha n_A, \beta n_B) \qquad (b)$$

$$+ \text{resto}$$

Considerando sólo el primer término de la serie de Taylor se tiene que

$$\begin{aligned} \text{Var}(\hat{Y}) &\doteq E[\text{Var}(\hat{Y}/n'_{ab}, n''_{ab})] \\ &= E[f(\alpha n_A, \beta n_B)] \\ &= N_A^2 (1-\alpha)\sigma_a^2 / n_A + N_B^2 (1-\beta) \sigma_b^2 / n_B + \\ &\quad + N_{ab}^2 \sigma_{ab}^2 / (\alpha n_A + \beta n_B) \end{aligned}$$

El orden de aproximación de (12) es el mismo que utiliza Hartley para la variancia del estimador del total. Así, se puede observar que ambas expresiones son idénticas, lo cual indica que ambos estimadores tienen igual eficiencia para este orden de aproximación.

La desviación de la aproximación de la variancia (12) respecto al verdadero valor de la variancia de Y_L fue estimada por Lund considerando hasta el segundo orden del desarrollo de Taylor. Entonces el primer término de (12) podría ser corregido multiplicándolo por $[1 + \alpha/(1-\alpha)n_A]$ y el tercero por $[1 + \beta/(1-\beta)n_B]$. La corrección para el segundo término es $[1 + \delta/(\alpha n_A + \beta n_B)]$ donde δ es el promedio ponderado de $(1-\alpha)$ y $(1-\beta)$, siendo los pesos αn_A y βn_B.

Demostración

Tomando la expresión (b) de la demostración anterior y expresando los términos de las derivadas en los puntos de aproximación, se tiene:

$$\begin{aligned} \text{Var}(\hat{Y}/n'_{ab}, n''_{ab}) = {} & N_A^2 (1-\alpha)\sigma_a^2 / n_A + N_B^2 (1-\beta) \sigma_b^2 / n_B + \\ & N_{ab}^2 \sigma_{ab}^2 / (\alpha n_A + \beta n_B) + (n'_{ab} - \alpha n_A) [N_a^2 \sigma_a^2 / (n_A (1-\alpha))^2 - \\ & N_{ab}^2 \sigma_{ab}^2 / (\alpha n_A + \beta n_B)^2] + (n''_{ab} - \beta n_B)[N_b^2 \sigma_b^2 / (n_B (1-\beta))^2 - \\ & N_{ab}^2 \sigma_{ab}^2 / (\alpha n_A + \beta n_B)^2] + (n'_{ab} - \alpha n_A)^2 [N_a^2 \sigma_a^2 / (n_A (1-\alpha))^3 \end{aligned}$$

$$+ N_{ab}^2\sigma_{ab}^2/(\alpha n_A + \beta n_B)^3] + (n''_{ab} - \beta n_B)^2 [N_b^2 \sigma_b^2/(n_B (1-\beta))^3$$

$$+ N_{ab}^2\sigma_{ab}^2/ (\alpha n_A + \beta n_B)^3] + 2(n'_{ab} - \alpha n_A)(n''_{ab} - \beta n_B) N_{ab}^2\sigma_{ab}^2$$

$$(\alpha n_A + \beta n_B)^{-3} + \text{resto}$$

$$= N_a^2 \sigma_a^2 / [n_A (1-\alpha)] [1 + (n'_{ab} - \alpha n_A) / (n_A (1-\alpha)) +$$

$$+ (n'_{ab} - \alpha n_A)^2 / (n_A (1-\alpha))^2] + N_b^2 \sigma_b^2/[n_B(1-\beta)]$$

$$[1 + (n''_{ab} - \beta n_B)/(n_B (1-\beta)) + (n''_{ab} - \beta n_B)^2/(n_B (1-\beta))^2] +$$

$$+ N_{ab}^2\sigma_{ab}^2/(\alpha n_A + \beta n_B)^2[(\alpha n_A + \beta n_B) - (n'_{ab}-\alpha n_A) -(n''_{ab}-\beta n_B) +$$

$$+ (n'_{ab} - \alpha n_A)^2 / (\alpha n_A + \beta n_B) + (n''_{ab} - \beta n_B)^2 / (\alpha n_A + \beta n_B)+$$

$$+ 2 (n'_{ab} - \alpha n_A)(n''_{ab} - \beta n_B) / (\alpha n_A + \beta n_B)]$$

Luego, aplicando esperanza:

$$V(\hat{Y}_L) = E[V(\hat{Y} / n'_{ab} , n''_{ab})$$

$$= N_a^2 \sigma_a^2 / (n_A (1-\alpha)) + 0 +$$

$$N_a^2 \sigma_a^2 / (n_A (1-\alpha))^3 V(n'_{ab}) + N_b^2 \sigma_b^2 / (n_B (1-\beta)) + 0 +$$

$$N_b^2 \sigma_b^2/(n_B (1-\beta))^3 V(n''_{ab}) + N_{ab}^2\sigma_{ab}^2/ (\alpha n_A + \beta n_B)^2 [\alpha n_A + \beta n_B$$

$$+ [V(n'_{ab}) + V(n''_{ab})] / (\alpha n_A + \beta n_B)]$$

$$= N_a^2 \sigma_a^2 / (n_A (1-\alpha)) [1+ \alpha(1-\alpha) n_A / n_A^2(1-\alpha)^2] +$$

$$+ N_b^2 \sigma_b^2 /(n_B (1-\beta)) [1 + \beta(1-\beta) n_B / n_B^2 (1-\beta)^2] +$$

$$+ N_{ab}^2\sigma_{ab}^2/(\alpha n_A + \beta n_B) [1 + (\alpha n_A (1-\alpha) + \beta n_B (1 - \beta))$$

$$/ (\alpha n_A + \beta n_B)^2]$$

$$= N_a^2 \sigma_a^2 / [n_A (1-\alpha)] [1 + \alpha / ((1-\alpha) n_A)] +$$

$$+ N_b^2 \sigma_b^2 / [n_B (1-\beta)] [1 + \beta / ((1-\beta) n_B)] +$$

$$+ N_{ab}^2\sigma_{ab}^2/ (\alpha n_A + \beta n_B) [1 + \delta/ (\alpha n_A + \beta n_B)] =$$

$$= N_a^2 \sigma_a^2/[n_A(1-\alpha)][1 + \alpha/((1-\alpha)n_A)] + N_b^2 \sigma_b^2/[n_B (1-\beta)]$$
$$[1 + \beta/((1-\beta) n_B)] + N_{ab}^2\sigma_{ab}^2/ (\alpha n_A + \beta n_B)$$
$$[1 + \delta / (\alpha n_A + \beta n_B)]$$

donde $\delta = [\alpha n_A (1-\alpha) + \beta n_B (1-\beta)] / (\alpha n_A + \beta n_B)$

Nota 1: En el cálculo de las variancias de n'_{ab} y n''_{ab} se despreció el factor de corrección por finitud.

En el caso de la aproximación de la variancia del total propuesta por Hartley, es válida la misma corrección que propone Lund para el primer y tercer término. Para el segundo término δ se convierte en la suma de $(1-\alpha)$ y $(1-\beta)$.

Luego, la variancia aproximada (12) (o (5)) es razonablemente exacta y la ganancia en eficiencia al usar el p dado por Lund en lugar del dado por Hartley, resulta poco significativa excepto para muestras extremadamente pequeñas.

En cuanto al problema de asignación óptima de la muestra entre los marcos, una solución general puede expresarse por el sistema iterativo:

$$r_1 = (c_B/c_A \beta/\alpha)^{1/2}$$
$$r_{i+1}^2 = c_B/c_A (\beta/\alpha)^2 [(r_i + \beta/\alpha)^2 (1-\alpha)\sigma_a^2 + r_i^2 \sigma_{ab}^2]/[(r_i +$$
$$\beta/\alpha)^2 (1-\beta)\sigma_b^2 + (\beta/\alpha)^2 \beta\sigma_{ab}^2]$$

donde c_A y c_B son los costos unitarios muestrales de los marcos A y B respectivamente y $r = n_A/n_B$.

Probando el sistema para distintos valores de los parámetros se observó que pocas iteraciones son necesarias en la mayoría de los casos (Ver Anexo 1).

Para determinar la sensibilidad del estimador cuando no se tiene la asignación óptima, el valor óptimo de r y la correspondiente variancia se computaron para un amplio rango de valores de los parámetros. Se realizó una comparación con las variancias correspondientes a desviaciones de un 10 por ciento en ambas direcciones del óptimo. La variancia se incrementó en más de un uno por ciento sólo en muy pocos casos.

Caso de N_a N_{ab} y N_b desconocidos

En este caso, primeramente Lund determina estimadores para N_a , N_{ab} y N_b a partir de datos muestrales y de los tamaños conocidos N_A y N_B. Estimadores insesgados de N_a y N_b son, respectivamente, $N_A(n_a/n_A)$ y $N_B(n_b/n_B)$. Dos estimadores insesgados de N_{ab} son $N_A(n'_{ab}/n_A)$ y $N_B(n''_{ab}/n_B)$. Ponderando estos estimadores por p y (1-p) y sustituyendo en (11), se obtiene un estimador insesgado para el total:

$$\hat{Y} = N_A/n_A n_a \bar{y}_a + [N_A/n_A n'_{ab} p + N_B/n_B n''_{ab}(1-p)]\, \bar{y}_{ab} + N_B/n_B n_b\, \bar{y}_b \qquad (13)$$

donde $\bar{y}_{ab}$ se define como antes, la media muestral de todos los elementos del dominio ab.

Una aproximación a la variancia es:

$$\begin{aligned} Var(\hat{Y}) = \; & N_A^2 (1-\alpha)\, \sigma_a^2 / n_A + N_A N_B\, \alpha\beta\, \sigma_{ab}^2 / (\alpha\, n_A + \beta\, n_B) \\ & + N_B^2 (1-\beta)\, \sigma_b^2 / n_B + \\ & + N_A^2 (1-\alpha)\alpha / n_A\, [\bar{y}_a - p\bar{y}_{ab}]^2 + \\ & + N_B^2 (1-\beta)\beta / n_B\, [\bar{y}_b - (1-p)\bar{y}_{ab}]^2 \end{aligned} \qquad (14)$$

Los dos últimos términos representan el incremento en la variancia por no conocer N_a, N_b y N_{ab}. Estos términos hacen un aporte significativo en (14) excepto cuando el dominio ab contiene a casi todos los elementos de los marcos o cuando es relativamente pequeño.

Demostración de (14)

Como en el caso de N_a, N_b y N_{ab} conocido, se parte de que:

$$Var(\hat{Y}) = E[Var(\hat{Y}/n'_{ab}, n''_{ab})] + Var[E(\hat{Y}/n'_{ab}, n''_{ab})] \qquad (c)$$

Comenzando por la primera expresión de (c) se tiene que

$$\begin{aligned} Var(\hat{Y}/n'_{ab}, n''_{ab}) &= f(n'_{ab}, n''_{ab}) \\ &= (N_A^2/n_A^2\, n_a \sigma_a^2 + [N_A/n_A n'_{ab} p + N_B/n_B n''_{ab}\, (1-p)]^2 \\ & \quad \sigma_{ab}^2 / (n'_{ab} + n''_{ab}) + N_B^2 / n_B^2\, n_b\, \sigma_b^2 \end{aligned}$$

Desarrollando por Taylor $\mathrm{Var}(\hat{Y}/n'_{ab}, n''_{ab})$ en $(\alpha n_A, \beta n_B)$ y considerando sólo el primer término de la serie $(f(\alpha n_A, \beta n_B))$ se obtiene:

$$\mathrm{Var}(\hat{Y}/n'_{ab}, n''_{ab}) = N_A^2 / n_A (1-\alpha)\, \sigma_a^2 + N_B^2 / n_B (1-\beta)\, \sigma_b^2 + \\ + [N_A/n_A \alpha n_A p + N_B/n_B \beta n_B - N_B/n_B\, \beta n_B p]^2 \sigma_{ab}^2 / \\ /(\alpha n_A + \beta n_B)$$

Como la expresión anterior es una constante resulta igual a su esperanza, y además sabiendo que $\alpha N_A = \beta N_B = N_{ab}$ se tiene:

$$E[\mathrm{Var}(\hat{Y}/n'_{ab}, n''_{ab})] \doteq N_A^2/n_A(1-\alpha)\sigma_a^2 + N_A N_B \alpha\beta\sigma_{ab}^2/(\alpha n_A + \beta n_B) \\ + N_B^2 / n_B (1-\beta)\, \sigma_b^2 \qquad (d)$$

Ahora bien,

$$\begin{aligned} E(\hat{Y}/n'_{ab}, n''_{ab}) &= E[(\hat{N}_a\bar{y}_a + \hat{N}_b\bar{y}_b + \hat{N}_{ab}\bar{y}_{ab}) / n'_{ab}, n''_{ab}] \\ &= E(\hat{N}_a\bar{y}_a/n'_{ab}, n''_{ab}) + E(\hat{N}_b\bar{y}_b/n'_{ab}, n''_{ab}) + \\ &\quad + E(\hat{N}_{ab}\bar{y}_{ab}/n'_{ab}, n''_{ab}) \\ &= N_A / n_A\, n_a\, E(\bar{y}_a/n'_{ab}, n''_{ab}) + N_B / n_B\, n_b\, E(\bar{y}_b/n'_{ab}, n''_{ab}) \\ &\quad + \hat{N}_{ab}\, E(\bar{y}_{ab}/n'_{ab}, n''_{ab}) \\ &= N_A / n_A\, n_a\, \bar{Y}_a + N_B / n_B\, n_b\, \bar{Y}_b + N_A / n_A\, n'_{ab}\, p\, \bar{Y}_{ab} + \\ &\quad N_B / n_B\, n''_{ab}\, \bar{Y}_{ab}\,(1-p) \\ &= N_A\bar{Y}_a - N_A/n_A\, n'_{ab}\bar{Y}_a + N_B\bar{Y}_b - N_B/n_B\, n''_{ab}\bar{Y}_b + N_A/n_A \\ &\quad n'_{ab}\, p\, \bar{Y}_{ab} + N_B/n_B\, n''_{ab}\, \bar{Y}_{ab}\,(1-p) \\ &= N_A\bar{Y}_a + N_B\bar{Y}_b - N_A/n_A\, n'_{ab}(\bar{Y}_a - p\bar{Y}_{ab}) - N_B/n_B n''_{ab}(\bar{Y}_b - (1-p)\,\bar{Y}_{ab}) \end{aligned}$$

Luego,

$$\mathrm{Var}[E(\hat{Y}/n'_{ab}, n''_{ab})] = N_A^2/n_A^2\, n_A\alpha(1-\alpha)\,(\bar{Y}_a - p\bar{Y}_{ab})^2 + \\ + N_B^2/n_B^2\, n_B\beta(1-\beta)\,(\bar{Y}_b - (1-p)\bar{Y}_{ab})^2 \qquad (e)$$

Entonces, reemplazando (d) y (e) en (c) se obtiene la variancia aproximada (14).

El grado de aproximación en la variancia (14) no es el mismo que la (12) o la (5).

En este caso todos los términos son exactos excepto el segundo. Nuevamente, estimando la verdadera $\mathrm{Var}(\hat{Y}/n'_{ab}, n''_{ab})$ a través del segundo orden de la serie de Taylor, y examinando el segundo término, Lund sugiere corregirlo multiplicándolo por un valor menor o igual a $[1 + (1-\alpha)/\alpha n_A + (1-\beta)/\beta n_B]$. Así, la aproximación resulta razonablemente exacta en prácticamente todos los casos.

Minimizando (14) como función de p, n_A y n_B sujeto a la ecuación de costo

$$\text{Costo total} = n_A c_A + n_B c_B ,$$

donde c_A y c_B son los costos unitarios muestrales para cada marco respectivamente, se obtiene que:

$$p_o = [\, N_A(1-\alpha)/n_A \bar{Y}_a + N_B(1-\beta)/n_B (\bar{Y}_{ab} - \bar{Y}_b)]/[\, (N_A(1-\alpha)/n_A + N_B(1-\beta)/n_B)\, \bar{Y}_{ab} \,] \qquad (15)$$

La asignación óptima de la muestra en los dos marcos se expresa nuevamente como un sistema iterativo

$$r_1 = [c_B/c_A(\beta/\alpha)]^{1/2}$$

$$r_{i+1}^2 = c_B/c_A\, (\beta/\alpha)^2 \{ (1-\alpha)\sigma_a^2 + r_i^2\, \alpha\, \sigma_{ab}^2/ (r_i + \beta/\alpha)^2 + + r_i^2\, \alpha\, (1-\alpha)\, (\bar{Y}_a + \bar{Y}_b - \bar{Y}_{ab})^2 / [r_i + (\beta\, (1-\alpha) / \alpha (1-\beta))]^2 \} * \{ (1-\beta)\sigma_b^2 + (\beta/\alpha)^2\, \beta\, \sigma_{ab}^2 / (r_i + \beta/\alpha)^2 + + [\beta/\alpha\, (1-\alpha)/(1-\beta)]^2\, \beta(1-\beta)\, (\bar{Y}_a + \bar{Y}_b - \bar{Y}_{ab})^2 / [r_i + \beta/\alpha\, (1-\alpha)/(1-\beta)]^2 \}^{-1}$$

Utilizando el sistema con varios valores de los parámetros, se observa que normalmente se requieren pocas iteraciones.

Al igual que en el caso de N_a, N_b y N_{ab} conocidos, una investigación empírica para medir la eficiencia del estimador mostró que éste es bastante insensible a desviaciones moderadas de la asignación óptima.

Ahora, para determinar el valor de "p" uno puede usar las estimaciones de los parámetros usadas en la determinación de la asignación óptima. Sin embargo, se obtendría una ganancia en eficiencia si se utilizaran los datos de la muestra. Un estimador del "p" óptimo usando información muestral es:

$$\hat{p} = [\, N_A/n_A^2\, n_a\bar{y}_a + N_B/n_B^2\, n_b(\bar{y}_{ab} - \bar{y}_b)\,] \,/\, [\, (N_A/n_A^2\, n_a + N_B/n_B^2\, n_b)\, \bar{y}_{ab}\,] \qquad (16)$$

Obviamente, el uso de (16) hace que el estimador (13) sea sesgado, ya que "p" es ahora función de n'_{ab} y n''_{ab}. No obstante, puede probarse que el grado del sesgo se puede considerar poco significativo. El valor esperado de (13) con el "p" dado por (16), fue aproximado utilizando el desarrollo en serie de Taylor hasta el segundo orden. En base a esta aproximación se encontró que el sesgo de (13) puede reducirse multiplicando al segundo término de dicho estimador por el factor $[1-\delta]$, donde δ es un promedio ponderado de $(1/n_A)$ y $(1/n_B)$, siendo los pesos $N_A(1-\alpha)/n_A$ y $N_B(1-\beta)/n_B$.

Finalmente, puede probarse que el estimador propuesto por Lund para el caso de N_a, N_b y N_{ab} desconocidos, posee igual o mayor eficiencia que el estimador propuesto por Hartley.

1.7 Estimadores propuestos por otros autores

Wayne Füller y Leon Burmeister (1972) abordan el tema de "marcos múltiples" considerando que sólo se conocen N_A y N_B, por lo tanto comienzan por dar alternativas para la estimación de N_{ab}.

Se trata, primeramente, el caso en que no se identifican en la muestra los elementos comunes a ambos marcos. En este sentido y como se deduce de (2), Hartley propone el estimador:

$$\hat{N}_{ab,H} = p\, n'_{ab}\, N_A / n_A + q\, n''_{ab}\, N_B / n_B\,, \qquad (17)$$

donde "p" y "q" son valores fijos, con la condición de que $(p + q) = 1$.

La variancia de este estimador viene dada por:

$$\mathrm{Var}(\hat{N}_{ab,H}) = p^2 f_A^2\, \mathrm{Var}(n'_{ab}) + q^2 f_B^{-2}\, \mathrm{Var}(n''_{ab}) \qquad (18)$$

donde $f_A = n_A / N_A$ y $f_B = n_B / N_B$

Bajo distribución hipergeométrica, se deducen las variancias de n'_{ab} y n''_{ab}. Reemplazando éstas en (18), se obtiene el "p" que minimiza la variancia y entonces ésta se transforma en:

$$Var_o(\hat{N}_{ab,H}) = N_{ab}N_aN_bg_Ag_B / (n_AN_bg_B + n_BN_ag_A)$$

donde $g_A = (N_A-n_A)/(N_A-1)$ y $g_B = (N_B-n_B)/(N_B-1)$

Ahora bien, el estimador propuesto por Hartley, con "p" fijo, no siempre cae en el intervalo de valores posibles de N_{ab}. Luego, se propone un estimador que sí lo hace, basándose sólo en los datos de la muestra.

Entonces, reemplazando la expresión del "p" que minimiza la variancia en (17), se llega a la expresión cuadrática en N_{ab}

$$[n_Ag_B + n_Bg_A] \hat{N}^2_{ab,m} - [n_AN_Bg_B + n_BN_Ag_A + n'_{ab}N_Ag_B + n''_{ab}N_Bg_A] \hat{N}_{ab,m} + [n'_{ab}g_B + n''_{ab}g_A] N_A N_B = 0 \quad (19)$$

Se puede demostrar que (19) arroja siempre raíces reales y que la raíz menor está en el intervalo cero y mínimo (N_A, N_B). Por lo tanto, el estimador de N_{ab} es la raíz más pequeña de la ecuación. La variancia de este estimador es igual a la de Hartley en la aproximación realizada para su obtención.

Además, el orden del sesgo de $\hat{N}_{ab,m}$ resulta algo menor comparado con el de su variancia, por ello se recomienda su uso cuando no resulta posible identificar los elementos de la muestra en ambos marcos.

En el caso en que los elementos duplicados en la muestra son identificados, esto es, se comparan los n'_{ab} elementos del marco A con los n''_{ab} elementos del marco B y se encuentra que n_d de estos elementos son comunes, se proponen tres estimadores de N_{ab}.

Uno de estos estimadores es contruído como una combinación lineal de tres estimadores insesgados:

$$\hat{N}_{ab,l} = p\, f_A^{-1}\, n'_{ab} + r\, f_B^{-1}\, n''_{ab} + (1 - p - r)\, f_A^{-1}\, f_B^{-1}\, n_d \quad (20)$$

La variancia de $\hat{N}_{ab,l}$ es

$$\begin{aligned} Var(\hat{N}_{ab,l}) = {} & p^2\, g_A\, n_A^{-1}\, N_a\, N_{ab} + r^2\, g_B\, n_B^{-1}\, N_b\, N_{ab} + (1 - p - r) \\ & [g_A\, n_A^{-1}\, N_a\, N_{ab} + g_B\, n_B^{-1}\, N_b\, N_{ab} + N_{ab}\, f_A^{1}\, f_B^{-1}\, g_A\, g_B \\ & (1 - 1/N_A - 1/N_B + N_{ab}/N_A/N_B)] + 2p(1- r- p)\, g_A \\ & n_A^{-1}\, N_a\, N_{ab} + 2\, r\, (1- r - p)\, g_B\, n_B^{-1}\, N_b\, N_{ab} \end{aligned} \quad (21)$$

Una vez obtenidos "p" y "r" que minimizan $Var(\hat{N}_{ab,I})$ se llega a una expresión de la variancia de la siguiente forma:

$$Var_o(\hat{N}_{ab,I}) = N_{ab}N_aN_b\, g_Ag_B /(f_Af_B\, N_a\, N_b + n_B\, N_a\, g_A + n_A\, N_b\, g_B) \qquad (22)$$

Puede demostrarse que $Var_o(\hat{N}_{ab,I})$ resulta inferior a la variancia obtenida por Hartley, cuando no se identifican los elementos repetidos en la muestra. Esta reducción depende de las fracciones muestrales y de la proporción de los elementos de la población que están en el dominio ab.

Otro estimador propuesto es el estimador de máxima verosimilitud de N_{ab}. La probabilidad de obtener una muestra con un número dado de elementos seleccionados del dominio ab es:

$$L(n'_{ab}, n''_{ab}, n_d ; N_{ab}) = L(n'_{ab} ; N_{ab})\, L(n''_{ab}, n_d ; N_{ab})$$

$$= \frac{\binom{N_a}{n_a}\binom{N_{ab}}{n'_{ab}}}{\binom{N_A}{n_A}} \cdot \frac{\binom{N_b}{n_b}\binom{N_{ab}-n'_{ab}}{n''_{ab}-n_d}\binom{n'_{ab}}{n_d}}{\binom{N_B}{n_B}}$$

Haciendo la razón entre $L(n'_{ab}, n''_{ab}, n_d ; N_{ab})$ y $L(n'_{ab}, n''_{ab}, n_d ; N_{ab}-1)$ igual a 1, se obtiene el estimador máximo verosimil de N_{ab} resolviendo la ecuación cuadrática

$$(n_a + n_{ab} + n_b)\, \hat{N}^2_{ab,mv} - [n_a\, N_B + n_{ab}\, (N_A + N_B) + n_b\, N_A - n_a\, n_b]$$
$$\hat{N}_{ab,mv} + n_{ab}\, N_A\, N_B = 0$$

donde $n_{ab} = n'_{ab} + n''_{ab} - n_d$

Se demuestra que la estimación de N_{ab} es dada por la raíz izquierda de la ecuación anterior. La variancia de este estimador es aproximadamente igual a (22).

El estimador de máxima verosimilitud puede usarse para obtener una estimación de los pesos óptimos calculados a partir de la minimización de (21) (tanto "p" como "r" quedan en función de N_{ab}). Así, un estimador de N_{ab} podría obtenerse reemplazando estos pesos estimados en (20). Con ésto, se logra una reducción en el sesgo del estimador respecto al de máxima verosimilitud, pero la misma no es significativa.

Por último se plantea el estimador de Horvitz-Thompson, estimador insesgado de N_{ab}. Este es:

$$\hat{N}_{ab,HT} = (n'_{ab} + n''_{ab} - n_d) / (f_A + f_B - f_A f_B)$$

siendo la variancia del estimador:

$$\begin{aligned} Var(\hat{N}_{ab,HT}) = {} & (f_A + f_B - f_A f_B)^{-2} [\, (1-f_A)(1-f_B)^2 n_A \alpha(1-\alpha) + \\ & + (1-f_B)(1-f_A)^2 n_B \beta(1-\beta) + N_{ab} f_A f_B (1-f_A) \\ & (1-f_B) \,] + O(1) \end{aligned}$$

En general, la variancia de este estimador resulta aproximadamente similar a la variancia del estimador máximo verosímil.

En cuanto a la estimación del total de una característica Y, una vez estimado N_{ab} y en el caso de que no se identifican los elementos comunes en la muestra, se considera:

$$\hat{Y} = (N_A - \hat{N}_{ab})\, \bar{y}_a + \hat{N}_{ab}\, \bar{y}_{ab} + (N_B - \hat{N}_{ab})\, \bar{y}_b \qquad (23)$$

donde:

$$\bar{y}_{ab} = [p\, \bar{y}'_{ab} + (1-p)\, \bar{y}''_{ab}]$$

y

$$p = n'_{ab}(1-f_B) / [n'_{ab}(1-f_B) + n''_{ab}(1-f_A)]$$

Claramente se deduce que los pesos se determinaron minimizando la variancia de $\bar{y}_{ab}$.

La variancia de este estimador es aproximadamente igual a:

$$\begin{aligned} Var(\hat{Y}) = {} & N_a(f_A^{-1} - 1)\, \sigma_a^2 + [\, (1-f_B)f_A + (1-f_A)f_B \,]^{-1} \\ & (1-f_A)(1-f_B) \\ & N_{ab}\, \delta_{ab} + N_b(f_B^{-1} - 1)\, \sigma_b^2 + (\bar{Y}_{ab} - \bar{Y}_a - \bar{Y}_b)^2 \\ & N_{ab}\, N_a\, N_b\, g_A\, g_B / (n_A\, N_b\, g_B + n_B\, N_a\, g_A) \end{aligned}$$

En el caso en que los elementos muestrales comunes son identificados el estimador del total resulta igual a (23), excepto que:

$$\bar{y}_{ab} = 1/n_{ab} \sum_{i=1}^{n_{ab}} y_i$$

donde $n_{ab} = n'_{ab} + n''_{ab} - n_d$

Este estimador posee menor variancia que el estimador que no identifica los elementos. La reducción en la variancia depende de la fracción de muestreo.

Debido a la variedad de situaciones que pueden presentarse al intentar aplicar esta metodología, como así tambien al interés por mejorar la eficiencia de los estimadores, se han desarrollado, además de los ya comentados, diversos trabajos sobre "marcos múltiples".

Dentro de los trabajos que hemos revisado se encuentra el de **Bankier** (1983). Este estudio propone estimadores alternativos a los de Hartley. Restringe el análisis al caso en que una muestra aleatoria simple estratificada se extrae independientemente de cada marco, pero la técnica de estimación puede extenderse a diseños más complejos. Se asume que la unidad seleccionada en más de una muestra puede ser identificada.

La técnica de estimación propuesta por Bankier puede proveer estimadores con variancias significativamente menores. Además, computacionalmente y algebraicamente es fácil extender el método a un mayor número de marcos.

Primeramente la metodología es desarrollada para obtener estimaciones basadas en dos muestras independientes extraídas del mismo marco, donde las estratificaciones pueden ser distintas. Luego se muestra como dicho resultado puede extenderse al caso de marcos múltiples.

Entonces, se supone que g es el número de estratos de donde se saca la muestra A y h es el número de estratos de donde se extrae la muestra B. De esta forma puede obtenerse una tabla g por h, de elementos que caen en el estrato g de A y en el h de B. Así, el valor de la característica puede calcularse en cada estrato cruzado y sumar a través de éstos. Una primera estimación se realiza a través del estimador de Horvitz Thompson.

Luego plantea la posibilidad de aprovechar información auxiliar para obtener variancias más pequeñas. Si se tiene dicha información para cada estrato cruzado (gh) entonces se usa un estimador de razón separado. En el caso de no poseerla o alguno de los n_{gh} es cero o cercano a cero entonces no se puede utilizar el estimador separado y se usa un estimador combinado que suma a través de todos los estratos. Ahora, si no puede aplicarse el estimador separado pero se conocen los totales de cada estrato ($X_{g.}$ y $X_{.h}$), se usa el estimador de razón de rangos (procedimiento iterativo), el cual provee menor variancia que el estimador combinado. La forma de este estimador es:

$$\hat{Y}^{(p)} = \sum_g \sum_h w_{gh}^{(p)} y_{gh}$$

donde, $w_{gh}^{(0)} = 1 / [1- (1-f_{Ag})(1-f_{Bh})]$

$$w_{gh}^{(p)} = w_{gh}^{(p-1)} X_{g.}/\hat{X}_{g.}^{(p-1)} \quad \text{si p es impar}$$

$$= w_{gh}^{(p-1)} X_{.h}/\hat{X}_{.h}^{(p-1)} \quad \text{si p es par}$$

$$\hat{X}_{gh}^{(p-1)} = w_{gh}^{(p-1)} x_{gh}$$

Luego, la fórmula de la variancia es:

$$V(Y^{(p)}) = V \left(\sum_g \sum_h w_{gh}^{(p-1)} (y_{gh} - R^* x_{gh})\right),$$

donde $R^* = R_{g.} = Y_{g.}/X_{g.}$ si p es par
$= R_{.h} = Y_{.h}/X_{.h}$ si p es impar

El uso de estos estimadores se generaliza para el caso de dos marcos, considerando en ambos un estrato adicional que cubre la parte del otro marco que no quedó superpuesta, y del cual se sacan cero elementos en la muestra.

En el caso de conocer N_{gh} considera otro estimador con un p para cada estrato cruzado. Esto es, se tiene una estimación del estrato cruzado a partir de A y otra a partir de B; entonces el estimador definitivo es una ponderación de éstos.

Si N_{gh} no se conoce o alguno de los n_{ghA} o n_{ghB} es cero o cercano a cero, no es posible usar el estimador anterior; se propone uno similar al de Hartley que difiere en los pesos.

Por último plantea estimadores de razón con $X_{gh} = N_{gh}$, lo cual demuestra con un ejemplo numérico que reduce notablemente la variancia.

Así, siendo siete los estimadores presentados, se comparan sus variancias a través de un ejemplo numérico.

El estudio concluye que en el caso de un muestreo estratificado, utilizando dos marcos, se obtiene una variancia considerablemente menor con el estimador iterativo de razón que con el sugerido por Hartley.

Bosecker and Ford (1976) extienden el estimador desarrollado por Hartley para probar si obtiene ventajas al estratificar dentro del dominio superpuesto. Supone que se tiene un marco de área y un listado. El marco constituído por el listado está estratificado y entonces los estratos del dominio superpuesto se definen a partir de éste.

Como el listado está dentro del marco de área, el estimador de Hartley queda:

$$\hat{Y} = \hat{Y}_a + p\,\hat{Y}'_{ab} + q\,\hat{Y}''_{ab}$$

donde $Y''_{ab} = Y_l$ es el total estimado a partir del listado.

Haciendo la extensión al caso de un muestreo estratificado en el listado se tiene que:

$$\hat{Y}_l = \hat{Y}_{l(1)} + ... + \hat{Y}_{l(k)}$$

donde $Y_{l(h)}$ es la estimación del total en el h-ésimo estrato del listado y k es el número de estratos del mismo.

Luego, para hacer un pareo entre ambos marcos se estratifica el área superpuesta.

Entonces :

$$\hat{Y}'_{ab} = \hat{Y}'_{ab(1)} + ... + \hat{Y}'_{ab(k)}$$

Así,

$$\hat{Y}_{estrato} = \hat{Y}_a + \sum_{h=1}^{k} p_h * \hat{Y}'_{ab(h)} + \sum_{h=1}^{k} q_h * \hat{Y}_{l(h)}$$

donde $p_h + q_h = 1$ para cada h.

También puede escribirse:

$$\hat{Y}_{estrato} = \hat{Y}_{area} + \sum^{k} q_h * [\hat{Y}_{l(h)} - \hat{Y}_{a(h)}]$$

donde: $\hat{Y}_{area} = \hat{Y}^a + \hat{Y}'_{ab}$

Luego demuestra que:

$$Var(\hat{Y}_{estrato}) < Var(\hat{Y}_{area})$$

$$Var(\hat{Y}_{estrato}) \leq Var(\hat{Y}_H)$$

No obstante, a través de un ejemplo numérico demuestra que la ganancia al estratificar no es mucha comparada con el de Hartley.

En cuanto a la aplicabilidad en casos concretos de la técnica de "marcos múltiples", **Kalton & Anderson** (1986) consideraron entre posibles opciones la de utilizar esta metodología para estimar características en "poblaciones raras" (pequeño subgrupo de una población total).

Puesto que en general no se dispone de una lista completa de la población rara, sí puede contarse con listas parciales que registran casos raros (como ser registros de hospitales sobre determinadas enfermedades). Estas pueden complementarse con un muestreo de áreas para dar representatividad a aquéllos no incluídos en las listas.

Ahora bien, puede suceder que una persona esté incluída en más de una lista y en el marco de la muestra de área. Hay dos caminos a seguir: uno, redefinir los marcos a fin de no tener elementos superpuestos; otro, hacer compensaciones en el análisis.

En el primer caso, se crea una lista combinada con todos los listados disponibles, eliminando las duplicaciones de elementos y se diseña una muestra de área para obtener información de aquéllos que no figuran en las listas.

Un ejemplo de lo anterior es el estudio sobre población sorda realizado en Washington en 1960. Para ello, se formó una lista, lo más completa posible, de personas sordas a partir de organizaciones para sordos, escuelas para sordos, informantes, etc. Adicionalmente, se utilizó una muestra de área con una fracción de muestreo total de 1/120 para cubrir a las personas sordas no representadas en el listado. Nótese que el problema de identificar a los individuos de ambos marcos (lista y área) en la muestra se resuelve en forma relativamente fácil durante el trabajo de campo.

La otra alternativa consiste en aplicar ajustes de ponderación en las medias, de tal forma de compensar la posible inclusión en la muestra de los elementos de la población desde los distintos marcos considerados. Los estudios al respecto parten desde el estimador de marcos múltiples propuesto por Hartley (1962, 1974). Estos, como se detalló anteriormente, tienen la ventaja de que permiten determinar la fracción de muestreo óptima y el óptimo valor del peso (p) a ser utilizado.

Es interesante considerar el caso presentado en the Agriculture Division of Statistics Canada, ya que tiene cierta relación con el interés práctico de la presente investigación, si bien se desarrolla en otra disciplina. Dicha institución conduce anualmente la Encuesta de Enumeración Agrícola (AES), brindando estimaciones para producción de granos, existencia de ganado y costos de operación; se utiliza un muestreo de áreas. El estudio, al que se hará referencia, fue realizado por **Barbara Armstrong** en 1978.

El problema consistía en que los tamaños muestrales asignados a las provincias pequeñas era insuficiente para producir buenas estimaciones a dicho nivel. Para ésto se decidió probar la técnica de marcos múltiples en una de las provincias, eligiéndose la de New Brunswick (año 1978), a fin de mejorar la eficiencia de los estimadores y estudiar problemas operacionales relacionados con la metodología.

La AES emplea cierto tipo de marcos múltiples: el marco de áreas para asegurar la cobertura completa y la complementación con un listado para mejorar la eficiencia. Este listado incluye un grupo de establecimientos agropecuarios de gran tamaño, tomados desde el Censo Agropecuario de 1976 y son incluídos en la muestra con probabilidad 1. La justificación de su inclusión forzosa radica en que si se las considera sólo en el marco de

áreas pueden resultar poco representadas para el cálculo de los estimadores. Además estas granjas contribuyen significativamente en los totales provinciales, la no consideración puede proporcionar subestimaciones.

En el estudio para la provincia de New Brunswick se conservó el diseño muestral de la AES y además se empleó un listado a partir del censo de 1976, excluyéndose las granjas de menor tamaño. Luego se estratificó la lista y se consideró un muestreo simple al azar en cada estrato. La asignación a cada estrato se hizo en base a varias características. La asignación de la muestra a los estratos se hizo buscando la mejor combinación de los coeficientes de variación para tres variables conocidas.

En cuanto al marco de áreas, se definieron áreas de enumeración, con las cuales se formaron estratos y luego se aplicó un muestreo bietápico dentro de cada estrato; seleccionando áreas de enumeración primero y segmentos de tierra, en segunda etapa. Además, se consideró un diseño con replicaciones en cada estrato. El estimador utilizado fue el que correspondía al diseño. Al aplicar marcos múltiples se emplearon los estimadores de Hartley y el "screening". Este último, considerando el total de una característica dada, resulta de sumar la estimación del total a partir del listado y la estimación a partir de la muestra de área excluyendo los elementos del dominio ab. Claramente se ve que dicho estimador desprecia la información que proporcionan los elementos de la muestra de área que pertenecen a la intersección de los marcos.

La experiencia en New Brunswick demostró obtener ganancias al utilizar marcos múltiples, a través de la comparación de los coeficientes de variación del estimador "screening" con el de área (el utilizado por la AES). Esto puede deberse a que los listados proporcionan estimadores más eficientes y por ello es de esperar que, los coeficientes de variación disminuyan con su inclusión.

Luego se incluye la comparación con el de Hartley y se visualiza que la ganancia en la reducción de la variancia no es significativa respecto a la obtenida con el "screening". No obstante, si se dispone de la información, recomienda el uso del estimador de Hartley, especialmente si existen unidades del dominio ab ineficientemente captados por la lista.

Finalmente, destaca la importancia de una buena determinación de la intersección entre los marcos pues si no se identifican unidades de la misma se produciría una estimación "inflada". Esto se observó a través de diferencias encontradas entre las primeras estimaciones de marcos múltiples y unas segundas estimaciones hechas después de una revisión de los marcos donde se corrigió el tamaño de la intersección.

Si bien los estudios y aplicaciones de la técnica de marcos múltiples no se agotan en lo presentado aquí, trató de mostrarse que su utilización puede arrojar estimaciones confiables, con una reducción de costos respecto a la utilización de un solo marco.

De hecho, diversas investigaciones continúan desarrollándose, en busca de mejoras en los estimadores y de soluciones a problemas metodológicos, como podría ser la determinación de los dominios.

2. APLICACION PRACTICA

2.1 Objetivos y población en estudio

Con el fin de probar la aplicabilidad de la técnica estudiada en un área pequeña, se decidió trabajar en la comuna de Conchalí (ver Anexo 2), la cual participa del proceso de descentralización actual y requiere información que les permita tomar decisiones sobre una base concreta de la realidad de sus habitantes.

El objetivo es hacer estimaciones demográficas confiables y lo más actualizadas posibles, evaluando a su vez la técnica de marcos múltiples para tal fin. Se harán estimaciones del total poblacional, del total por sexo y por grandes grupos de edades. Se analizarán los resultados obtenidos y se realizará una comparación con las proyecciones existentes elaboradas por el CELADE.

2.2 Fuentes de datos

Puesto que se considera a la comuna de Conchalí como un "área pequeña", por definición existe información a través de una encuesta por muestreo de una población que la contiene. En este caso particular, se trata de la Encuesta Nacional de Empleo (ENE).

Además, se cuenta con información parcial de la propia comuna -lo cual nos lleva a utilizar la metodología de marcos múltiples- registrada a través de las fichas CAS.

Dado que las estimaciones se harán en base a estas dos fuentes de información, se describirá en que consiste cada una de ellas.

Encuesta Nacional de Empleo (ENE)

Esta encuesta es llevada a cabo por el Instituto Nacional de Estadística (INE), en forma regular y a nivel nacional, desde 1976. Se producen estimaciones para el total del país, regiones y provincias; también por áreas urbana y rural.

Entre 1985 y 1986 se produjo una actualización de la muestra. Si bien esta nueva muestra (Programa Integrado De Encuestas de Hogares) está utilizándose para la obtención de estadísticas laborales, su fin es mucho más amplio ya que constituye una infraestructura muestral de própositos múltiples.

Dicho formulario, además de averiguar sobre las características laborales para los mayores de 15 años, investiga sobre las variables jefe/no jefe del hogar,relación de parentesco, sexo, edad, estado civil y nivel educacional, como así también datos sobre la vivienda y el hogar.

El diseño muestral contempla una estratificación de carácter geográfico y por tamaño poblacional. A nivel de cada región y/o provincia se definen cuatro tipos de estratos, para los cuales se obtienen estimaciones que luego se componen para generar valores de las variables a los niveles correspondientes de estimación. En el caso de la provincia de Santiago, los estratos IF (de inclusión forzosa) corresponden a las Comunas.

Cada estrato es a su vez subdividido en secciones, cuyo tamaño se define en términos de población y viviendas, tratando de obtenerse secciones con un determinado tamaño medio en función del número de viviendas.

El diseño muestral del PIDEH es bietápico. Las unidades de primera etapa son las secciones y las de segunda etapa, viviendas particulares en dichas secciones. Las viviendas particulares finalmente seleccionadas son consideradas como conglomerados. Las secciones seleccionadas son unidades de muestreo de primera etapa a través de toda la vida útil de la muestra, en cambio las viviendas dentro de dichas secciones son sometidas a procesos de rotación a través del tiempo.

En cuanto a la determinación del tamaño de la muestra, trató de ser compatible con los niveles de estimación (provincia, región y país) y la importancia de los estratos definidos, y con las principales variables que son objeto de estimación. El diseño privilegia los niveles nacional y regional, buscando en éstos bajos errores de muestreo. También se tienen en cuenta las principales variables en estudio (tasa de desocupación, por ejemplo) y se considera la mayor variabilidad que provoca el trabajar con un muestreo bietápico.

La selección de la muestra se realizó en dos etapas. Primeramente fueron seleccionadas secciones en forma sistemática, a partir de un arranque aleatorio, y con probabilidad proporcional al tamaño (medido éste por medio del número de viviendas). Una vez seleccionadas las secciones se actualizó su tamaño para dejar definido el marco muestral para la segunda etapa, donde se seleccionó un tamaño de muestra, constante e igual, en promedio, a 15 viviendas dentro de cada sección.

La muestra total corresponde a un período trimestral pero distribuída en tres muestras de tamaño similar para cada mes. Esto indica que dicha encuesta se produce en forma continua donde cada mes se encuesta una muestra pequeña de viviendas.

En cuanto a los estimadores, éstos se calculan teniendo en cuenta la estratificación, el muestreo bietápico, probabilidades de selección, etc. Además, interviene información exógena -como proyecciones de población- que es tenida en cuenta utilizando estimadores de razón.

Ficha CAS

Este programa está a cargo de la Secretaría de Desarrollo y de Asistencia Social (dependiente del Ministerio de Planificación), a través del cual se asignan distintos tipos de subsidios a la población, como por ejemplo subsidios a la tercera edad (PASIS), Subsidio Unico Familiar, subsidios para vivienda, etc. Cada Municipio se encarga de recibir solicitudes de personas carenciadas que concurren a solicitar los subsidios. Luego, en una

segunda etapa se visitan las viviendas de dichas personas, intentando obtener información lo más veraz posible para determinar puntajes que establecen un orden de prioridad entre todos los postulantes. Esta tarea se desarrolla en forma permanente desde 1987, contando además con una actualización año a año de aquéllos que ya pertenecen al programa.

Su objetivo radica en una forma de focalización de programas sociales hacia la población con menores recursos, es por ello la importancia que debería darse al buen desempeño de la tarea de encuesta. Puesto que es bien sabido que no siempre son los más pobres los que recurren a solicitar ayuda y los medios no son suficientes para cubrir a todos, debe tratarse de seleccionar entre todos los necesitados registrados a aquéllos más carenciados.

El registro CAS consiste en una ficha por vivienda, distinguiendo las distintas familias en su interior y recabando información sobre la vivienda y sobre todos los moradores de la misma.

En base a la información sobre características de la vivienda, educación, ocupación, ingreso y patrimonio, se construye el puntaje que resume la condición socio-económica del individuo y su familia.

2.3 Elección de las muestras y variables de estudio

Para el presente estudio se utilizó la información de la ENE recogida en el último trimestre de 1990 para la comuna de Conchalí. De la ficha CAS se consideró la base de datos constituída por las actualizaciones e inscripciones nuevas registradas durante el mismo período, de tal forma de hacerla comparable con la otra fuente.

Como la CAS posee un registro permanente, en un período dado se tiene información antigua y además se recoge información de las fichas que corresponde actualizar (trancurrió un año desde la encuesta anterior) y de las nuevas inscripciones. Si bien uno podría diseñar una muestra a partir del listado completo de las viviendas de la CAS, se pensó aprovechar el relevamiento antes mencionado, estudiando la posibilidad de considerarlo como una muestra aleatoria, y así disminuir costos.

Para ello se tuvo en cuenta la naturaleza de las variables en estudio, observando si éstas dependían del tiempo y si existían diferencias entre los datos de las actualizaciones y de las nuevas inscripciones, correspondientes a dichas variable.

Puesto que para estimar totales de población y distribuciones por edad y por sexo, no hay indicios para suponer que el tiempo afecta dichas variables o que existen diferencias entre las actualizaciones y las nuevas inscripciones, es posible trabajar con la información recogida en el último trimestre de 1990 considerándola como una muestra aleatoria del total de registros. Más aún, puede pensarse en un muestreo sistemático donde el determinante de la muestra no es la primer unidad seleccionada sino, una unidad de tiempo.

El marco del INE consistía de 34691 viviendas, las cuales fueron distribuídas en 113 secciones. La muestra de la ENE consta de 8 secciones seleccionadas en forma sistemática, en una primer etapa. En una segunda etapa, de cada una de las 8 secciones de tamaño M_i (i=1...8) se extraen 15 viviendas, también en forma sistemática.

En el caso de la ficha CAS, el tamaño del listado se consideró como el número de inscriptos registrados en un año, lo cual proporcionaba un marco de 8500 viviendas. Luego, la muestra registraba 1208 viviendas para el período seleccionado.

2.4 Presentación de estimadores

Teniendo en cuenta la simbología utilizada en el primer capítulo, llamaremos

Marco A, al correspondiente a la Encuesta de Empleo (ENE).
Marco B, al correspondiente a la Ficha CAS.

M_A = Total de viviendas en el marco A (ENE).
M_B = Total de viviendas en el marco B (CAS).
M_a = Total de viviendas sólo en A.
M_{ab} = Total de viviendas en A y B.
m''_{ab} = número de viviendas tomadas en la muestra del marco B.

Dado que el marco del listado está contenido en el del INE, entonces:

$$M_B = M_{ab}$$

y

$$M_b = 0.$$

Considerando el diseño muestral de la ENE, se tiene

N = número de secciones en la población, del marco A.
n = número de secciones en la muestra, del marco A.

Luego, las M_i pueden ser distribuídas en M_{ai} y M_{abi}, según pertenezcan a los dominios a ó ab, respectivamente. Lo mismo sucede con las m_i de cada sección de la muestra, distribuyéndose en m_{ai} y m'_{abi}.

El reconocimiento de unidades en ambas muestras se hizo a través de las direcciones de viviendas.

Considerando el estimador de Hartley, se tiene:

$$\hat{Y}_H = \hat{Y}_a + p\,\hat{Y}'_{ab} + q\,\hat{Y}''_{ab}$$

$$= N/n \sum_{i=1}^{n} M_{ai}/m_{ai} \sum_{j=1}^{m_{ai}} y_{aij} +$$

$$p\,N/n \sum_{i=1}^{n} M_{abi}/m'_{abi} \sum_{j=1}^{m'_{abi}} y_{abij} +$$

$$q\,M_{ab}/m''_{ab} \sum_{k=1}^{m''_{ab}} y_k$$

donde:

$(p + q) = 1$

$\hat{Y}_a$ es el total estimado para el dominio "a"

$\hat{Y}'_{ab}$, $\hat{Y}''_{ab}$ son los totales estimados para el dominio ab a partir de los marcos A y B, respectivamente.

y_{aij} es el total de personas con la característica analizada, en una vivienda j de la sección i-ésima de la muestra en el dominio a.

y_k es el total de personas con la característica analizada, en una vivienda k de la muestra del marco B.

Para el cálculo de la variancia debe tenerse en cuenta la forma del estimador de marcos múltiples, el muestreo bietápico de la ENE y en el caso de la CAS, puede considerarse aquélla de un muestreo simple al azar. Esta última aseveración se basa en que $E[Var(y_{sia})] = Var(y_{mas})$, teorema demostrado por Madow en 1944 (el desarrollo en detalle se presenta en Anexo 3).

Finalmente:

$$Var(\hat{Y}_H) = N^2 (1-f_A)/[n(n-1)] \sum_{i=1}^{n} (M_{ai}\bar{y}_{ai} - \bar{y}_a)^2 +$$

$$+ N/n \sum_{i=1}^{n} M_{Ai}^2/\, m_{Ai} (1-\alpha_i) (1-f_{Ai})\, s_{ai}^2 +$$

$$+ p^2 \{ N^2(1-f)/[n(n-1)] \sum_{i=1}^{n} (M_{abi}\, \bar{y}_{abi} - \bar{y}'_{ab})^2 +$$

$$+ N/n \sum_{i=1}^{n} (1-f_{Ai})\, M_{Ai}^2/\, m_{Ai}\, \alpha_i\, s^{2'}_{abi}\} +$$

$$+ q^2\ (1-f_B)\, M_{ab}{}^2/m''_{ab}\, s^{2''}_{ab}$$

donde:

$f_A = n/N$ $\qquad f_{Ai} = m'_{ab}/M_{ab}$ $\qquad f_B = m''_{ab}/M_{ab}$

$\bar{y}_a = 1/n \sum_{i=1}^{n} M_{ai}\, \bar{y}_{ai}$ $\qquad \bar{y}'_{ab} = 1/n \sum_{i=1}^{n} M_{abi}\, \bar{y}_{abi}$

$$\bar{y}_{ai} = 1/m_{ai} \sum_{j=1}^{m_{ai}} y_{aij} \qquad \bar{y}'_{abi} = 1/m'_{abi} \sum_{j=1}^{m_{abi}} y'_{abij}$$

$$y_B = 1/m''_{ab} \sum_{k=1}^{m''_{ab}} y_k$$

$$\alpha_i = M_{abi}/M_{Ai}$$

$$\alpha = M_{ab}/M_A$$

$$ß = M_{ab}/M_B = 1$$

$$p = \alpha m_A/ (\alpha m_A + ß m_B) \qquad q = 1-p$$

$$s^2_{ai} = 1/(m_{ai}-1) \sum_{j=1}^{m_{ai}} (y_{aij} - \bar{y}_{ai})^2$$

$$s^{2'}_{abi} = 1/(m'_{abi}-1) \sum_{j=1}^{m_{abi}} (y_{abij} - \bar{y}_{abi})^2$$

$$s^{2''}_{ab} = 1/(m''_{ab}-1) \sum_{k=1}^{m''_{ab}} (y_{Bk} - \bar{y}_B)^2$$

Para comparar los estimadores de la nueva técnica con los obtenidos solamente a través de los datos de la Encuesta de Empleo se consideró el estimador del total con igual probabilidad, ya que los tamaños de las secciones no varían significativamente.

$$\hat{Y}_{área} = N/n \sum_{i=1}^{n} M_{Ai}/m_{Ai} \sum_{j=1}^{m_{Ai}} y_{Aij}$$

$$Var(\hat{Y}_{área}) = N^2 (1-f_A)/[n(n-1)] \sum_{i=1}^{n} (M_{Ai}\bar{y}_{Ai} - \bar{y}_A)^2 +$$

$$+ N/n (1-f) \Sigma M_{Ai}^2/m_{Ai} (1-f_{Ai}) s^{2''}_{Ai}$$

siendo:

$$\bar{y}_A = 1/n \sum_{i=1}^{n} M_{Ai} \bar{y}_{Ai} \qquad \bar{y}_{Ai} = 1/m_{Ai} \sum_{j=1}^{m_{Ai}} M_{Ai} y_{Aij}$$

$$s^{2''}_{Ai} = 1/(m_{Ai}-1) \sum_{j=1}^{m_{Ai}} (y_{Aij} - \bar{y}_{Ai})^2$$

2.5 Procedimiento de cálculos y análisis de resultados.

De acuerdo al diseño muestral de la ENE, para la estimación del total de una característica Y se requiere conocer los M_{ai} , M_{abi} , m_{ai} y m'_{abi}, en cada una de las 8 secciones de la muestra. Para ello se necesitaría un listado con las direcciones de las 8500 viviendas de la CAS y observar cuántas caen en las secciones muestreadas y luego cotejarlas con las viviendas de la muestra de la ENE.

Como no fue posible contar con dicho listado, se tomaron distribuciones en base a los datos disponibles. A partir de la muestra de la CAS -de la cual sí se poseían los domicilios

se cotejaron las 1208 viviendas con los listados de vivienda de la ENE y se determinaros así parte de las M_{ai} y M_{abi}, como así también las m_d (viviendas comunes en ambas muestras).

De esta manera, utilizando la distribución de las viviendas de la muestra de la CAS que cayeron en las 8 secciones, se estimaron los primeros M_{ai} y M_{abi} (distribución I).

Por otra parte, las secciones de la ENE están contenidas en Unidades Vecinales (divisiones geográficas consideradas por el Municipio). Entonces, dado que se conocen las unidades vecinales que envuelven a las 8 secciones de la muestra y que las fichas CAS identifican a las mismas, se calcularon otros M_{ai} y M_{abi} a partir de la distribución de las viviendas de la muestra de la CAS en todas las Unidades Vecinales (distribución II).

Así, en cada sección se desagregó el marco entre los dominios "a" y "ab" en las dos formas que se presentan:

		Distribución I		Distribución II	
Sección	M_i	M_{ai}	M_{abi}	M_{ai}	M_{abi}
101	305	251	54	228	77
102	337	284	53	261	76
103	312	254	58	227	85
104	306	296	10	292	14
105	343	328	15	322	21
106	294	168	126	164	130
107	318	279	39	262	56
108	241	236	5	234	7

Respecto a los m_{ai} y m'_{abi}, sólo se conocen las viviendas que en ambas muestras pertenecen al dominio ab (elementos comunes). Entonces, las viviendas que se encuestaron por el INE se distribuyeron en los dominios a y ab de cuatro formas distintas. La primer forma consistió en suponer que las viviendas comunes en las muestras constituían las únicas que pertenecían al dominio ab. Esto es, de las 15 viviendas seleccionadas de cada una de las 8 secciones, sólo las comunes con la muestra de la ficha CAS caían en la intersección de los marcos.

A partir de esta distribución de la muestra del INE entre los dominios a y ab, se formaron otras incrementando los tamaños del dominio ab y cambiando la forma de distribución. Esto, para probar la sensibilidad del estimador. En estos casos siempre se incluyeron las unidades que ya pertenecían a la intercepción y las nuevas se seleccionaron aleatoriamente entre las restantes, en cada sección. Con este criterio se determinaron las siguientes:

Distribuciones de la muestra

SECCION		D1		D2		D3		D4	
	m_i	m_{ai}	m'_{abi}	m_{ai}	m'_{abi}	m_{ai}	m'_{abi}	m_{ai}	m'_{abi}
101	15	14	1	13	2	12	3	13	2
102	13	13	0	12	1	11	2	9	4
103	15	14	1	13	2	12	3	12	3
104	14	13	1	12	2	12	2	11	3
105	15	15	0	14	1	14	1	15	0
106	15	12	3	12	3	11	4	10	5
107	14	13	1	13	1	12	2	12	2
108	15	15	0	15	0	14	1	14	1

Para el cálculo de estimadores y variancias primeramente se calculó el valor de p,

$$p = \alpha m_A/(\alpha m_A + \beta m_B)$$

$$= (M_{ab} * m_A/M_A)/[(M_{ab} * m_A/M_A) + (M_{ab} * m_B/M_B)]$$

$$= (M_{ab} * m_A/M_A)/[(M_{ab} * m_A/M_A) + m_B)]$$

$$= (8500 * 120/34691)/[(8500 * 120/34691) + 1208]$$

$$= 0.024$$

De esta manera, el estimador obtenido a partir de las unidades de la muestra del INE que caen en la intersección serán ponderados por 0.024. Por lo tanto, los procedentes de la ficha CAS se ponderarán por 0.976.

A partir de las distribuciones propuestas se calcularon los valores del estimador para cada característica, arrojando los resultados que se presentan a continuación. Denotaremos con I y II a las distribuciones del marco del INE entre los dominios a y ab, y con 1, 2, 3, 4 a las distribuciones de la muestra.

Además, se presentan para las mismas características las proyecciones realizadas por CELADE (Y_{proy}) y se considera un estimador de área basado en los datos proporcionados por la ENE para Conchalí ($Y_{área}$) 2/.

2/ La ENE utiliza un estimador de área corregido por las actualizaciones de las secciones muestreadas; no proporciona estimaciones para áreas pequeñas.

Estimaciones para el total de población

Distribuciones	$\hat{Y}_a$	$\hat{Y}_{ab}$	$\hat{Y}_H$	$\hat{\sigma}(\hat{Y}_H)$	c.v(%)
I.1	134745	31739	177220	10592.62	5.98
I.2	135679	34564	178222	11021.89	6.18
I.3	136760	30439	179204	10995.44	6.14
I.4	135687	29387	178106	10216.17	5.74
II.1	127660	39569	170323	9974.64	5.86
II.2	128544	43630	171305	10386.85	6.06
II.3	129575	38302	172208	10359.32	6.02
II.4	128552	37316	171161	9598.50	5.61

$\hat{Y}''_{ab}$ = 42739 $\quad \hat{V}(\hat{Y}''_{ab})$ = 294893.8

Y_{proy} = 170667

$\hat{Y}_{área}$ = 164407 $\quad \hat{\sigma}(\hat{Y}_{área})$ = 10251.88 $\quad$ C.V. = 6.24%

Estimaciones para totales por sexo

MUJERES

Distribuciones	$\hat{Y}_a$	$\hat{Y}_{ab}$	$\hat{Y}_H$	$\hat{\sigma}(\hat{Y}_H)$	c.v(%)
I.1	69572	18348	91673	5418.43	5.91
I.2	69859	19069	91977	5387.85	5.86
I.3	70519	16232	92569	5324.56	5.75
I.4	70533	15035	92554	4811.64	5.20
II.1	66057	22755	88264	5354.59	6.41
II.2	66289	23758	88520	5259.23	5.94
II.3	66910	20225	89056	5189.34	5.83
II.4	66939	18991	89056	4722.87	5.30

$\hat{Y}''_{ab}$ = 22193 $\quad \hat{V}(\hat{Y}''_{ab})$ = 108618.8

Y_{proy} = 88193

$\hat{Y}_{área}$ = 85684 $\quad \hat{\sigma}(\hat{Y}_{área})$ = 6036.28 $\quad$ C.V.= 7.045%

HOMBRES

Distribuciones	$\hat{Y}_a$	$\hat{Y}_{ab}$	$\hat{Y}_H$	$\hat{\sigma}(\hat{Y}_H)$	c.v(%)
I.1	65172	13391	85547	6235.57	7.30
I.2	65820	15495	86245	6479.14	7.51
I.3	66241	14349	86639	6600.29	7.62
I.4	65154	14351	85552	6458.36	7.55
II.1	61603	16813	82059	5651.38	6.89
II.2	62255	19872	82785	5956.02	7.19
II.3	62664	18275	83156	6084.93	7.31
II.4	61613	18325	82106	5938.98	7.23

$\hat{Y}''_{ab}$ = 20546 $\quad \hat{V}(\hat{Y}''_{ab})$ = 121396.8

Y_{proy} = 82474

$\hat{Y}_{área}$ = 78723 $\quad \hat{\sigma}(\hat{Y}_{área})$ = 6088.47 $\quad \hat{C.V.}$ = 7.734%

Estimaciones por grandes grupos de edad

EDAD 0-14 años

Distribuciones	$\hat{Y}_a$	$\hat{Y}_{ab}$	$\hat{Y}_H$	$\hat{\sigma}(\hat{Y}_H)$	c.v(%)
I.1	33327	13800	49351	3024.95	6.13
I.2	33256	13419	49270	2741.29	5.56
I.3	34261	10412	50204	3027.68	6.03
I.4	33544	10234	49482	2725.64	5.51
II.1	31633	16964	47733	2936.43	6.15
II.2	31560	16427	47647	2672.50	5.61
II.3	32473	12449	48465	2866.20	5.91
II.4	31807	12555	47801	2595.73	5.43

$\hat{Y}''_{ab}$ = 16078 $\quad \hat{V}(\hat{Y}''_{ab})$ = 94320.98

Y_{proy} = 49416

$\hat{Y}_{área}$ = 43265 $\quad \hat{\sigma}(\hat{Y}_{área})$ = 4318.32 $\quad \hat{C.V.}$ = 9.98%

EDAD 15-64 años

Distribuciones	$\hat{Y}_a$	$\hat{Y}_{ab}$	$\hat{Y}_H$	$\hat{\sigma}(\hat{Y}_H)$	c.v(%)
I.1	94668	17204	119783	8032.08	6.71
I.2	95598	20411	120790	8238.23	6.82
I.3	96323	18417	121467	8278.73	6.81
I.4	96066	17866	121197	8148.33	6.72
II.1	89645	21785	114870	7425.31	6.46
II.2	90546	26393	115882	7688.55	6.63
II.3	91265	23728	116537	7754.08	6.65
II.4	90992	23061	116248	7564.37	6.51

$\hat{Y}''_{ab}$ = 25310 $\quad \hat{V}(\hat{Y}''_{ab})$ = 145911.20

Y_{proy} = 111077

$\hat{Y}_{área}$ = 113131 $\quad \hat{\sigma}(\hat{Y}_{área})$ = 8238.38 $\quad \hat{C.V.}$ = 7.28%

EDAD 65 y más años

Distribuciones	$\hat{Y}_a$	$\hat{Y}_{ab}$	$\hat{Y}_H$	$\hat{\sigma}(\hat{Y}_H)$	c.v(%)
I.1	6750	735	8086	1545.77	19.11
I.2	6826	735	8162	1537.20	18.83
I.3	6176	1610	7533	1385.38	18.39
I.4	6076	1287	7426	1420.19	19.13
II.1	6382	810	7720	1455.15	18.85
II.2	6438	810	7776	1422.36	18.29
II.3	5833	2126	7206	1288.78	17.88
II.4	5753	1700	7112	1335.12	18.77

$\hat{Y}''_{ab}$ = 1351 $\quad \hat{V}(\hat{Y}''_{ab})$ = 9415.267

Y_{proy} = 10174

$\hat{Y}_{área}$ = 8012 $\quad \hat{\sigma}(\hat{Y}_{área})$ = 1759.27 $\quad \hat{C.V.}$ = 22 %

Del análisis de los resultados presentados para cada una de las variables que se estimaron, podría decirse que el estimador de marcos múltiples utilizado no resulta sensible a la distribución de los elementos de la muestra (viviendas) entre los dominios a y ab.

En cambio, diferencias algo marcadas se estarían detectando al cambiar la distribución de las unidades del marco constituído por las fichas CAS dentro de las secciones muestreadas por el INE. Con la distribución I se obtienen estimaciones mayores que con la distribución II.

Esto último podría deberse a que, en la distribución II el tamaño del dominio a se reduce en cada una de las 8 secciones de la muestra del INE y se incrementa el del dominio ab. Así, las estimaciones del total hechas a partir de sólo INE (Y_a) disminuyen y las de la intersección (Y'_{ab}) se incrementan. Por otra parte, la estimación del total a partir de las fichas CAS (Y''_{ab}) se mantiene constante, en el estimador de marcos múltiples, para ambas distribuciones. Luego, la baja ponderación aplicada a Y'_{ab} hace que su incremento no compense la disminución de Y_a.

No obstante, las diferencias observadas oscilan entre un 3 y un 4 por ciento. Posiblemente la distribución II está más próxima a la verdadera distribución entre los dominios a y ab, puesto que se empleó mayor información para su confección. Sería importante, entonces, determinar la real distribución del listado de la CAS en las secciones y así verificar las estimaciones.

En cuanto a los coeficientes de variación, éstos resultan aceptables para todas las estimaciones, a excepción del grupo de edades "65 y mas" donde superan al 10 por ciento en todos los casos. Un incremento importante en la variancia se debe a las estimaciones que provienen de la información de la ENE ya que ésta, además de poseer una muestra pequeña para Conchalí, contiene muy pocas observaciones como para lograr estimaciones confiables en este grupo de edad y menos aún cuando se la divide en los dominios a y ab.

Un aspecto interesante a mencionar es el del "mejoramiento" en las estimaciones, si se comparan los coeficientes de variación estimados a partir del estimador de Hartley con el obtenido de la ENE ($Y_{área}$). En todos los casos el uso de dos marcos provoca disminuciones en los coeficientes, presentándose en algunas estimaciones bajas realmente importantes. Además, parecería que cuando existe un mayor número de elementos interceptados en la muestra, las estimaciones resultan mas confiables.

Considerando ahora las proyecciones del CELADE para el total poblacional y los totales por sexo, las estimaciones de marcos múltiples resultan cercanas a ellas, siendo aproximadamente de un 4 por ciento la mayor diferencia encontrada. No obstante, varios de los resultados difieren en menos de un 0.5 por ciento.

Respecto a las estimaciones por grupos de edades, las diferencias resultan superiores, excepto para el grupo "0-14". Aquí aparece nuevamente el problema del tamaño pequeño de la muestra del INE, y más aún al dividir las viviendas en los dominios a y ab. Si además se toman subpoblaciones pequeñas en relación a la población total, no sólo puede distorsionarse la distribución de éstas en las muestras sino arrojar estimaciones muy poco precisas.

El grupo de edades "65 y más" sería una subpoblación que presenta este inconveniente, al igual que el de "0-14", por lo cual las estimaciones para el grupo "15-64" se ven "infladas".

En cuanto al grupo "0-14", no se reflejan mayores discrepancias con la respectiva proyección porque la subrepresentatividad logra compensarse con la información de la CAS. La misma arroja una distribución por los tres grandes grupos de edades donde el primer grupo tiene más peso que en la distribución de los datos del INE.

Ahora, una forma de validar la proyección se hará a través de docimasias de hipótesis, tomando a cada una de las muestras de marcos múltiples y sus respectivas estimaciones para el total poblacional y los totales por sexo.

Para ello se considera un test, bajo supuesto de normalidad, de la forma

$$H_0)\ Y = Y_{proy}$$

$$H_1)\ Y \neq Y_{proy}$$

fijándose $\alpha = 0.05$.

Si bien no se puede probar el supuesto de normalidad del estimador de marcos múltiples (sólo se cuenta con una muestra), se tratará de evaluar el no cumplimiento mediante simulación. En el Anexo 4 se presentan estimaciones simuladas teniendo en cuenta las distribuciones del número de personas por vivienda de la muestra de la CAS y del INE a partir de las cuales se fueron generando posibles muestras en forma aleatoria. Esto se hizo 1000 veces para el total poblacional y totales por sexo, tomando las distribuciones I.2 e II.4.

En base a las estimaciones obtenidas se hizo un test de Bondad de Ajuste para cada variable y distribución considerada, concluyendo que no se rechaza la suposición de normalidad, a un nivel de significación de 0.05.

Los resultados de las dócimas para las proyecciones (Anexo 5), permiten decir que ante la evidencia muestral puede considerarse a éstas como los verdaderos valores de los parámetros poblacionales.

Si bien la validación es recíproca, es decir uno puede decir que la muestra constituída por marcos múltiples estima adecuadamente a las proyecciones, el pensar en éstas como valores paramétricos se evidencia también de la comparación con las estimaciones hechas con los datos de la ENE.

Las diferencias de las $Y_{área}$ respecto a las proyecciones son considerables, cosa que no ocurre cuando se realiza la comparación a un nivel de agregación mayor3/, como se muestra en el siguiente cuadro.

3/ Se consideraron las estimaciones que calcula el INE para Total País, Región Metropolitana y Provincia de Santiago. Suponemos que a nivel comunal el $Y_{área}$ estaría próximo a la estimación que obtendría el INE.

	Ambos Sexos	Hombres	Mujeres
Total País			
Proyección	13173348	6505617	6667731
Est. INE	12897100	6313900	6588100
Región Metropolitana			
Proyección	5236321	2514106	2722215
Est. INE	5122500	2438700	2684500
Provincia de Santiago			
Proyección	4388313	2090454	2297859
Est. INE	4380800	---	---

Fuente: Proyección: -INE, CELADE "Chile, Proyección de Población por Sexo y Edad". Provincias 1980-2000. Fascículo F/CHI.4. INE, Santiago, Chile, 1988.
-INE, CELADE "Chile, Proyección de Población por Sexo y Edad". Regiones 1980-2000. Fascículo F/CHI.3. INE, Santiago, Chile, 1987.
-INE, CELADE "Chile, Proyección de Población por Sexo y Edad". Totla del País 1950-2025. Fascículo F/CHI.1. INE, Santiago, Chile, 1987.

Estimación del INE: Inédito.

Puesto que la ENE está diseñada para arrojar estimaciones confiables a nivel país, regiones y provincias, es claro que las cifras observadas en el cuadro cumplirían dicho propósito. Las discrepancias respecto de las proyecciones son en promedio de un 1.8 por ciento. Esto demuestra, por un lado la validez de las proyecciones; por otro, que mientras la ENE no permite obtener buenas estimaciones a nivel comunal, la utilización de marcos múltiples sí lo hace. Esto último se deduce de la comparación de las estimaciones realizadas a partir de la Distribución II (como se explicó anteriormente, ésta resulta más precisa) con las Y_{area} y las proyecciones.

2.6 Consideraciones finales.

De las variables que se analizaron en este trabajo, no fue posible realizar una mayor desagregación -grupos quinquenales de edad para el total y por sexo- debido a las pocas observaciones que caerían dentro de cada sección muestral, dado el pequeño tamaño de muestra utilizado en la ENE, lo cual arrojaría resultados poco confiables y precisos.

Hubiera resultado interesante calcular tasas u otras medidas referidas a la ocupación, a fin de reflejar el perfil socio-económico de la Comuna, dado que ambas fuentes (INE y CAS) cuentan con información al respecto. En este caso debería analizarse si los datos de las fichas CAS para un período dado pueden ser considerados como una muestra aleatoria del total de fichas, observando si existe relación entre el tiempo y esta variable.

Además, deberían compararse las fichas de las viviendas actualizadas y aquéllas que recién ingresan al programa para determinar si existen diferencias significativas dadas por el momento de entrada 4/. Para ello se sugeriría el mantener información que permita identificar las actualizaciones de los recién ingresados. Esto último no sólo para permitir realizar estimaciones sino que serviría para propósitos más amplios. Por ejemplo, se podría evaluar el programa desde el punto de vista de su cobertura y de los posibles cambios en la demanda de cada tipo de subsidios.

Ahora bien, supongamos se pruebe la validez del listado CAS de un período como muestra, de tal forma de poder aplicar marcos múltiples para estimar la desocupación. Obviamente, esta variable es una de las más relevantes para los gobiernos comunales, y lamentablemente sólo existen estimaciones a niveles de agregación mayores que la comuna. Puesto que creemos que la metodología presentada aquí permitiría obtener estimaciones confiables (vimos que el tamaño de las subpoblaciones en la muestra del INE referidas al empleo "alcanzaría" para arrojar estimaciones precisas usando ambos marcos), sería conveniente para fines estadísticos que las fichas CAS desagregaran -en la pregunta sobre ocupación- la categoría "No tiene ocupación" de tal forma que puedan identificarse cesantes, los que buscan trabajo por primera vez y los inactivos (amas de casa, estudiantes, etc.). Así, la información se haría compatible a la recogida por la ENE.

Por último, sería interesante seguir trabajando en la aplicación de esta metodología, probando resultados con otras variables de interés y en otras comunas. Para ello, se sugiere tener presentes los comentarios previos y, desde el punto de vista de la metodología, se tendría que ver la forma de que las fichas CAS identifiquen la sección de la ENE a la cual pertenecen. Estimamos que ésto es factible de realizar y que de esa manera las comunas podrían contar con buenas estimaciones a un costo inferior que resulta de utilizar información disponible.

4/ En este caso, el análisis podría complicarse ya que la base de datos no contiene información acerca de la fecha de entrada al programa; debería indagarse en registros administrativos o de otra índole para detectar dichos grupos (nuevos y actualizados).

CONCLUSION

Norman Beller (1979) explica en uno de sus trabajo relacionado con marcos múltiples lo que llama una paradoja estadística desde el punto de vista del muestreo; la idea se resume en que: "esfuerzos continuos para disminuir el error de muestreo (mejorar la precisión) a menudo requieren diseños muestrales más complicados, los cuales pueden incrementar los errores ajenos al muestreo (disminuir la exactitud) y entonces tener un error total mayor".

Si bien la metodología de marcos múltiples podría caer en esta paradoja, deben tenerse presente las experiencias positivas que se obtuvieron con la técnica (confiabilidad y bajo costo a la vez), a través de los diversos estudios que la utilizaron.

No obstante hay que considerar que una buena identificación de los elementos que pertenecen a la intersección de los marcos es imprescindible para que las estimaciones sean confiables y precisas. Si bien ésto requiere un cierto costo, generalmente resulta insignificante respecto a tener que elaborar un marco completo para la población a estudiar, y más aún si se trata de una población muy cambiante en el tiempo.

Ahora bien, en el caso particular de la presente aplicación, la identificación de los elementos del dominio ab puede hacerse de manera confiable y a un costo reducido. Por ejemplo, los responsables de la CAS en cada una de las comunas podrían incluir en las fichas una variable cuyo código indique las secciones de la muestra del INE a la cual pertenece la vivienda, facilitando su identificación mediante la combinación de la cartografía de ambas fuentes. Por otra parte, el INE podría incluir en el cuestionario de la ENE una pregunta que indague sobre la inscripción de los encuestados al programa CAS.

Por lo tanto, se estima que la técnica puede ser utilizada en el cálculo de estimaciones en áreas pequeñas, trabajando con el marco de la ENE y con el listado de la CAS. Habría que implementar algunas modificaciones o agregar información en las fichas CAS (por ejemplo, compatibilización con los datos del INE en el caso de la ocupación, etc.) lo cual, con costos adicionales mínimos, produciría ganancias sustanciales.

Finalmente, para una evaluación más objetiva de esta metodología se recomienda aplicarla en un período cercano al censo de Población y Vivienda a realizarse en 1992, y así comparar resultados. Para ello, se reitera la importancia de que exista coordinación entre el INE y los responsables de la CAS y de lograr la identificación de las viviendas del listado CAS en las secciones de la muestra de la ENE.

BIBLIOGRAFIA

ARMSTRONG, B. (1979). Test of multiple frame sampling techniques for agricultural surveys: New Brunswick, 1978. A.S.A., pág. 295-300.

AZORIN, F. y SANCHEZ CRESPO, J.L. (1986). Métodos y aplicaciones del muestreo. Alianza Editorial. Madrid.

BANKIER, M. (1983). Estimators based on several samples with applications to multiple frame surveys. A.S.A, pág. 91-96.

BELLER, N. (1979). Error profile - Multiple frame designs. A.S.A., pág. 221-222.

BOSECKER, R. and FORD, B. (1976). Multiple frame estimation with stratified overlap domain. A.S.A., pág. 219-224.

FULLER, W. and BURMEISTER, L. (1972). Estimators for samples selected from two overlapping frames. Meetings of the American Statistical Association.

HANSEN, M.; HURWITZ, W.; MADOW, W. (1960). Sample Survey Methods and Theory, Volume I: Methods and applications. U.S.A.

HARTLEY, H. O. (1962). Multiple frame surveys, A.S.A. Proceeding of the Social Statistics Section, pág. 203-206.

HUANG, E.; HOGUE, C. and ISAKI, C (1980). Comparisons of multi-frame with single-frame sample and voting survey data. A.S.A., pág. 711-715.

INSTITUTO NACIONAL DE ESTADÍSTICAS. CHILE: Proyecciones y estimaciones de población, por sexo y edad. Región Metropolitana, Comunas (1980-1995).

INSTITUTO NACIONAL DE ESTADÍSTICAS. División de Estadísticas Demográficas y Sociales Contínuas. Aspectos metodológicos del Programa Integrado de Encuestas de Hogares (1987). Santiago, Chile.

KALTON, G. and ANDERSON,D. (1986). Sampling Rare Populations. J. R. Statist. Soc. A, 149. Part. 1, pág. 65-82.

LUND, R. (1968). Estimators in Multiple Frame Surveys. A.S.A Proceeding of the Social Statistics Section, pág. 282-286.

SUPERINTENDENCIA DO DESENVOLVIMENTO DO NORDESTE, RECIFE. Importaçöes e Exportaçöes do Nordeste do Brasil (1980). Metodología estatística, pág. 12-19.

UNITED NATIONS, New York: A Short Manual on Sampling (1960). Volume I, Elements of sample survey theory. Serie F, Nro. 9.

VOGEL, R. and BROGAN, D. (1984). Integrated multiple frame sample surveys. A.S.A., pág. 233-236.

BIBLIOGRAFIA COMPLEMENTARIA

DALLA ZUANNA, G. (1989). Stima di dati su fecondita ed abortivita in piccole aree. Statistica, pág. 89-107.

GONZALEZ, M. E. and HOZA, C. (1978). Small Area Estimation with Application to Unemployement and Housing Estimates. Journal of the American Statistical Association, 73, pág. 7-15.

HIDIROGLOU, M. and SARNDAL, C. (1987). Conditional Inference for Small Area Estimation. A.S.A., pág. 147-156.

LUI, KUNG-JONG (1988). A Model-based approach: Composite Estimators for Small Area Estimation. A.S.A., pág. 191-195.

MACGIBBON, B. and TOMBERLIN, T. (1987). Small Area Estimates of Proportions via Empirical Bayes Techniques. A.S.A., pág. 341-346.

PRASAD, N. and RAO, J. (1986). On the estimation of mean square error of small area predictors. A.S.A., pág. 108-116.

ANEXOS

ANEXO 1

PROBLEMA DE ASIGNACION OPTIMA DE LA MUESTRA ENTRE LOS MARCOS

Caso de N_a, N_b y N_{ab} conocidos

I. Programa en basic del sistema iterativo (realizado en calculadora)

```
10    " A "
20    CLEAR
30    DIM R(100), Q(100)
40    INPUT "PRECISION= "; G
50    INPUT "N_A= "; A, "N_B= "; B, "Nab= "; N, "Ca= "; C, "Cb= "; D, "Va= "; V, "Vb= "; W,
"Vab= "; X
60    O= D/C
70    T= N/B
80    F=N/A
90    S= T/F
100   PRINT "Alfa= "; F
110   PRINT "Beta= "; T
120   L= 1-F
130   M= 1-T
140   Q(1)= 0
150   FOR I= 1 TO 100
160   R(1)=  (O*S)
170   Q(I+1)= O*S^2*((R(I)+S)^2*L*V + R(I)^2*X) / ((R(I)+S)^2*M*W+S^2*T*X)
180   R(I+1)=  Q(I+1)
190   K= ABS(R(I+1)-R(I))
200   IF K<G THEN PRINT "RE= "; R(I+1); ", I"; I
210   IF K<G GOTO 230
220   NEXT I
230   END
```

II. Prueba de la convergencia del método

En este ejemplo serán tomadas distintas combinaciones entre tamaños poblacionales de marcos (N_A y N_B), tamaño de intersección (N_{ab}), costos por unidad de encuesta (c_A y c_B) y variancias en los dominios. De esta manera pueden estimarse α, ß y el r óptimo, el cual determina la razón adecuada entre los tamaños de muestra de los marcos, para los parámetros específicados.

Para observar mejor el comportamiento de dicho sistema operativo se trabajó con un nivel bastante exigente de precisión.

PRECISION= 0.0001

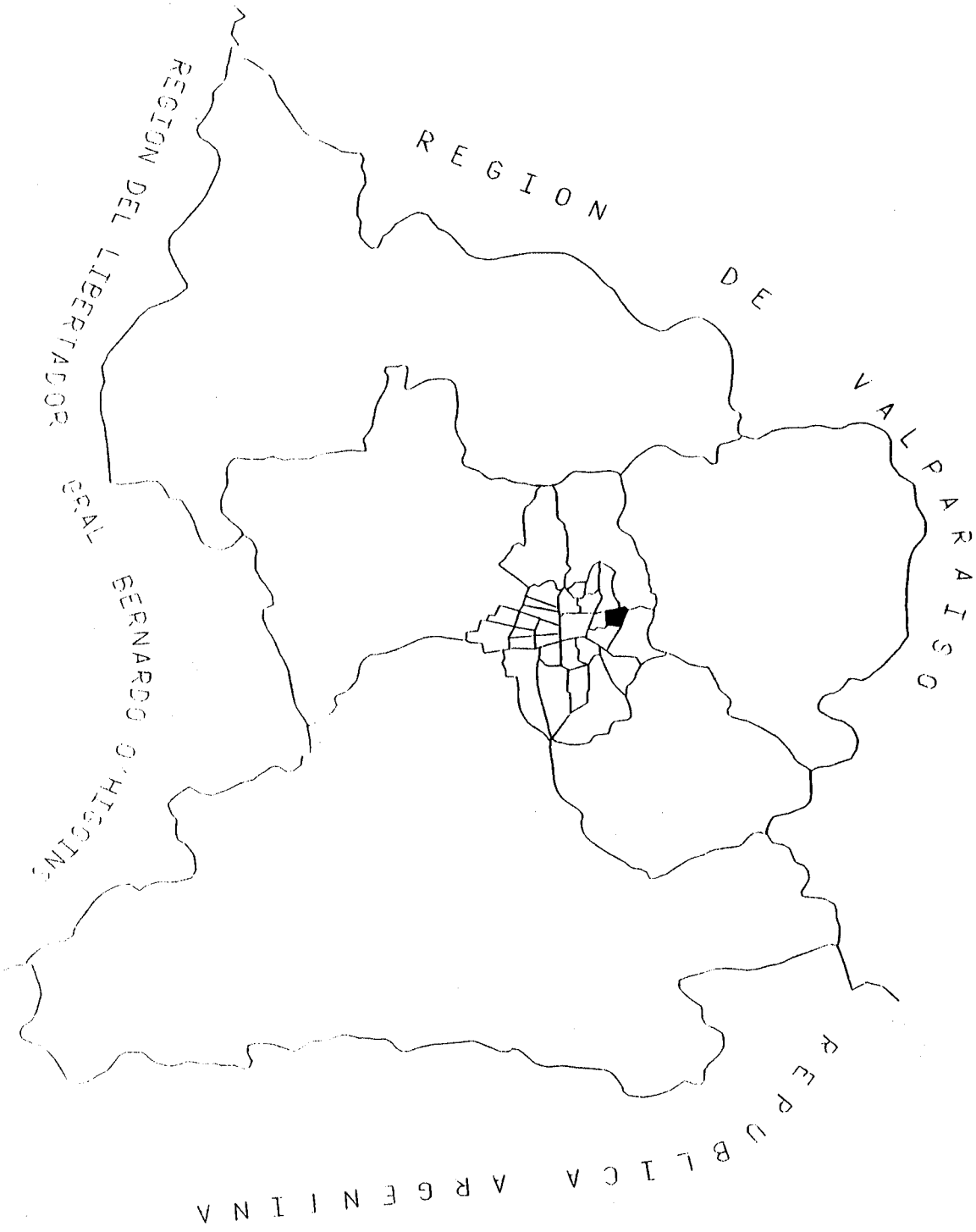
Variables	Prueba Nro.					
	1	2	3	4	5	(cont.)
N_A	150	150	150	150	150	
N_B	230	230	230	230	230	
N_{ab}	60	60	20	20	100	
c_A	2	2	2	2	2	
c_B	4	4	4	4	4	
σ^2_a	0.16	0.08	0.16	0.08	0.16	
σ^2_b	0.25	0.25	0.25	0.25	0.25	
σ^2_{ab}	0.20	0.12	0.20	0.12	0.20	
ß	0.26	0.26	0.133	0.133	0.67	
α	0.40	0.40	0.087	0.087	0.43	
$r_{opt}=n_A/n_B$	0.833	0.571	0.865	0.597	0.774	
Nro. iter.	6	7	5	6	8	

Variables	Prueba Nro.			
	6	7	8	9
N_A	150	150	220	220
N_B	230	230	290	290
N_{ab}	100	60	60	60
c_A	2	3	2	3
c_B	4	3.50	4	3.5
σ^2_a	0.08	0.16	0.16	0.08
σ^2_b	0.25	0.25	0.25	0.25
σ^2_{ab}	0.12	0.20	0.20	0.12
ß	0.67	0.26	0.21	0.21
α	0.43	0.40	0.27	0.27
$r_{opt}=n_A/n_B$	0.522	0.594	1.007	0.50
Nro. iter.	8	6	6	6

Como puede observarse sólo se requiere de un reducido número de iteraciones para obtener el r óptimo.

En el caso de N_a, N_b y N_{ab} desconocidos el sistema es bastante similar sólo que ahora intervienen también las medias correspondientes a cada dominio. Por dicho motivo no se consideró necesario incorporar la correspondiente simulación.

ANEXO 2


REGION DEL LIBERTADOR
REGION
DE
VALPARAISO
REGION METROPOLITANA DE SANTIAGO
REPUBLICA ARGENTINA

ANEXO 3

Teorema: Si se consideran las N! poblaciones finitas formadas por las N! permutaciones de cualquier conjunto de números $Y_1, Y_2, ..., Y_N$, entonces

$$E[Var(\bar{y}_{sis})] = Var(\bar{y}_{mas})$$

donde $\bar{y}_{sis}$ representa la media muestral de un muestreo sistemático, $\bar{y}_{mas}$ la media muestral de un muestreo aleatorio simple y E la esperanza tomada a través de las N! poblaciones equiprobables.

Demostración: se sabe que $V(\bar{y}_{mas}) = S^2/n\ (N-n)/N$

donde: $S^2 = 1/(N-1) \sum_{i=1}^{K} (y_i - \bar{Y})^2$

y

$$V(\bar{y}_{sis}) = 1/k \sum_{i=1}^{N} (\bar{y}_i - \bar{Y})^2$$

donde: $\bar{y}_i = \sum_{j=1}^{n} y_{ij}/n$

y donde "y_{ij}" se refiere a los elementos del cuadro siguiente:

Muestras posibles

1	2	...	i	...	k
y_{11}	y_{21}	...	y_{i1}	...	y_{k1}
y_{12}	y_{22}	...	y_{i2}	...	y_{k2}
y_{1j}	y_{2j}	...	y_{ij}	...	y_{kj}
y_{1n}	y_{2n}	...	y_{in}	...	y_{kn}

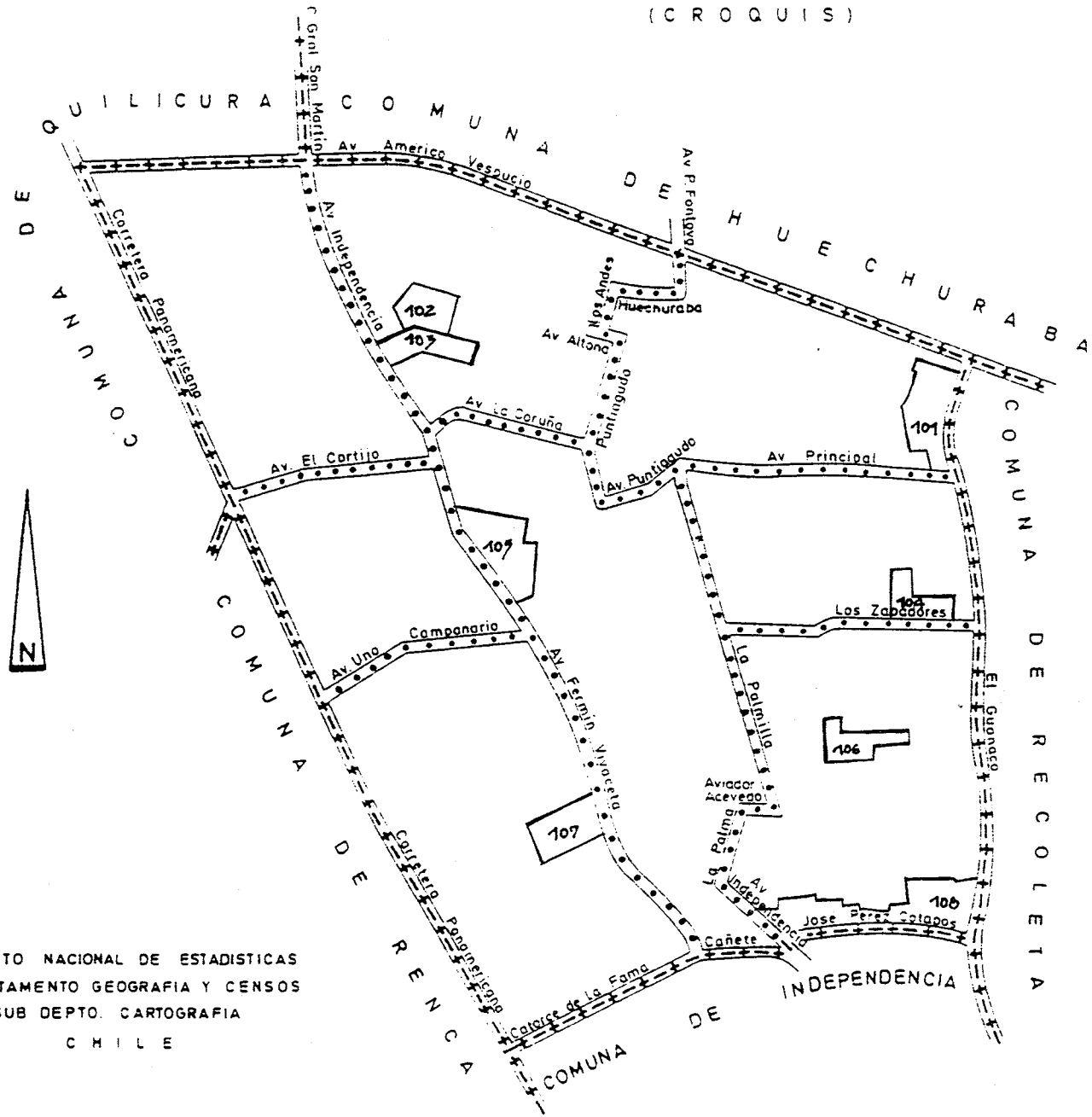
Si se considera que las N! permutaciones de la población son equiprobables (se indica con el supraíndice "s" cada una de las N! permutaciones equiprobables).

$$E[Var(\bar{y}_{sis})] = \sum_{s=1}^{N!} \{1/k \sum_{i=1}^{k} [\bar{y}_i^{(s)} - \bar{Y}]^2\}\ 1/N!$$

$$= 1/(N!k) \sum_{s=1}^{N!} \sum_{i=1}^{k} \{ \sum_{j=1}^{n} y_{ij}/n - \bar{Y} \}^2$$

COMUNA DE CONCHALI

(C R O Q U I S)



INSTITUTO NACIONAL DE ESTADISTICAS
DEPARTAMENTO GEOGRAFIA Y CENSOS
SUB DEPTO. CARTOGRAFIA
C H I L E

$$= \quad 1/(N!n^2k) \sum_{s=1}^{N!} \sum_{i=1}^{k} \{ \sum_{j=1}^{n} y_{ij}^{(s)} - n\bar{Y} \}^2$$

$$= \quad 1/(N!n^2k) \sum_{s}^{N!} \sum_{i}^{k} \{ \sum_{j}^{n} (y_{ij}^{(s)} - \bar{Y}) \}^2$$

$$= \quad 1/(N!n^2k) \sum_{s}^{N!} \sum_{i}^{k} \{ \sum_{j}^{n} (y_{ij}^{(s)} - \bar{Y})^2$$

$$+ \sum_{j \neq j'}^{n} (y_{ij}^{(s)} - \bar{Y})(y_{ij'}^{(s)} - \bar{Y}) \}$$

$$= \quad 1/(N!n^2k) \sum_{s}^{N!} \{ \sum_{i}^{k} \sum_{j}^{n} (y_{ij}^{(s)} - \bar{Y})^2 +$$

$$+ \sum_{i}^{k} \sum_{j \neq j'}^{n} (y_{ij}^{(s)} - \bar{Y})(y_{ij'}^{(s)} - \bar{Y}) \}$$

$$= \quad 1/(N!n^2k) \{ \sum_{s}^{N!} (N-1)S^2 +$$

$$+ kn(n-1)(N-2)! \sum_{u \neq u'}^{N!} (y_u - \bar{Y})(y_{u'} - \bar{Y}) \}$$

y como:

$$\sum_{u \neq u'}^{N!} (y_u - \bar{Y})(y_{u'} - \bar{Y}) \quad = \quad - (N-1)S^2$$

resulta:

$$\mathbf{E[Var(\bar{y}_{sis})]} \quad = \quad 1/(N!n^2k)\, S^2 \{N!\,(N-1) - kn(n-1)(N-2)!\,(N-1)\}$$

$$= \quad 1/(N!n^2k)\, S^2 \{N!\,(N-1) - N!(n-1)\}$$

$$= \quad S^2/Nn \{N-1-n+1\}$$

$$= \quad S^2/n\,(N-n)/N \;=\; \mathbf{Var(\bar{y}_{sis})}$$

ANEXO 4

El siguiente programa fue utilizado para generar muestras aleatorias, en base a la distribución del número de personas por vivienda, para el total y los totales por sexo. Para cada variable, se generaron 1000 muestras considerando las distribuciones I.2 e II.4.

Así, para cada caso, se calcularon 1000 valores posibles del estimador de Hartley, aplicándose luego un test de bondad de ajuste para probar el supuesto de normalidad.

```
'PROGRAMA: FORMULA.BAS
'FUNCION  : ESTIMACION DE TOTAL Y POR SEXO
'INPUT    : NINGUNO
'OUTPUT   : FORMULA.EXE , ESTIMADOS.DAT
'*****************************************

'Constantes

Cte1  = (113/8)  : Cte2  = (251/13)        : Cte3 =(284/12) : Cte4  = (254/13)
Cte5  = (296/12) : Cte6  = (328/14)        : Cte7 =(168/12) : Cte8  = (279/13)
Cte9  = (236/15) : Cte10 = (0.024*113/8)   : Cte11 =(54/2)  : Cte12 = 53
Cte13 = (58/2)   : Cte14 = (10/2)          : Cte15 = 15     : Cte16 = (126/3)
Cte17 = (39/1)   : Cte18 = (0.976*8500/1208)

' ** MAIN PROGRAM **

MTIMER

INPUT "deme el número de veces a repetir el proceso";cuantos

DIM YESTIMADO(cuantos)
RANDOMIZE MTIMER


FOR VECES = 1 TO CUANTOS 'cantidad de veces que repite proceso

    VAR1 = 0 : VAR5 = 0 : VAR9  = 0 : VAR13 = 0
    VAR2 = 0 : VAR6 = 0 : VAR10 = 0 : VAR14 = 0
    VAR3 = 0 : VAR7 = 0 : VAR11 = 0 : VAR15 = 0
    VAR4 = 0 : VAR8 = 0 : VAR12 = 0 : VAR16 = 0
```

```
'********************* Calculos ****************************
 for contador = 1 to 13
 GOSUB GENERA2
 VAR1 = VAR1 + sumando
 next contador
'-------------------------------------
 for contador = 14 to 25
 GOSUB GENERA2
 VAR2 = VAR2 + sumando
 next contador
'-------------------------------------
 for contador = 26 to 38
 GOSUB GENERA2
 VAR3 = VAR3 + sumando
 next contador
'-------------------------------------
 for contador = 39 to 50
 GOSUB GENERA2
 VAR4 = VAR4 + sumando
 next contador
'-------------------------------------
 for contador = 51 to 64
 GOSUB GENERA2
 VAR5 = VAR5 + sumando
 next contador
'-------------------------------------
 for contador = 65 to 76
 GOSUB GENERA2
 VAR6 = VAR6 + sumando
 next contador
'-------------------------------------
 for contador = 77 to 89
 GOSUB GENERA2
 VAR7 = VAR7 + sumando
 next contador
'-------------------------------------
 for contador = 90 to 104
 GOSUB GENERA2
 VAR8 = VAR8 + sumando
 next contador
'-------------------------------------
 for contador = 105 to 106
 GOSUB GENERA2
 VAR9 = VAR9 + sumando
 next contador
'-------------------------------------
 for contador = 107 to 107
 GOSUB GENERA2
 VAR10 = VAR10 + sumando
 next contador
'-------------------------------------
```

```
        for contador = 108 to 109
        GOSUB GENERA2
        VAR11 = VAR11 + sumando
        next contador
       '-----------------------------------
        for contador = 110 to 111
        GOSUB GENERA2
        VAR12 = VAR12 + sumando
        next contador
       '-----------------------------------
        for contador = 112 to 112
        GOSUB GENERA2
        VAR13 = VAR13 + sumando
        next contador
       '-----------------------------------
        for contador = 113 to 115
        GOSUB GENERA2
        VAR14 = VAR14 + sumando
        next contador
       '-----------------------------------
        for contador = 116 to 116
        GOSUB GENERA2
        VAR15 = VAR15 + sumando
        next contador
       '-----------------------------------
        for contador = 117 to 1324
        GOSUB GENERADOR
        VAR16 = VAR16 + sumando
        next contador
       '-----------------------------------
        '** Resolución de la Formula **
        Y = 0 : y1 = 0 : y2 = 0 : y3 =0

        y1 = Cte1  *

(Cte2*VAR1+Cte3*VAR2+Cte4*VAR3+Cte5*VAR4+Cte6*VAR5+
Cte7*VAR6+Cte8*VAR7+Cte9*VAR8)

        y2 = Cte10 *
 (Cte11*VAR9+Cte12*VAR10+Cte13*VAR11+Cte14*VAR12
+Cte15*VAR13+Cte16*VAR14+Cte17*VAR15)

        y3 = Cte18 * VAR16

        Y  = (y1 + y2 + y3)

        YESTIMADO(VECES) = Y

        ?"Y estimado (";VECES;") = "; Y,

NEXT VECES
```

```
?"GRABANDO DATOS"
' ******  graba  ESTIMADOS.DAT  *****
OPEN "O",#1,"A:\ESTIMADOS.DAT"
FOR i = 1 to cuantos
   WRITE #1,YESTIMADO(i)
NEXT i
CLOSE #1
?"FIN PROCESO , su disk contiene el archivo ESTIMADOS.DAT"


                                        END
'=========================================>>>>>>>

'*********************************************

 GENERADOR :
'genera un número aleatorio entero entre 1 y 1000
'y asigna sumando (valor) segun tabla

'*********************************************

aleatorio = int(RND(contador + 1) * 1000)

SELECT CASE aleatorio

        CASE 1   to 12
           sumando = 1
        CASE 13  to 99
           sumando = 2
       CASE 100 to 263
           sumando = 3
       CASE 264 to 495
           sumando = 4
       CASE 496 to 677
           sumando = 5
       CASE 678 to 787
           sumando = 6
       CASE 788 to 862
           sumando = 7
       CASE 863 to 915
           sumando = 8
       CASE 916 to 943
           sumando = 9
       CASE 944 to 966
           sumando = 10
       CASE 967 to 979
           sumando = 11
       CASE 980 to 986
           sumando = 12
       CASE 987 to 993
           sumando = 13
       CASE 994 to 997
           sumando = 14
```

```
        CASE 998
            sumando = 15
        CASE 999
            sumando = 16
        CASE 1000
            sumando = 18

END SELECT

                                    RETURN
'=========================================================>>>>>>>

'***********************************************

 GENERA2 :
'genera un número aleatorio entero entre 1 y 1000
'y asigna sumando (valor) segun tabla

'***********************************************

aleatorio = int(RND(contador + 1) * 1000)

SELECT CASE aleatorio

         CASE 1 to 27
            sumando = 1
         CASE 28 to 137
            sumando = 2
        CASE 138 to 294
            sumando = 3
        CASE 295 to 493
            sumando = 4
        CASE 494 to 678
            sumando = 5
        CASE 679 to 829
            sumando = 6
        CASE 830 to 883
            sumando = 7
        CASE 884 to 931
            sumando = 8
        CASE 932 to 979
            sumando = 9
        CASE 980 to 993
            sumando = 11
        CASE 994 to 1000
            sumando = 13
END SELECT
                                    RETURN
'=========================================================>>>>>>>
```

Resultados de la simulación y aplicación de test de Bondad de Ajuste

Las hipótesis a plantear en cada caso son:

H_0) $Y_H \sim N$
H_1) $Y_H \not\sim N$

Se utilizará un nivel de significación $\alpha = 0.05$ y la estadística X^2 para testar la hipótesis nula, esto es:

$$u = X^2_{ob} = \sum_{i=1}^{k} [(o_i - e_i)^2 / e_i]$$

Siendo o_i las frecuencias observadas y e_i son las frecuencias esperadas.

$$X^2_{ob} \sim X^2_{k-1-r}$$

donde k es el número de clases y r es el número de parámetros a estimar en caso que no estén especificados.

Se rechazará H_0) si $X^2_{ob} \geq X^2_{(k-1-r),(1-\alpha)}$

Total de personas (distribución I.2)

$\bar{y}$ = 181073.0

$\hat{\sigma}$ = 6573.226

Clases	o_i	z_i	pp_i	p_i	$e_i=np_i$	$[o_i-e_i]^2/e_i$
160300.8-164404.0	5	-3.16	0.0000	0.0055	5.5	0.045454
		-2.54	0.0055			
164404.0-166455.5	6	-2.54	0.0055	0.0077	7.7	0.375324
		-2.22	0.0132			
166455.5-168507.1	19	-2.22	0.0132	0.0149	14.9	1.128187
		-1.91	0.0281			
168507.1-170558.7	22	-1.91	0.0281	0.0267	26.7	0.827340
		-1.60	0.0548			
170558.7-172610.2	49	-1.60	0.0548	0.0437	43.7	0.642791
		-1.29	0.0985			
172610.2-174661.8	55	-1.29	0.0985	0.0650	65.0	1.538461
		-0.98	0.1635			
174661.8-176713.4	82	-0.98	0.1635	0.0911	91.1	0.909001
		-0.66	0.2546			
176713.4-178764.9	129	-0.66	0.2546	0.1086	108.6	3.832044
		-0.35	0.3632			
178764.9-180816.5	133	-0.35	0.3632	0.1208	120.8	1.232119
		-0.04	0.4840			
180816.5-182868.1	123	-0.04	0.4840	0.1224	122.4	0.002941
		0.27	0.6064			
182868.1-184919.6	112	0.27	0.6064	0.1160	116.0	0.137931
		0.59	0.7224			
184919.6-186971.2	80	0.59	0.7224	0.0935	93.5	1.949197
		0.90	0.8159			
186971.2-189022.8	78	0.90	0.8159	0.0710	71.0	0.690140
		1.21	0.8869			
189022.8-191074.3	39	1.21	0.8869	0.0488	48.8	1.968032
		1.52	0.9357			
191074.3-193125.9	29	1.52	0.9357	0.0307	30.7	0.094136
		1.83	0.9664			
193125.9-195177.5	18	1.83	0.9664	0.0178	17.8	0.002247
		2.15	0.9842			
195177.5-197229.1	8	2.15	0.9842	0.0089	8.9	0.091011
		2.46	0.9931			
197229.1-199280.6	9	2.46	0.9931	0.0041	4.1	5.856097
		2.77	0.9972			
199280.6-201332.2	4	2.77	0.9972	0.0027	2.7	0.625925
		3.08	0.9999			
TOTAL	1000					

u = 21.94838

$X^2_{16;0.95}$ = 26.3

Total de personas (distribución II.4)

$\bar{y}$ = 185749.3

$\hat{\sigma}$ = 6701.185

Clases	o_i	z_i	pp_i	p_i	$e_i=np_i$	$[o_i-e_i]^2/e_i$
166466.4-168609.4	2	-2.88	0.0020	0.0032	3.2	0.450000
		-2.56	0.0052			
168609.4-170752.4	4	-2.56	0.0052	0.0074	7.4	1.562162
		-2.24	0.0126			
170752.4-172895.4	17	-2.24	0.0126	0.0148	14.8	0.327027
		-1.92	0.0274			
172895.4-175038.4	26	-1.92	0.0274	0.0274	27.4	0.071532
		-1.60	0.0548			
175038.4-177181.4	45	-1.60	0.0548	0.0455	45.5	0.005494
		-1.28	0.1003			
177181.4-179324.4	70	-1.28	0.1003	0.0682	68.2	0.047507
		-0.96	0.1685			
179324.4-181467.4	100	-0.96	0.1685	0.0926	92.6	0.591360
		-0.64	0.2611			
181467.4-183610.4	114	-0.64	0.2611	0.1134	113.4	0.003174
		-0.32	0.3745			
183610.4-185753.3	142	-0.32	0.3745	0.1255	125.5	2.169322
		0.00	0.5000			
185753.3-187896.3	130	0.00	0.5000	0.1255	125.5	0.161354
		0.32	0.6255			
187896.3-190039.3	97	0.32	0.6255	0.1134	113.4	2.371781
		0.64	0.7389			
190039.3-192182.3	83	0.64	0.7389	0.0926	92.6	0.995248
		0.96	0.8315			
192182.3-194325.3	70	0.96	0.8315	0.0682	68.2	0.047507
		1.28	0.8997			
194325.3-196468.3	44	1.28	0.8997	0.0455	45.5	0.049450
		1.60	0.9452			
196468.3-198611.3	26	1.60	0.9452	0.0274	27.4	0.071532
		1.92	0.9726			
198611.3-200754.3	13	1.92	0.9726	0.0148	14.8	0.218918
		2.24	0.9874			
200754.3-202897.3	8	2.24	0.9874	0.0074	7.4	0.048648
		2.56	0.9948			
202897.3-205040.3	4	2.56	0.9948	0.0032	3.2	0.200000
		2.88	0.9980			
205040.3-209326.3	5	2.88	0.9980	0.002	2.0	4.500000
		3.52	0.9998			
TOTAL	1000				u =	13.89202

$X^2_{16;0.95}$ = 26.3

Total hombres (distribución I.2)

$\bar{y}$ = 88048.42

$\hat{\sigma}$ = 4338.621

Clase	o_i	z_i	pp_i	p_i	$e_i=np_i$	$[o_i-e_i]^2/e_i$
73980.8-75332.8	1	-3.24	0.0000	0.0017	1.7	0.288235
		-2.93	0.0017			
75332.8-76684.7	1	-2.93	0.0017	0.0027	2.7	1.070370
		-2.62	0.0044			
76684.7-78036.6	4	-2.62	0.0044	0.0060	6.0	0.666666
		-2.31	0.0104			
78036.6-79388.6	13	-2.31	0.0104	0.0124	12.4	0.029032
		-2.00	0.0228			
79388.6-80740.5	27	-2.00	0.0228	0.0237	23.7	0.459493
		-1.68	0.0465			
80740.5-82092.4	30	-1.68	0.0465	0.0388	38.8	1.995876
		-1.37	0.0853			
82092.4-83444.3	62	-1.37	0.0853	0.0593	59.3	0.122934
		-1.06	0.1446			
83444.3-84796.3	93	-1.06	0.1446	0.0820	82.0	1.475609
		-0.75	0.2266			
84796.3-86148.2	111	-0.75	0.2266	0.1034	103.4	0.558607
		-0.44	0.3300			
86148.2-87500.1	122	-0.44	0.3300	0.1183	118.3	0.115722
		-0.13	0.4483			
87500.1-88852.0	128	-0.13	0.4483	0.1270	127.0	0.007874
		0.19	0.5753			
88852.0-90203.9	101	0.19	0.5753	0.1162	116.2	1.988296
		0.50	0.6915			
90203.9-91555.9	90	0.50	0.6915	0.0995	99.5	0.907035
		0.81	0.7910			
91555.9-92907.8	78	0.81	0.7910	0.0776	77.6	0.002061
		1.12	0.8686			
92907.8-94259.7	53	1.12	0.8686	0.0550	55.0	0.072727
		1.43	0.9236			
94259.7-95611.6	41	1.43	0.9236	0.0355	35.5	0.852112
		1.74	0.9591			
95611.6-96963.6	25	1.74	0.9591	0.0207	20.7	0.893236
		2.05	0.9798			
96963.6-98315.5	11	2.05	0.9798	0.0113	11.3	0.007964
		2.37	0.9911			
98315.5-99667.4	3	2.37	0.9911	0.0052	5.2	0.930769
		2.68	0.9963			
99667.4-101019.3	6	2.68	0.9963	0.0023	2.3	5.952173
		2.99	0.9986			
TOTAL	1000				u =	18.39680

$X^2_{17;0.95}$ = 27.59

Total hombres (distribución II.4)

$\bar{y}$ = 90770.18

$\hat{\sigma}$ = 4246.188

Clases	o_i	z_i	pp_i	p_i	$e_i=np_i$	$[o_i-e_i]^2/e_i$
72579.8-79604.2	6	-4.28	0.0000	0.0043	4.3	0.672093
		-2.63	0.0043			
79604.2-81360.3	5	-2.63	0.0073	0.0089	8.9	1.708988
		-2.22	0.0132			
81360.3-83116.3	24	-2.22	0.0132	0.0227	22.7	0.074449
		-1.80	0.0359			
83116.3-84872.4	46	-1.80	0.0359	0.0464	46.4	0.003448
		-1.39	0.0823			
84872.4-86628.5	74	-1.39	0.0823	0.0812	81.2	0.638423
		-0.98	0.1635			
86628.5-88384.6	129	-0.98	0.1635	0.1242	124.2	0.185507
		-0.56	0.2877			
88384.6-90140.6	145	-0.56	0.2877	0.15271	152.7	0.388277
		-0.15	0.4404			
90140.6-91896.7	191	-0.15	0.4404	0.1660	166.0	3.765060
		0.27	0.6064			
91896.7-93652.8	144	0.27	0.6064	0.1453	145.3	0.011631
		0.68	0.7517			
93652.8-95408.9	95	0.68	0.7517	0.1104	110.4	2.148188
		1.09	0.8621			
95408.9-97164.9	76	1.09	0.8621	0.0724	72.4	0.179005
		1.51	0.9345			
97164.9-98921.0	38	1.51	0.9345	0.0381	38.1	0.000262
		1.92	0.9726			
98921.0-100677.1	15	1.92	0.9726	0.0175	17.5	0.357142
		2.33	0.9901			
100677.1-102433.1	7	2.33	0.9901	0.0069	6.9	0.001449
		2.75	0.9970			
102433.1-107701.4	5	2.75	0.9970	0.0030	3.0	1.333333
		3.99	1.0000			
TOTAL	1000					

u = 11.46726

$X^2_{12,0.95}$ = 21.03

Total mujeres (distribución I.2)

$\bar{y} = 92478.71$

$\hat{\sigma} = 3979.405$

Clases	o_i	z_i	pp_i	p_i	$e_i=np_i$	$[o_i-e_i]^2/e_i$
79595.28-80897.88	2	-3.24	0.0000	0.0018	1.8	0.022222
		-2.91	0.0018			
80897.88-82200.48	3	-2.91	0.0018	0.0031	3.1	0.003225
		-2.58	0.0049			
82200.48-83503.09	10	-2.58	0.0049	0.0070	7.0	1.285714
		-2.26	0.0119			
83503.09-84805.69	9	-2.26	0.0119	0.0149	14.9	2.336241
		-1.93	0.0268			
84805.69-86108.30	28	-1.93	0.0268	0.0280	28.0	0.000000
		-1.60	0.0548			
86108.30-87410.90	46	-1.60	0.0548	0.0472	47.2	0.030508
		-1.27	0.1020			
87410.90-88713.51	77	-1.27	0.1020	0.0691	69.1	0.903183
		-0.95	0.1711			
88713.51-90016.11	89	-0.95	0.1711	0.0965	96.5	0.582901
		-0.62	0.2676			
90016.11-91318.71	126	-0.62	0.2676	0.1183	118.3	0.501183
		-0.29	0.3859			
91318.71-92621.32	128	-0.29	0.3859	0.1301	130.1	0.033897
		0.04	0.5160			
92621.32-93923.92	144	0.04	0.5160	0.1246	124.6	3.020545
		0.36	0.6406			
93923.92-95226.53	90	0.36	0.6406	0.1143	114.3	5.166141
		0.69	0.7549			
95226.53-96529.13	91	0.69	0.7549	0.0912	91.2	0.000438
		1.02	0.8461			
96529.13-97831.74	69	1.02	0.8461	0.0654	65.4	0.198165
		1.35	0.9115			
97831.74-99134.34	42	1.35	0.9115	0.0410	41.0	0.024390
		1.67	0.9525			
99134.34-100436.9	22	1.67	0.9525	0.0247	24.7	0.295141
		2.00	0.9772			
100436.9-101739.5	14	2.00	0.9772	0.0129	12.9	0.093798
		2.33	0.9901			
101739.5-103042.1	5	2.33	0.9901	0.0059	5.9	0.137288
		2.65	0.9960			
103042.1-105647.3	5	2.65	0.9960	0.0040	4.0	0.250000
		3.31	1.0000			
TOTAL	1000					u = 14.88498

$X^2_{16;0.95} = 26.3$

Total mujeres (disribución II.4)

$\bar{y}$ = 95422.06

$\hat{\sigma}$ = 4203.726

Clases	o_i	z_i	pp_i	p_i	$e_i=np_i$	$[o_i-e_i]^2/e_i$
83767.78-84990.72	3	-2.77	0.0028	0.0038	3.8	0.168421
		-2.48	0.0066			
84990.72-86213.66	8	-2.48	0.0066	0.0077	7.7	0.011688
		-2.19	0.0143			
86213.66-87436.60	19	-2.19	0.0143	0.0144	14.4	1.469444
		-1.90	0.0287			
87436.60-88659.54	22	-1.90	0.0287	0.0250	25.0	0.360000
		-1.61	0.0537			
88659.54-89882.48	46	-1.61	0.0537	0.0397	39.7	0.999748
		-1.32	0.0934			
89882.48-91105.41	60	-1.32	0.0934	0.0581	58.1	0.062134
		-1.03	0.1515			
91105.41-92328.35	96	-1.03	0.1515	0.0782	78.2	4.051662
		-0.74	0.2297			
92328.35-93551.29	82	-0.74	0.2297	0.0967	96.7	2.234643
		-0.45	0.3264			
93551.29-94774.23	91	-0.45	0.3264	0.1140	114.0	4.640350
		-0.15	0.4404			
94774.23-95997.17	113	-0.15	0.4404	0.1153	115.3	0.045880
		0.14	0.5557			
95997.17-97220.11	109	0.14	0.5557	0.1107	110.7	0.026106
		0.43	0.6664			
97220.11-98443.04	112	0.43	0.6664	0.0978	97.8	2.061758
		0.72	0.7642			
98443.04-99665.98	89	0.72	0.7642	0.0796	79.6	1.110050
		1.01	0.8438			
99665.98-100888.9	56	1.01	0.8438	0.0594	59.4	0.194612
		1.30	0.9032			
100888.9-102111.8	40	1.30	0.9032	0.0409	40.9	0.019804
		1.59	0.9441			
102111.8-103334.8	25	1.59	0.9441	0.0259	25.9	0.031274
		1.88	0.9700			
103334.8-104557.7	14	1.88	0.9700	0.0150	15.0	0.066666
		2.17	0.9850			
104557.7-105780.6	7	2.17	0.9850	0.0081	8.1	0.149382
		2.46	0.9931			
105780.6-107003.6	4	2.46	0.9931	0.0040	4.0	3.2E-30
		2.76	0.9971			
107003.6-108226.5	4	2.76	0.9971	0.0027	2.7	0.625925
		3.05	0.9998			
TOTAL	1000				u = 18.32955	
					$X^2_{17;0.95}$ = 27.59	

De la comparación de cada X^2_{ob} con su respectivo $X^2_{crit.}$, se concluye que en ningún caso se rechaza el supuesto de normalidad, a un nivel de significación del 5 por ciento.

ANEXO 5

I. Test de hipótesis para las proyecciones

Para evaluar la validez de la proyección de CELADE, se usaron test de validación mediante las hipótesis nula y alternativa:

$$H_0)\ Y = Y_{proyectado}$$

$$H_1)\ Y \neq Y_{proyectado}$$

Se utiliza un nivel de significación del 5 por ciento. La estadística a usar es:

$$z_{obs.} = (\hat{Y}_{dimj} - Y_{proy.})/\sigma(\hat{Y}_{dimj})$$

la cual bajo la hipótesis nula se distribuye normalmente (en los test de bondad de ajuste no se rechazó la hipótesis de normalidad). Entonces, de acuerdo al nivel de significación prefijado se rechaza H_0, toda vez que $z_{obs.}$ es mayor o igual a 1.96 o menor o igual a -1.96.

Luego, se utiliza la estimación más alejada de cada distribución para testar la validez de la proyección.

TOTAL

$H_0)\ Y = 170667$ (Y proyectado)
$H_1)\ Y \neq 170667$

$\hat{Y}_{dlm3} = 179204$ $\qquad Z_{1obs.} = 0.776412$
$\hat{Y}_{dllm3} = 172208$ $\qquad Z_{2obs.} = 0.148754$

En este caso, la evidencia muestral lleva a no rechazar la hipótesis de que el valor correspondiente al total de población en Conchalí sea 170667, con una confianza del 95 por ciento.

MUJERES

$H_0)\ Y = 88193$ (Y proyectado)
$H_1)\ Y \neq 88193$

$\hat{Y}_{dlm3} = 92569$ $\qquad Z_{1obs.} = 0.821851$
$\hat{Y}_{dllm3} = 89056$ $\qquad Z_{2obs.} = 0.166302$

HOMBRES

H_0) Y = 82474 (Y proyectado)
H_1) Y $\neq$ 82474

$\hat{Y}_{dIm3} = 86639$ $\quad Z_{1obs.} = 0.631032$

$\hat{Y}_{dIIm3} = 83156$ $\quad Z_{2obs.} = 0.112080$

También en el caso de desagregar la población por sexo, las estimaciones permiten aceptar las proyecciones realizadas para dicha Comuna.

No se consideró conveniente aplicar test para analizar las proyecciones por grandes grupos de edad ya que las estimaciones en general no resultaban confiables y/o precisas.