UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE ESTUDIOS SUPERIORES
"ACATLAN"

CALCULO DE LA VARIANZA EN UNA ENCUESTA COMPLEJA

T E S I N A

QUE PARA OBTENER EL TITULO DE:

A C T U A R I O

P R E S E N T A :

ERNESTO REYES GUTIERREZ

ASESOR: LUIS ALEJANDRO TAVERA PEREZ

2004

DEDICATORIA:

A LA ABUELITA Y A MIS PADRES

ÍNDICE

RESUMEN

El método de los conglomerados últimos, es empleado para estimar la varianza de las estimaciones generadas a través de una encuesta por muestreo, probabilística y compleja. En este trabajo, dicho método es aplicado a la Encuesta Nacional de Empleo Urbano que se realizó en el Área Metropolitana de Chihuahua durante el primer trimestre del 2003.

PALABRAS CLAVE: conglomerados últimos; varianza; encuesta compleja; estimación.

ABSTRACT

The ultimate clusters method is used to estimate the variance of the estimations generated from a probabilistic complex sample survey. In this work, the method is applied to the National Urban Employment Survey carried out at the Metropolitan Area of Chihuahua during the first trimester of 2003.

KEY WORDS: ultimate clusters; variance; complex survey; estimation.

INTRODUCCIÓN

En los últimos años en nuestro país, las encuestas por muestreo han tenido una importante demanda por parte de los diferentes usuarios de información estadística. Siendo las encuestas por muestreo del tipo probabilístico las más demandadas. Estas se caracterizan entre otras cosas, por generar estimaciones de totales, razones, medias, medianas, proporciones y demás parámetros poblacionales, a partir de una muestra extraída de una población de interés. En consecuencia, es necesario que dichas estimaciones cuando son publicadas o presentadas al usuario de la información, estén acompañadas de una medida de su error, el error estándar, raíz cuadrada de la varianza.

En encuestas por muestreo probabilísticas del tipo complejo su diseño de muestreo con frecuencia se caracteriza por ser estratificado, por conglomerados y polietápico. En especial, la última característica complica la formulación del estimador de la varianza, debido a que se tiene que calcular la varianza en cada una de las etapas de muestreo para obtener la varianza del estimador del parámetro poblacional correspondiente.

El presente trabajo tiene como objetivo hacer la presentación y aplicación del método de los conglomerados últimos para estimar la varianza en una encuesta por muestreo del tipo complejo. Dicho método se caracteriza por generar estimaciones de la varianza sin tener que estimarla en cada una de las etapas de muestreo. Y así, permitir al usuario de la información decidir si son o no confiables, las estimaciones de los valores poblacionales. De manera breve se puede decir, que el método de los conglomerados últimos se caracteriza por ser estadísticamente confiable y fácil de aplicar, obteniéndose de esta manera estimaciones de la varianza rápidas, insesgadas o sesgadas y consistentes.

La aplicación en este trabajo del método de los conglomerados últimos, consiste en estimar la varianza para algunas estimaciones de totales y razones simples generadas por la Encuesta Nacional de Empleo Urbano, levantada por el Instituto Nacional de Estadística Geografía e Informática (INEGI), en el Área Metropolitana de Chihuahua.

A continuación se describe brevemente el contenido de cada uno de los capítulos que conforman la presente tesina.

En el Capítulo 1 se presenta la utilidad de la encuesta por muestreo, además, de una serie de conceptos y definiciones, que serán de utilidad para la comprensión del método de los conglomerados últimos y el ámbito de su aplicación. El Capítulo 2, presenta los estimadores del total y la razón simple para un diseño de muestreo estratificado, por conglomerados y trietápico, así como sus correspondientes estimadores de varianza bajo el método de los conglomerados últimos. Debido a que es importante tener una estimación del efecto de diseño, también son definidos los estimadores de varianza bajo el muestreo aleatorio simple, para los estimadores del total y la razón simple. El estimador de la varianza por el método de los conglomerados últimos para el estimador del total es presentado junto con los supuestos que debe cumplir. Todos los estimadores antes mencionados son acompañados de la bibliografía respectiva en notas a pie de página, con el fin de que el lector interesado recurra a ella en busca de mayor información. En el

Capítulo 3, se presenta la utilidad, instalación y forma de empleo del programa de cómputo VARCU (Varianza por Conglomerados Últimos), que permite obtener estimaciones del error estándar, coeficiente de variación y efecto de diseño, de las estimaciones del total y la razón simple. En el Capítulo 4, son presentadas y comentadas entre otras cosas, las precisiones estadísticas de las distintas estimaciones de totales y razones simples, generadas por la Encuesta Nacional de Empleo Urbano que se levanta en el Área Metropolitana de Chihuahua, así como un resumen del diseño de muestreo de la encuesta.

Finalmente, se presentan las conclusiones con respecto al método de los conglomerados últimos y de las estimaciones que de él se generan.

Sin duda el presente documento le será de utilidad a estudiantes de las carreras de Actuaría, Estadística, Matemáticas y otras carreras afines, que estén en busca de aplicaciones de los conocimientos adquiridos en materias como estadística, probabilidad, muestreo, cálculo diferencial y computación. Así como, para aquellos usuarios de la información estadística proveniente de una encuesta por muestreo probabilística, compleja o no.

1 MARCO TEÓRICO

1.1 Utilidad de la encuesta por muestreo

La necesidad de información, en particular de la estadística, ha obligado al desarrollo de técnicas para su recolección y análisis: censos, registros administrativos y encuestas por muestreo conforman las técnicas de recolección de datos. En este trabajo se presentan fórmulas que permiten medir la precisión de la información recabada a través de una encuesta por muestreo.

"*Una encuesta por muestreo es una técnica que permite hacer inferencias sobre la población de la que fue seleccionada la muestra*".[1] Su finalidad es "*obtener información para satisfacer una necesidad*",[2] que no queda satisfecha con las estadísticas existentes. Si a esto se agrega que dicha información se puede obtener de manera confiable a un bajo costo y en un corto tiempo, esta técnica de generación de información estadística, se convierte en una importante herramienta para la toma de decisiones, tanto en el sector público como en el privado.

En la mayoría de los países, existe una oficina gubernamental encargada de obtener información estadística a través de las encuestas por muestreo. Los temas que cubren están relacionados con "*la población total; su distribución por área, sexo, edad y otras características socioeconómicas; la tasa de crecimiento de la población; la migración interna y varios otros aspectos*";[3] "*la proporción de la tierra dedicada a la agricultura, las áreas bajo diferentes cultivos, las que sostienen pastizales y bosques, la producción de alimentos (cereales, frutas, etc.) y el número y calidad del ganado*",[4] así como del empleo, la industria y varios temas más. Generalmente los gobiernos emplean la información recabada para evaluar y diseñar programas de desarrollo nacional.

A continuación, se describirá brevemente parte de la información estadística que generan las encuestas por muestreo en hogares que levanta el Instituto Nacional de Estadística, Geografía e Informática (INEGI) en nuestro país.

Tres de las principales encuestas a nivel nacional son la Encuesta Nacional de Empleo Urbano (ENEU), la Encuesta Nacional de Empleo (ENE) y la Encuesta Nacional de los Ingresos y Gastos de los Hogares (ENIGH). Las dos primeras cubren la temática de empleo y desempleo para la población de 12 años y más, residente habitual de la vivienda, así como de aspectos demográficos generales a nivel de áreas metropolitanas y entidades federativas respectivamente. De manera particular se obtienen entre otras, estimaciones por sexo y grupos de edad de los totales correspondientes a las poblaciones económicamente activa e inactiva, desocupada abierta y ocupada, siendo esta última de la cual se deriva una cantidad importante de estimaciones. Sin embargo, es la tasa de desempleo abierto, cociente de la población desocupada abierta entre la población económicamente activa, una de las estimaciones más importante sobre el desempleo que se publica a la fecha.

[1] Azorín, Francisco. Métodos y Aplicaciones de Muestreo. Alianza Editorial. Madrid, 1986. p. 34.
[2] Raj, Des. Teoría del Muestreo. FCE. México, 1992. p. 34.
[3] Idem, p. 34.
[4] Idem, p. 35.

Por su parte la ENIGH, genera información estadística a nivel nacional para localidades de 2,500 y más habitantes y menores de 2,500. Las estimaciones que proporciona son entre otras, totales para el ingreso y gasto, monetario y no monetario, de las percepciones y erogaciones totales, así como del ingreso y gasto total.[5]

1.2 Concepto de población, marco muestral y muestra

Una vez que se ha decidido emplear la encuesta por muestreo para obtener determinada información estadística, se debe definir de manera concreta dentro de los objetivos de la encuesta, la población a cubrir y la información a recabar.

La definición de la población a estudiar o población objetivo, se hace en forma conjunta con la de sus elementos, debido a que "*los elementos de una población son las unidades de las que se busca la información, es decir, son las unidades del análisis, y su naturaleza se determina mediante los objetivos de la encuesta; la población es el agregado de los elementos*".[6]

Definida conceptualmente la población, el siguiente paso es seleccionar la muestra. Para realizar esta importante actividad, "*la población debe ser dividida en partes llamadas unidades de muestreo o unidades. Estas deben cubrir la totalidad de la población y no traslaparse en el sentido de que todo elemento de la población pertenezca a una y solamente una unidad*".[7] Así, cuando se lleva a cabo el muestreo directo de elementos, cada unidad de muestreo contendrá solamente a un elemento. En otros casos, la unidad de muestreo corresponderá a un conjunto de elementos llamado conglomerado, dicha unidad se emplea comúnmente en el muestreo polietápico, el cual más adelante será explicado

Ya identificadas o creadas las unidades de muestreo, estas son agrupadas y organizadas siguiendo algún criterio, con el fin de conformar el marco muestral o marco de la encuesta por muestreo, a partir del cual se seleccionará la muestra (marcos de viviendas, empresa, etc.).

A continuación, se dará una definición más detallada de lo que es un marco muestral citada por Särndal (1993) y debida a Lessler (1982):

> "*Marco. Los materiales o recursos que delimitan, identifican y permiten el acceso a los elementos de la población objetivo. En una encuesta por muestreo, las unidades del marco son las unidades a las cuales el esquema de muestreo probabilístico es aplicado.*
>
> *El marco también incluye cualquier información auxiliar (medidas de tamaño, información demográfica) que es empleada en:*
>
> a. *Técnicas de muestreo especiales, tales como la estratificación y selección muestral con probabilidad proporcional al tamaño.*
> b. *Técnicas de estimación especiales, tales como la razón o estimación de regresión.*"[8]

De la definición anterior se desprende que en el muestreo probabilístico, la muestra es un conjunto de unidades de muestreo seleccionadas con un esquema de muestreo probabilístico del marco muestral. En algunas situaciones esta selección se llega a realizar a partir de varios marcos.

[5] Para mayor información consulte la página de internet que tiene dicho Instituto.
[6] Kish, Leslie. Muestreo de Encuestas. Editorial Trillas. México, 1982. p. 27.
[7] Cochran, William. Técnicas de Muestreo. CECSA. México, 1992. p. 26.
[8] Särndal, Carl-Erick, et. al. Model Assisted Survey Sampling. Springer-Verlag. New York, 1993. p. 9.

1.3 Tipos de muestreo

Con base en el esquema que se sigue para seleccionar la muestra de unidades de muestreo, se distinguen dos tipos de muestreo: el probabilístico y el no probabilístico. Este último, caracterizado principalmente por no emplear la teoría de probabilidades en la selección de las unidades de muestreo. Nosotros nos avocaremos exclusivamente al probabilístico.

1.3.1 Muestreo probabilístico

Aunque en líneas anteriores se ha mencionado que las unidades de muestreo del marco muestral son seleccionadas con algún esquema o método de muestreo probabilístico. No se ha establecido en que consiste este procedimiento.

El muestreo probabilístico es una forma de seleccionar la muestra, que satisface las siguientes condiciones:

1. Permite definir el conjunto de muestras $L = \{ s_1, s_2, s_3, \ldots\ldots\ldots, s_m \}$ que son posibles de obtener con el procedimiento de muestreo.
2. Asocia una probabilidad conocida de selección $p(s)$ con cada muestra posible s, definiendo una distribución de probabilidad.
3. Asigna a cada unidad de muestreo una probabilidad de selección distinta a cero (probabilidad de inclusión).
4. Al seleccionar una muestra con un mecanismo aleatorio cada posible muestra recibe exactamente la probabilidad $p(s)$.

Una muestra obtenida bajo estos cuatro requerimientos es llamada una muestra probabilística.[9]

De la lista de condiciones antes mencionada la número tres tiene especial importancia, debido a que la probabilidad de selección de cada unidad de muestreo puede ser igual o desigual. Esta última, determinada con base en alguna medida de tamaño de la unidad de muestreo. En la práctica, una muestra probabilística no se obtiene proporcionando el conjunto de muestras L y la probabilidad de selección de sus elementos $p(s)$. Lo que se hace es especificar la probabilidad de inclusión de las unidades de muestreo en la muestra, extrayéndose una a la vez o en grupos hasta constituir la muestra del tamaño y tipo deseado.[10]

Algunos métodos de muestreo probabilísticos son: el muestreo aleatorio simple con o sin reemplazo, el muestreo sistemático, el muestreo con probabilidad proporcional al tamaño con o sin reemplazo, el método de Midzuno y varios más.[11]

Todos los métodos de muestreo probabilísticos se pueden aplicar a marcos muestrales de elementos o conglomerados, que pueden estar o no estratificados. Más adelante se describirán con detalle las características de los más empleados en la práctica.

---

[9] Särndal, Carl-Erick, et. al. Model Assisted Survey Sampling. Springer-Verlag. New York, 1993. p. 8.
[10] Cochran, William. Op cit., p. 29.
[11] Véase:K.R.W. Brewer and Muhammad Hanif. Sampling with Unequal Probabilities. Springer-Verlag. New York Heidelberg Berlin. Donde se lista un número importante de procedimientos de muestreo.

1.4 Fuentes de error en las encuestas

Es requisito básico que los resultados obtenidos a través de una encuesta por muestreo probabilística, sean acompañados de medidas que cuantifiquen y permitan evaluar su precisión. Esto como consecuencia de las fuentes de error que los afectan, las cuales son agrupadas en dos categorías: error de muestreo y errores no de muestreo (de cobertura, en las observaciones, etc.).[12] Siendo el primero al que se encuentra dedicado el presente trabajo.

1.4.1 Error de muestreo

Se ha mencionado en párrafos previos, que un esquema de muestreo probabilístico aplicado a la población de interés para obtener una muestra de tamaño n, podría producir una enorme cantidad de muestras posibles, las cuales generarían sus respectivas estimaciones del valor poblacional de interés. Todas ellas variando en diferentes cantidades con respecto al valor poblacional. Esta variación es lo que determina la magnitud del error de muestreo, pues el error de muestreo se debe al hecho de observar únicamente a la muestra y no a la población completa,[13] y su magnitud depende de la variación existente entre las diferentes estimaciones producidas por las distintas muestras posibles y el valor poblacional.[14] Dicha magnitud disminuye conforme se incrementa el tamaño de muestra n. La forma más comúnmente empleada para medir el error de muestreo, es a través del error estándar de la estimación, definido como la raíz cuadrada de la varianza.[15]

La varianza verdadera generalmente no es posible calcularla, debido a que involucra valores poblacionales desconocidos. Por lo que, se obtiene una estimación a partir de la información contenida en la muestra. Es importante señalar que de los puntos que caracterizan a una muestra probabilística se desprende que el tipo de esquema de muestreo probabilístico empleado para seleccionar la muestra, determina la forma del estimador del parámetro poblacional y su varianza. El presente trabajo proporciona un método para estimar la varianza del total y la razón simple en encuestas por muestreo complejas.

1.5 Algunos conceptos básicos y notación

A continuación se proporcionan definiciones y notaciones, algunas de ellas tratadas con anterioridad, que serán de utilidad en lo sucesivo.

Población y variables

Se considerará una población finita de N elementos etiquetados por $k=1,\ldots..,N$ o simplemente una población a encuestar

$$U = \{ u_1, \ldots\ldots\ldots, u_k, \ldots\ldots\ldots, u_N \}$$

con el fin de simplificar, el elemento k-ésimo será representado por su etiqueta k, así que la población finita será denotada por

$$U = \{ 1, \ldots\ldots\ldots, k, \ldots\ldots\ldots, N \}$$

[12] Azorín, Francisco. Op cit., p. 44.
[13] Kish, Leslie. Op cit., p. 30.
[14] Idem, p. 32.
[15] Idem, p. 32.

Sean

y = *Variable de interés o bajo estudio de la población*
Y_i = *Valor de y en el elemento i-ésimo de la población*

Letras como *x, y, z* y otras, se emplearán para identificar las variables de interés de la población a encuestar.[16]

Parámetros poblacionales

Un parámetro es una función de una o más variables de interés.[17] Algunos de los principales parámetros de una población finita son:

$$Y = \sum_{i=1}^{N} Y_i \qquad \textit{Total poblacional}$$

$$\bar{Y} = \frac{1}{N} Y \qquad \textit{Media poblacional por elemento}$$

$$R = \frac{Y}{X} \qquad \textit{Razón de totales poblacionales}$$

$$D = \frac{Y}{X} - \frac{W}{Z} \qquad \textit{Diferencia entre razones}$$

Donde X, W y Z se definen de manera análoga a Y, de manera general se representará al parámetro poblacional con la letra griega θ.

Diseño muestral

El objetivo de una encuesta por muestreo probabilística, como se menciono al principio del capítulo, es el de obtener información estadística con base en una muestra de la población finita de interés. En otras palabras, obtener estimaciones de los parámetros poblacionales con base en una muestra de la población U, denotada por s. La cual es una de las muestras posibles del conjunto L, que es extraída de U a través de un esquema de selección muestral o muestreo probabilístico, estableciendo una probabilidad de selección $p(s)$ para cada muestra posible. De manera general la función $p(\cdot)$ es llamada *diseño muestral*, la cual permite determinar las propiedades estadísticas de la estimación.[18] Un diseño muestral se llama no informativo si y sólo si, $p(\cdot)$ no depende de los valores de la variable de interés.[19]

Estimador

El estimador de un parámetro poblacional es una fórmula empleada para obtener la estimación con base en la muestra seleccionada. Un estimador $\hat{\theta}$ se dice que es insesgado con respecto al diseño muestral, cuando su valor esperado es igual al valor poblacional θ, esto es $E(\hat{\theta})=\theta$, de otra manera se dice que es sesgado.[20]

[16] Lehtonen, Risto and Pahkinen, Erkki J. Practical Methods for Design of Complex Surveys. John Wiley & Sons. England, 1995. p. 10.
[17] Idem, p. 10.
[18] Särndal, Carl-Erick, et. al. Op cit., p. 27.
[19] Wolter, Kirk M. Introduction to Variance Estimation. Springer-Verlag. New York, 1985.p.8.
[20] Cochran, William. Op Cit., p. 46.

En la práctica los estimadores sesgados no son descartados por completo (razón y regresión), debido a que sus estimaciones son muy próximas al parámetro poblacional a estimar. El sesgo de un estimador $\hat{\theta}$ se define como la diferencia entre el valor esperado del estimador y el valor poblacional, es decir, $sesgo(\hat{\theta})=E(\hat{\theta})-\theta$. Generalmente este sesgo disminuye a medida que el tamaño de muestra n se incrementa.[21]

A continuación se presentan algunos de los principales estimadores de parámetros poblacionales:

$$\hat{Y} = Estimador\ del\ total\ poblacional\ Y$$
$$\hat{\bar{Y}} = Estimador\ de\ la\ media\ poblacional\ por\ elemento\ \bar{Y}$$
$$\hat{R} = Estimador\ de\ la\ razón\ poblacional\ R$$
$$\hat{D} = Estimador\ de\ la\ diferencia\ entre\ razones\ D$$

Es importante señalar que la formulación del estimador depende del esquema de selección muestral empleado para obtener la muestra. Por último, un diseño de muestreo es aquel en el cual se especifica el tipo de muestreo probabilístico y el método de estimación.[22]

Varianza del estimador

En los párrafos anteriores se estableció que un estimador $\hat{\theta}$ es insesgado si su valor esperado o valor promedio de la distribución muestral coincide con el parámetro poblacional θ. La distribución muestral representa la fluctuación aleatoria de la estimación debida al diseño muestral, y esa variabilidad es medida por el error estándar. Llamado también desviación estándar de la distribución muestral, definida como la raíz cuadrada de la varianza de la distribución muestral o simplemente la varianza del estimador $\hat{\theta}$.[23] Así que, para un estimador insesgado $\hat{\theta}$ su varianza se define como

$$V(\hat{\theta}) = E(\hat{\theta} - E(\hat{\theta}))^2 .$$

Generalmente, en la práctica no es posible calcular la varianza del estimador $V(\hat{\theta})$, debido a que se desconocen valores poblacionales y únicamente se cuenta con datos de una sola muestra. Por lo que, siempre se procede a obtener una estimación de la varianza verdadera del estimador $\hat{\theta}$, denotada por $v(\hat{\theta})$.

En el caso de estimadores sesgados, el diseño muestral $p(\cdot)$ también permite describir una distribución muestral del estimador, cuyo valor esperado en este caso diferirá cierta cantidad del parámetro poblacional. Esa cantidad se conoce con el nombre de sesgo del estimador $\hat{\theta}$, definido con anterioridad. Por tanto, una medida que permite conocer esta dispersión del estimador con respecto al parámetro poblacional, es el error cuadrático medio,[24] definido de la siguiente manera

$$\begin{aligned} ECM(\hat{\theta}) = E(\hat{\theta} - \theta)^2 &= V(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2 \\ &= V(\hat{\theta}) + [sesgo(\hat{\theta})]^2 \end{aligned}$$

[21] Kish, Leslie. Op cit., p. 34.
[22] Idem, p. 37.
[23] Idem, p. 32.
[24] Idem, p. 34.

En algunos estimadores como es el caso del estimador de razón $\hat{R}$, el sesgo se vuelve despreciable en muestras grandes. En cuyo caso el $ECM(\hat{R}) \approx V(\hat{R})$.[25] En relación a las fuentes de sesgo, Kish menciona las siguientes tres:

- el tipo de estimador empleado,
- empleo de marcos muestrales imperfectos, y
- errores en las observaciones.

Por tanto, el error cuadrático medio involucra además del error debido al muestreo, el error ajeno al muestreo.[26] En este trabajo se considerará únicamente a la primera fuente de sesgo para su cálculo.

Muestreo estratificado

En este tipo de muestreo los elementos de la población o unidades de muestreo que conforman el marco muestral, son agrupados en subpoblaciones o subgrupos no traslapados lo más homogéneos posibles, con el fin de mejorar la precisión de las estimaciones. Las variables más comúnmente utilizadas para la estratificación son aquellas de carácter regional (localidades, municipios), demográfico (sexo, grupo de edad) y socioeconómico (ingreso), todas ellas obtenidas a partir de un censo y que conforman la información auxiliar de los marcos muestrales.[27] Así que, la división de la población en subgrupos homogéneos llamados estratos, deberá observar la siguiente característica:

$$\sum_{h=1}^{L} N_h = N$$

donde

L = *Número de estratos*
N = *Tamaño de la población*
N_h = *Tamaño de la población en el estrato h-ésimo*

Una vez conformados los estratos, lo que sigue es distribuir el tamaño de muestra n en cada uno de ellos. De tal manera que

$$\sum_{h=1}^{L} n_h = n$$

donde

L = *Número de estratos*
n_h = *Tamaño de muestra en el estrato h-ésimo*

La distribución de la muestra total en cada uno de los estratos, se conoce con el nombre de **afijación de la muestra**. Para llevar a cabo esto, se cuenta con al menos tres procedimientos, la afijación igual, proporcional y óptima.[28]

Asignado el tamaño de muestra en los estratos, una muestra probabilística se selecciona de manera independiente en cada uno de ellos, calculándose posteriormente a la recolección de la información, las correspondientes estimaciones y varianzas por estrato, útiles en la obtención de la estimación global del parámetro poblacional y de su varianza.

[25] Cochran, William. Op cit., p. 56.
[26] Kish, Leslie. Op cit., p. 34.
[27] Lehtonen, Risto and Pahkinen, Erkki J. Op cit., p. 67.
[28] Idem, p. 70.

Muestreo por conglomerados

En encuestas de mediano o gran tamaño no es posible llevar a cabo el muestreo directo de los elementos de la población. Lo que hace necesario conformar unidades de muestreo que los agrupen, las cuales reciben el nombre de conglomerados de elementos.[29] Son dos las razones principales para la aplicación del muestreo por conglomerados:

i. No disponer de un marco muestral donde las unidades de muestreo identifiquen directamente a los elementos individuales de la población (marco de lista), cuya elaboración representaría un costo elevado.

ii. Una muestra proveniente de un marco de lista se caracteriza por una amplia dispersión, la cual representaría un alto costo económico para la recolección de la información. Siendo afectado también el trabajo de supervisión, lo que podría conducir a una alta no respuesta y a severos errores de medición.[30]

En la práctica los conglomerados pueden ser naturales y con un número de elementos distinto. Rara vez se encuentran conglomerados de igual tamaño. Ejemplos de conglomerados son escuelas, viviendas, casillas electorales, manzanas de viviendas en las ciudades, salones de clases, paquetes de productos resultado de un proceso manufacturero, etc.. El tamaño de los conglomerados generalmente responde a condiciones dependientes de los recursos disponibles y de la situación de la encuesta.[31]

Por otra parte, de manera breve se puede decir que el muestreo por conglomerados incrementa su eficiencia, si los conglomerados son internamente heterogéneos y la varianza entre conglomerados es baja. Ocurriendo lo contrario, sí los conglomerados son internamente homogéneos y la varianza entre conglomerados alta.[32]

Una vez conformado el marco muestral de conglomerados, se procede a obtener una muestra probabilística, que puede seleccionarse con probabilidad igual o desigual. La primera se utiliza cuando no se cuenta con información auxiliar en el marco muestral, complementaria al identificador del conglomerado o cuando estos son de igual tamaño. Si la información auxiliar existe, tal como el tamaño del conglomerado i-ésimo M_i, la selección se realiza en función de dicha información empleando probabilidades desiguales. De esta manera, si en los conglomerados que conforman la muestra todos sus elementos son encuestados, a este tipo de muestreo se le conoce con el nombre de muestreo por conglomerados de una sola etapa.

Por otra parte, el muestreo bietápico consiste en seleccionar una muestra de conglomerados a partir de un marco muestral de estos, y posteriormente en cada conglomerado seleccionado, conformar un nuevo marco de elementos o conglomerados de elementos para una segunda y final selección. A los conglomerados del primer marco muestral se les llama unidades primarias de muestreo (UPM). Los elementos o conglomerados de elementos que conforman el segundo marco muestral en cada UPM seleccionada, se les nombra unidades secundarias de muestreo (USM).

[29] Cochran, William. Op cit., p. 125.
[30] Särndal, Carl-Erick, et. al. Op cit., p. 124.
[31] Kish, Leslie. Op cit., p. 184.
[32] Lehtonen, Risto and Pahkinen, Erkki J. Op cit., p. 89.

"*El muestreo multietápico consiste de tres o más etapas de muestreo. Existe una jerarquía de las unidades de muestreo: unidades primarias de muestreo, unidades secundarias de muestreo dentro de las UPM's, unidades terciarias de muestreo dentro de las USM, y así sucesivamente*".[33]

Es importante señalar que una ventaja del muestreo multietápico o polietápico, es que los marcos de segunda etapa se requieren solamente para las UPM seleccionadas, y así sucesivamente.

1.6 La encuesta por muestreo probabilística y compleja

La encuesta por muestreo del tipo probabilístico debe de contar con un marco muestral que se ajuste lo mejor posible a la población objeto de estudio, de donde se pueda seleccionar una muestra probabilística, y las estimaciones obtenidas a través de la encuesta se puedan generalizar a la población, proporcionando además una medida de su precisión.[34]

Es Wolter quien propone el nombre de encuesta por muestreo compleja,[35] para la encuesta probabilística, buscando con ello distinguirla de la encuesta que entre otras cosas, se fundamenta en la teoría básica del muestreo. Cabe señalar que, aunque no se ha definido rigurosamente dicho término, las características principales de dichas encuestas son las siguientes:

i. El grado de complejidad del diseño de muestreo.
ii. El grado de complejidad del estimador.
iii. Características o variables de interés múltiples.
iv. Usos descriptivos y analíticos de los datos de la encuesta.
v. La escala o tamaño de la encuesta.

Un ejemplo de este tipo de encuesta se presenta cuando el muestreo es polietápico, por conglomerados, estratificado y las probabilidades de selección de las diferentes unidades de muestreo desiguales, siendo además de magnitud nacional con diferentes dominios de estudio y numerosas variables de interés por estimar, lo que requiere un tamaño de muestra lo suficientemente grande que permita generar estimaciones confiables. Gran cantidad de estas encuestas son de temas socioeconómicos, de gran interés para los gobiernos por lo que adquieren el carácter de continuas, es decir, son levantadas mes a mes, trimestre a trimestre o de acuerdo a alguna otra periodicidad, permitiendo un análisis de la información que va más allá del descriptivo.

1.7 Utilidad de la varianza

Aunque en subcapítulos previos se ha mencionado que la magnitud del error de muestreo de una estimación, se conoce a través del error estándar ($\sqrt{V(\hat{\theta})}$). Esta no es la única aplicación que tiene la varianza. Diferentes autores coinciden en que la varianza se requiere para:

i. Calcular el tamaño de muestra necesario para una encuesta que está siendo planeada.
ii. Conocer la precisión de la estimación.[36]
iii. Construir intervalos de confianza para los parámetros poblacionales, esto es,

[33] Särndal, Carl-Erick, et. al. Op cit., p. 125.
[34] Kish, Leslie. Op cit., p. 41.
[35] Wolter, Kirk M. Op cit., p. 2.
[36] Cochran, William. Op cit., p. 50.

$$\left(\hat{\theta} - t_{\alpha/2}\sqrt{v(\hat{\theta})}, \hat{\theta} + t_{\alpha/2}\sqrt{v(\hat{\theta})}\right)$$

donde

$$\hat{\theta} = Estimador\ del\ parámetro\ poblacional\ \theta$$
$$v(\hat{\theta}) = Estimador\ del\ error\ estándar\ de\ \hat{\theta}$$
$$\alpha = Nivel\ de\ significancia$$
$$t_{\alpha/2} = \begin{cases} Valor\ de\ t\ en\ la\ tabla\ de\ la\ distribución\ normal \\ para\ una\ confianza\ 1-\alpha \end{cases}$$

iv. Comparar la precisión obtenida por el diseño de muestreo empleado con otros diseños. Un tipo de comparación se realiza contrastando directamente las varianzas $v(\hat{\theta})$ o estimaciones de coeficientes de variación $cv_{\hat{\theta}} = \sqrt{v(\hat{\theta})}\,\hat{\theta}$.[37] Otra muy ampliamente utilizada, es el denominado efecto de diseño (deff) que compara la varianza del diseño de muestreo empleado con la varianza del muestreo aleatorio simple con o sin reemplazo mediante la relación[38]

$$deff = \frac{v(\hat{\theta})}{v(\hat{\theta})_{MAS}}.$$

Sin duda, es fundamental el cálculo de la varianza, al menos para las estimaciones más importantes, que genera una encuesta por muestreo probabilística.

1.8 Algunos métodos de muestreo probabilístico

El propósito principal de esta sección, es presentar algunos de los métodos de muestreo más empleados en la selección de una muestra a partir de un marco muestral.

Muestreo aleatorio simple sin reemplazo

Método que permite seleccionar n unidades de muestreo a partir de un marco muestral con N unidades, de tal modo que cada una de las $C_{N,n}$ muestras distintas tengan la misma probabilidad de ser seleccionadas. Dado que en la práctica la muestra de tamaño n se conforma seleccionando unidad por unidad, para cada selección el método de muestreo otorga la misma probabilidad de selección a todas las unidades de muestreo que no han sido seleccionadas. Esta selección se hace mediante la generación de números aleatorios.[39]

Muestreo sistemático

Esquema de muestreo probabilístico que permite seleccionar una muestra de tamaño n de un marco muestral, donde las unidades de muestreo pueden ser ordenadas bajo cierto criterio con base en información auxiliar y numeradas de *1* a *N*. El proceso de selección consiste en obtener el cociente $k = N/n$, seleccionar un número aleatorio i del intervalo $[1,k]$ e incluir en la muestra las unidades de muestreo con los números seriados $i, i+k, i+2k, \ldots\ldots, i+(n-1)k$.[40]

---

[37] Recuerdese que $CV = \sqrt{V(\hat{\theta})}\ E(\hat{\theta}) = \sqrt{V(\hat{\theta})}\ \theta$.

[38] Lehtonen, Risto and Pahkinen, Erkki J. Op cit., p. 14.

[39] Cochran, William. Op cit., p. 41.

[40] Raj, Des. Op cit.,, p. 41.

Muestreo con probabilidad proporcional al tamaño (ppt)

Método de muestreo que haciendo uso de información auxiliar de las unidades de muestreo del marco muestral, les asigna probabilidades desiguales en el proceso de selección de la muestra. Las unidades de muestreo son listadas de acuerdo a algún criterio, se acumula la medida de tamaño y se asigna para cada unidad de muestreo un rango de selección. Como se muestra en la siguiente tabla:

Unidad de Muestreo	Medida de Tamaño	Acumulado de la Medida de Tamaño	Rango de Selección
1	X_1	$A_1 = X_1$	$[1, A_1]$
2	X_2	$A_2 = A_1 + X_2$	$[A_1+1, A_2]$
3	X_3	$A_3 = A_2 + X_3$	$[A_2+1, A_3]$
:	:	:	:
:	:	:	:
i	X_i	$A_i = A_{i-1} + X_i$	$[A_{i-1}+1, A_i]$
:	:	:	:
:	:	:	:
N	X_N	$A_N = A_{N-1} + X_N$	$[A_{N-1}+1, A_N]$
Total	$X = \sum_{i=1}^{N} X_i$		

En el caso de llevar a cabo este proceso de muestreo con reemplazo, la probabilidad de seleccionar a la unidad de muestreo *i*-ésima será X_i/X. La selección de la muestra se efectúa escogiendo n números aleatorios no necesariamente distintos del intervalo $[1,X]$, las unidades de muestreo seleccionadas serán aquellas cuyo intervalo $[A_{i-1}+1,A_i]$ o $[1, A_1]$ contenga a uno o varios de los n números aleatorios. Sí el muestreo es sin reemplazo, se seleccionarán n unidades de muestreo distintas, para lo cual será necesario calcular n veces el valor de X y las probabilidades desiguales de selección.[41]

Muestreo sistemático con probabilidad desigual

Este método de muestreo probabilístico también se conoce con el nombre de muestreo sistemático con probabilidad proporcional al tamaño. Es una generalización del muestreo sistemático. Al igual que en el muestreo ppt se hace una acumulación de la medida de tamaño, generándose un rango de selección para cada unidad de muestreo $[1,A_1]$, $[A_1+1, A_2]$,....., $[A_{i-1}+1\ A_i]$,....., $[A_{N-1}+1\ A_N]$. Con el fin de seleccionar la muestra de tamaño n, se toma un número aleatorio entre 1 y $k=X/n$. Las unidades de muestreo seleccionadas serán aquellas en cuyo campo está el número aleatorio i y los demás números $i+k$, $i+2k$,..... obtenidos al sumar sucesivamente k a i.[42]

1.9 Algunos diseños de muestreo

A lo largo del presente trabajo se han considerado los términos: estimador del parámetro poblacional, varianza del estimador y el estimador de la varianza. Sin haber especificado las

[41] Raj, Des. Op cit.,, p. 61.
[42] Idem, p. 65.

fórmulas matemáticas correspondientes, las cuales como ya se mencionó son determinadas por el tipo de muestreo probabilístico empleado. Con el fin de completar la exposición del tema, a continuación se presentan algunos diseños de muestreo, considerando al estimador insesgado del total poblacional Y.

1. Muestreo aleatorio simple sin reemplazo.

Estimador [43]:
$$\hat{Y} = \frac{N}{n}\sum_{i=1}^{n} y_i$$

Varianza del estimador[44]:
$$V(\hat{Y}) = N^2\left(1 - \frac{n}{N}\right)\frac{S^2}{n}$$

donde
$$S^2 = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}{N-1}$$
$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

Estimador de la varianza[45]:
$$v(\hat{Y}) = N^2\left(1 - \frac{n}{N}\right)\frac{s^2}{n}$$

donde
$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$
$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

2. Muestreo ppt con reemplazo.

Estimador[46]:
$$\hat{Y} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}$$

donde
$$p_i = \begin{cases} \textit{Probabilidad de seleccionar a la} \\ \textit{unidad de muestreo } i - \textit{ésima} \end{cases}$$

Varianza del estimador[47]:
$$V(\hat{Y}) = \frac{1}{n}\sum_{i=1}^{N} p_i\left(\frac{Y_i}{p_i} - Y\right)^2$$

Estimador de la varianza[48]:
$$v(\hat{Y}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \hat{Y}\right)^2$$

[43] Cochran, William. Op cit., p.48.
[44] Idem, p. 48.
[45] Idem, p. 51.
[46] Idem, p. 313.
[47] Idem, p. 314.
[48] Idem, p. 315.

3. Muestreo con probabilidad desigual sin reemplazo para un tamaño de muestra *n* prefijado (πps).

Estimador[49]: $\hat{Y} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$ (Estimador Horvitz-Thompson)

donde $\pi_i = np_i = \begin{cases} Probabilidad\ de\ que\ la\ unidad\ de\ muestreo \\ i-ésima\ sea\ seleccionada \end{cases}$

Varianza del estimador[50]:

$$V(\hat{Y}) = \sum_{i=1}^{N} \sum_{j>i}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2$$

donde

$$\pi_{ij} = \begin{cases} Probabilidad\ conjunta\ de\ que\ las\ unidades\ de\ muestreo \\ i\ y\ j-ésima\ sean\ seleccionadas \end{cases}$$

Estimador de la varianza[51]:

$$v(\hat{Y}) = \sum_{i=1}^{n} \sum_{j>i}^{n} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$ (Estimador Yates-Grundy)

4. Muestreo bietápico, πps en la primera etapa y aleatorio simple sin reemplazo en la segunda etapa de muestreo.

Estimador[52]: $\hat{Y} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{M_i}{\pi_i m_i} y_{ij}$

donde

$\pi_i = np_i = Probabilidad\ de\ que\ la\ UPM\ i-ésima\ sea\ seleccionada$
$M_i = Número\ total\ de\ USM\ en\ la\ UPM\ i-ésima\ seleccionada$
$m_i = Número\ de\ USM\ seleccionadas\ en\ la\ UPM\ i-ésima$

Varianza del estimador[53]:

$$V(\hat{Y}) = \sum_{i=1}^{N} \sum_{j>i}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_{i\bullet}}{\pi_i} - \frac{Y_{j\bullet}}{\pi_j} \right)^2 + \sum_{i=1}^{N} \left(\frac{1}{\pi_i} \right) M_i^2 \left(1 - \frac{m_i}{M_i} \right) \frac{S_i^2}{m_i}$$

donde

$$\pi_{ij} = \begin{cases} Probabilidad\ conjunta\ de\ que\ las\ UPM's \\ i\ y\ j-ésima\ sean\ seleccionadas \end{cases}$$

$$Y_{i\bullet} = \sum_{j=1}^{M_i} Y_{ij}$$

[49] Cochran, William. Op cit., p. 322.
[50] Idem, p. 322.
[51] Idem, p. 323.
[52] Idem, p. 377.
[53] Idem, p. 377.

$$S_i^2 = \frac{\sum_{j=1}^{M_i} (Y_{ij} - \bar{Y}_{i\bullet})^2}{M_i - 1}$$

$$\bar{Y}_{i\bullet} = \frac{Y_{i\bullet}}{M_i}$$

Estimador de la varianza[54]:

$$v(\hat{Y}) = \sum_{i=1}^{n} \sum_{j>i}^{n} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{y_{i\bullet}}{\pi_i} - \frac{y_{j\bullet}}{\pi_j} \right)^2 + \sum_{i=1}^{n} \left(\frac{1}{\pi_i} \right) M_i^2 \left(1 - \frac{m_i}{M_i} \right) \frac{s_i^2}{m_i}$$

donde

$$y_{i\bullet} = \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij} = M_i \bar{y}_{i\bullet}$$

$$s_i^2 = \frac{\sum_{j=1}^{m_i} (y_{ij} - \bar{y}_{i\bullet})^2}{m_i - 1}$$

1.10 El método de los conglomerados últimos

En los diseños de muestreo que involucran el muestreo estratificado, por conglomerados y polietápico, el cálculo de la estimación de la varianza puede representar entre otras cosas, un alto costo y consumo excesivo de tiempo, debido a la complejidad del estimador de la varianza. Por ejemplo, si el muestreo es polietápico, para obtener la estimación global de la varianza, es necesario calcular la varianza para cada una de las etapas de selección, llamadas componentes de la varianza (véase el diseño 4 de la sección anterior).

En virtud de lo anterior Hansen, Hurwitz y Madow (1953) proponen el concepto y método de los conglomerados últimos con el fin de simplificar el cálculo de la varianza en encuestas polietápicas. Dicen "*un conglomerado último consiste de todas las unidades listadas en la muestra de una particular unidad primaria, i.e., el conglomerado último consiste de la muestra entera proveniente de la UPM ya sea obtenida por 1, 2 o más etapas de muestreo*",[55] agregando que, "*para estimar la precisión de los resultados los componentes de la varianza no se necesitan. Para obtener un diseño óptimo, sin embargo, las estimaciones de los componentes de la varianza son necesarios*".[56] Lo que hace considerar al muestreo polietápico como caso especial del muestreo por conglomerados de una sola etapa.[57] En otras palabras, para estimar la varianza de la estimación del parámetro de interés generada a través de una encuesta polietápica, se debe considerar para el calculo, simplemente al total ponderado de la variable bajo estudio proveniente de cada UPM en muestra. La fórmula correspondiente y demás detalles del método son presentados en el Capítulo 2.

[54] Cochran, William. Op cit., p. 378.
[55] Hansen, Morris H., et. al. Sample Survey Methods and Theory Vol. 1. John Wiley & Sons. New York, 1953. p. 257.
[56] Idem, p. 257.
[57] Azorín, Francisco. Op cit., p. 215.

2 HERRAMIENTA MATEMÁTICA

2.1 Estimación de la varianza por el método de los conglomerados últimos

En el capítulo anterior se mencionó que el método de los conglomerados últimos para estimar la varianza en encuestas polietápicas, consiste básicamente, en considerar al total de la variable de interés proveniente de las unidades de muestreo con que cuenta cada UPM en muestra, de tal manera, que al muestreo polietápico se le considere como caso especial del muestreo por conglomerados de una sola etapa.[58]

A continuación se presentara de forma más detallada en que consiste el método de los conglomerados últimos para obtener estimaciones de la varianza de los estimadores del total y la razón, en un diseño de muestreo estratificado, por conglomerados y trietápico, lo cual no representa perdida de generalidad, dado que se puede extender a más etapas.

2.1.1 Estimación de la varianza para el estimador de un total

Sea el estimador del total, o de forma general, el estimador lineal[59]:

$$\hat{\theta} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \sum_{k=1}^{r_{hij}} u_{hijk} \qquad (1)$$

donde

$$u_{hijk} = \begin{cases} Valor\ de\ la\ variable\ U\ ponderada\ en\ la\ UTM\ k-ésima, \\ de\ la\ USM\ j-ésima,\ en\ la\ UPM\ i-ésima\ en\ muestra \\ del\ estrato\ h-ésimo \end{cases}$$

Considerando los supuestos[60]:

i. Las muestras seleccionadas en los diferentes estratos son independientes.

ii. La muestra n_h de UPM en cada estrato h-ésimo, es seleccionada con reemplazo. De esta manera, si

$$p_{hi} = \begin{cases} Probabilidad\ fija\ de\ seleccionar\ a\ la\ UPM \\ i-ésima\ en\ el\ estrato\ h-ésimo \end{cases}$$

$$\Rightarrow \sum_{i=1}^{N_h} p_{hi} = 1$$

iii. Siendo $n_h \geq 2$ se reescribe la ecuación (1) como:

$$\hat{\theta} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} w_{hi}$$

[58] Azorín, Francisco. Op cit., p. 215.
[59] Skinner, C. J. and Smith, T.M.F. Analysis of Complex Surveys. John Wiley & Sons. England, 1989. p. 47.
[60] Idem, p. 47.

donde

$$w_{hi} = \sum_{j=1}^{m_{hi}} \sum_{k=1}^{r_{hij}} u_{hijk}$$

Sí los supuestos *i* y *ii* se cumplen w_{h1}, w_{h2}, ..., w_{hn_h} son independientes e idénticamente distribuidas dentro de los estratos, por lo que la varianza de $\hat{\theta}$ se define como[61]:

$$V(\hat{\theta}) = \sum_{h=1}^{L} n_h V(w_{hi}) \qquad (2)$$

En consecuencia, un estimador insesgado de $V(\hat{\theta})$ se obtiene al considerar los tres supuestos, definiéndolo como[62]:

$$v(\hat{\theta}) = \sum_{h=1}^{L} \left(\frac{n_h}{n_h - 1} \right) \sum_{i=1}^{n_h} (w_{hi} - \bar{w}_h)^2 \qquad (3)$$

donde

$$\bar{w}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} w_{hi} .$$

"*Si los tres supuestos se cumplen, no importa como el submuestreo ocurre dentro de la UPM,* $v(\hat{\theta})$ *permanece insesgado para* $V(\hat{\theta})$".[63] El estimador es calculado con las cantidades w_{hi}, formadas de los conglomerados últimos.

En cuanto a los supuestos, *i* siempre se respeta, pero cuando el *iii* no se cumple, $v(\hat{\theta})$ es imposible de calcular, teniéndose que colapsar los estratos produciéndose un estimador de varianza conservador. En relación al supuesto *ii*, este es regularmente incumplido, por lo que $v(\hat{\theta})$ es reformulado de manera compleja. Aunque existe una aproximación ampliamente usada, la cual asume que los n_h conglomerados últimos en el estrato h, forman una muestra aleatoria simple sin reemplazo de los estratos $h=1,2,.....,L$ y es obtenida al insertar una corrección por población finita en la ecuación (3), esto es[64]:

$$v_{sr}(\hat{\theta}) = \sum_{h=1}^{L} \left(1 - \frac{n_h}{N_h} \right) \left(\frac{n_h}{n_h - 1} \right) \sum_{i=1}^{n_h} (w_{hi} - \bar{w}_h)^2 \qquad (4)$$

"*Frecuentemente las fracciones de muestreo* n_h/N_h *son pequeñas y la diferencia entre* $v(\hat{\theta})$ *y* $v_{sr}(\hat{\theta})$ *o fórmulas alternativas sin reemplazo es insignificante (cf. Durbin, 1953). En cualquier caso* $v(\hat{\theta})$ *es generalmente un estimador de varianza conservador*".[65]

Para este trabajo, la fórmula considerada para estimar la varianza del estimador de un total, es la ecuación (3).

[61] Skinner, C. J. and Smith, T.M.F. Analysis of Complex Surveys. John Wiley & Sons. England, 1989. p. 47.
[62] Idem, p. 47.
[63] Idem, p. 47.
[64] Idem, p. 48.
[65] Idem, p. 48.

2.1.2 Estimación de la varianza de un estimador de razón simple

Para estadísticos que no tienen la forma de la ecuación (1), es decir, que en general no son funciones lineales de las observaciones, como es el caso del estimador de razón simple $\hat{R}$.[66] Se recurre a métodos que permiten aproximar la varianza de dichos estimadores.

Un método útil para la estimación de la varianza de un estimador no lineal $\hat{\theta}$, consiste en aproximarse a este con un estadístico lineal. Asumiendo que $\hat{\theta}$ es consistente para θ y adecuadamente diferenciable, esto se lleva a cabo empleando la aproximación con la serie de Taylor de primer orden (método de linealización o delta) para posteriormente aplicarle a la aproximación lineal, la fórmula de la varianza apropiada al diseño de muestreo específico.[67] En este caso, la fórmula de los conglomerados últimos para estimar la varianza del estimador del total, conduciendo a un estimador de varianza del estimador no lineal, sesgado pero consistente.[68]

De esta manera, para estimar la varianza del estimador de razón simple[69]:

$$\hat{\theta} = \frac{\hat{\theta}_1}{\hat{\theta}_2} = \frac{\sum_{h}^{L}\sum_{i}^{n_h} w_{1hi}}{\sum_{h}^{L}\sum_{i}^{n_h} w_{2hi}} \tag{5}$$

donde

$$\hat{\theta}_1 = Estimador\ lineal\ de\ \theta_1$$

$$\hat{\theta}_2 = Estimador\ lineal\ de\ \theta_2$$

$$w_{1hi} = \sum_{j=1}^{m_{hi}}\sum_{k=1}^{r_{hij}} u_{1hijk}$$

$$w_{2hi} = \sum_{j=1}^{m_{hi}}\sum_{k=1}^{r_{hij}} u_{2hijk}\ .$$

En un diseño de muestreo, estratificado, por conglomerados y trietápico, se cuenta con la fórmula[70]:

$$v(\hat{\theta}) = \frac{1}{\hat{\theta}_2^2}\sum_{h}^{L}\frac{n_h}{n_h-1}\sum_{i}^{n_h}\left[(w_{1hi}-\bar{w}_{1h})-\hat{\theta}(w_{2hi}-\bar{w}_{2h})\right]^2 \tag{6}$$

donde

$$\bar{w}_{1h} = \frac{1}{n_h}\sum_{i=1}^{n_h} w_{1hi}$$

$$\bar{w}_{2h} = \frac{1}{n_h}\sum_{i=1}^{n_h} w_{2hi}$$

[66] Skinner, C. J. and Smith, T.M.F. Analysis of Complex Surveys. John Wiley & Sons. England, 1989. p. 50.
[67] Wolter, Kirk M. Op cit., p. 221.
[68] Idem, p. 221.
[69] Lehtonen, Risto and Pahkinen, Erkki J. Op cit., p. 142.
[70] Skinner, C. J. and Smith, T.M.F. Analysis of Complex Surveys. John Wiley & Sons. England, 1989. p. 51.

2.2 Estimadores de varianza para el muestreo aleatorio simple

En el subcapítulo 1.7 se especificó que el efecto de diseño es una medida empleada para comparar la varianza del diseño de muestreo empleado, con la varianza del muestreo aleatorio simple, definiéndose al estimador del efecto de diseño como[71]:

$$deff = \frac{v(\hat{\theta})}{v(\hat{\theta})_{MAS}} \tag{7}$$

donde

$$v(\hat{\theta}) = Varianza\ bajo\ el\ diseño\ de\ muestreo\ empleado$$
$$v(\hat{\theta})_{MAS} = Varianza\ bajo\ el\ muestreo\ aleatorio\ simple$$

Para completar el cálculo del deff, a continuación se definirán los estimadores de la varianza bajo el muestreo aleatorio simple para los estimadores del total y la razón simple.

Sea el estimador de un total en un diseño de muestreo estratificado, por conglomerados y trietápico:

$$\hat{X} = \sum_{h}^{L}\sum_{i}^{n_h}\sum_{j}^{m_{hi}} w_{hij}\, x_{hij} \tag{8}$$

donde

$L =$ *Número de estratos*

$n_h =$ *Número de UPM en muestra*

$$m_{hi} = \begin{cases} Número\ de\ unidades\ elementales\ de\ observación \\ en\ la\ UPM\ i-ésima\ del\ estrato\ h-ésimo \end{cases}$$

$$w_{hij} = \begin{cases} Ponderador\ de\ la\ unidad\ elemental\ de\ observación \\ j-ésima,\ en\ la\ UPM\ i-ésima\ del\ estrato\ h-ésimo \end{cases}$$

$$x_{hij} = \begin{cases} Valor\ de\ la\ variable\ de\ análisis\ x\ para\ la\ unidad\ elemental\ de \\ observación\ j-ésima,\ en\ la\ UPM\ i-ésima\ del\ estrato\ h-ésimo \end{cases}$$

Es importante señalar, que w_{hij} es el ponderador definido como el inverso de la probabilidad de selección de la unidad terciaria de muestreo para el estrato *h*-ésimo. De la misma manera, x_{hij} es el valor obtenido para la variable *x* en la unidad elemental de observación *j*-ésima, contenida en la unidad terciaria de muestreo, seleccionada después de tres etapas de selección y que pertenece a la UPM *i*-ésima en el estrato *h*-ésimo. No se colocaron más subíndices, ni sumatorias con el fin de simplificar las expresiones.

El estimador de la varianza de $\hat{X}$, bajo el muestreo aleatorio simple se define como[72]:

$$v(\hat{X})_{MAS} = \left(1 - \frac{m_{\bullet\bullet}}{\hat{M}}\right)\left(\frac{\hat{M}}{m_{\bullet\bullet} - 1}\right)\left[\sum_{h}^{L}\sum_{i}^{n_h}\sum_{j}^{m_{hi}} w_{hij}\, x_{hij}^2 - \frac{1}{\hat{M}}\hat{X}^2\right] \tag{9}$$

[71] Lehtonen, Risto and Pahkinen, Erkki J. Op cit., p. 14.
[72] Statistical Laboratory, Iowa State University. PC CARP. 1986. p. 97.

donde

$$m_{\bullet\bullet} = \sum_{h}^{L} \sum_{i}^{n_h} m_{hi} = \textit{Número total de unidades elementales de observación}$$

$$\hat{M} = \sum_{h}^{L} \sum_{i}^{n_h} \sum_{j}^{m_{hi}} w_{hij} = \textit{Estimación del total de la población bajo estudio}$$

Sí lo que interesa es comparar la varianza bajo el diseño empleado con el muestreo aleatorio simple con reemplazo, la ecuación (9) se convierte en [73]:

$$v(\hat{X})_{MASCR} = \left(\frac{\hat{M}}{m_{\bullet\bullet} - 1}\right)\left[\sum_{h}^{L} \sum_{i}^{n_h} \sum_{j}^{m_{hi}} w_{hij} x_{hij}^2 - \frac{1}{\hat{M}} \hat{X}^2\right] \tag{10}$$

Siendo esta última fórmula, la empleada en este trabajo para estimar el efecto de diseño.

Considérese ahora, para el mismo diseño de muestreo al estimador de razón simple

$$\hat{R} = \frac{\hat{X}}{\hat{Y}} \tag{11}$$

donde, $\hat{X}$ y $\hat{Y}$ se definen de manera análoga a la ecuación (8).

Finalmente, la fórmula del estimador de la varianza de $\hat{R}$ bajo el muestreo aleatorio simple, se define como [74]:

$$v(\hat{R})_{MAS} = \left(1 - \frac{m_{\bullet\bullet}}{\hat{M}}\right)\left(\frac{1}{\hat{M}(m_{\bullet\bullet} - 1)}\right)\left\{\sum_{h}^{L} \sum_{i}^{n_h} \sum_{j}^{m_{hi}} w_{hij}\left[\left(x_{hij} - \bar{X}_{\bullet\bullet\bullet}\right) - \hat{R}\left(y_{hij} - \bar{Y}_{\bullet\bullet\bullet}\right)\right]^2\right\} \tag{12}$$

donde

$$\bar{X}_{\bullet\bullet\bullet} = \frac{1}{\hat{M}} \hat{X} \quad \text{y} \quad \bar{Y}_{\bullet\bullet\bullet} = \frac{1}{\hat{M}} \hat{Y}$$

y las demás variables, definidas igual que antes.

Sí se requiere hacer la comparación de la varianza bajo el diseño de muestreo empleado, con el muestreo aleatorio simple con reemplazo, la ecuación (12) se convierte en [75]:

$$v(\hat{R})_{MASCR} = \left(\frac{1}{\hat{M}(m_{\bullet\bullet} - 1)}\right)\left\{\sum_{h}^{L} \sum_{i}^{n_h} \sum_{j}^{m_{hi}} w_{hij}\left[\left(x_{hij} - \bar{X}_{\bullet\bullet\bullet}\right) - \hat{R}\left(y_{hij} - \bar{Y}_{\bullet\bullet\bullet}\right)\right]^2\right\} \tag{13}$$

que es la fórmula empleada en este trabajo para estimar el efecto de diseño.

[73] Statistical Laboratory, Iowa State University. PC CARP. 1986. p. 98.
[74] Idem, p. 98.
[75] Idem, p. 98.

3 MODELO

3.1 Utilidad del programa VARCU

El objetivo del presente trabajo es presentar y aplicar el método de los conglomerados últimos, para estimar los errores de muestreo de las estimaciones de totales y razones simples, generadas por una encuesta del tipo complejo. Para lograr esto, en el primer capítulo se presentó entre otras cosas, la definición de una encuesta compleja y el método de los conglomerados últimos. Y en el capítulo anterior, las fórmulas de los estimadores de la varianza para $\hat{Y}$ y $\hat{R}$ con el método de los conglomerados últimos y bajo el muestreo aleatorio simple.

Con el fin de llevar a cabo la aplicación del método de los conglomerados para estimar la varianza en una encuesta compleja. Se elaboró un programa de cómputo en Fox Pro para MS-DOS Ver 2.6a, llamado VARCU (Varianza por Conglomerados Últimos), el cual, data de 1999 aproximadamente.

En los siguientes subcapítulos se presenta la necesidad de hardware, el procedimiento de instalación y desinstalación y la forma de emplear el programa VARCU.

3.2 Instalación del programa VARCU

Requerimientos de hardware para instalar VARCU

- Cualquier PC compatible con IBM, con procesador Pentium o superior.
- Una unidad de disquete de 3 1/2.
- Por lo menos 16 MB de memoria.
- 3 MB en disco duro, para la instalación.

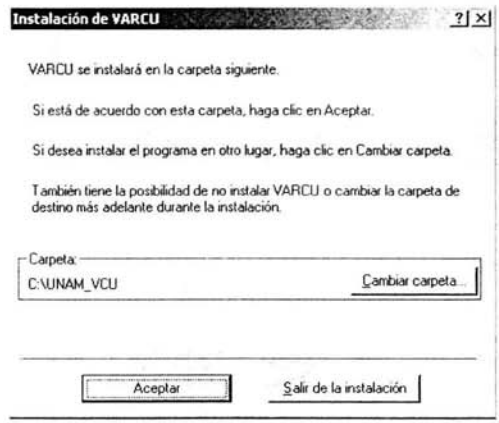
Instalación del programa

Para instalar el programa, deberá de ejecutar los programas de la instalación contenidos en 2 discos etiquetados. No podrá ejecutar el programa desde los disquetes de Instalar ya que los archivos están comprimidos.

Para instalar el programa:

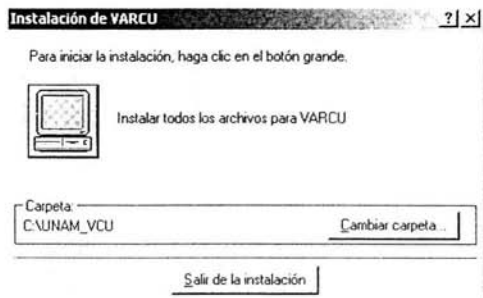
1. Inicie Microsoft Windows en modo estándar o mejorado.
2. Inserte el disquete 1 de instalación en la unidad A (unidad de inicio).
3. Elija **ejecutar** del menú inicio.
4. Escriba **a:\setup** y luego presione ENTRAR.
5. Siga las instrucciones que se presentan **en pantalla.**

La rutina de instalación coloca su aplicación en el directorio predeterminado **C:\UNAM_VCU** o si prefiere cambie el nombre o la ubicación de la carpeta.



El programa de instalación creará un grupo de programas para la aplicación con el nombre de **VARIANZA** cuando el usuario lo instale y hará que esté disponible a través del menú **Inicio** del usuario.

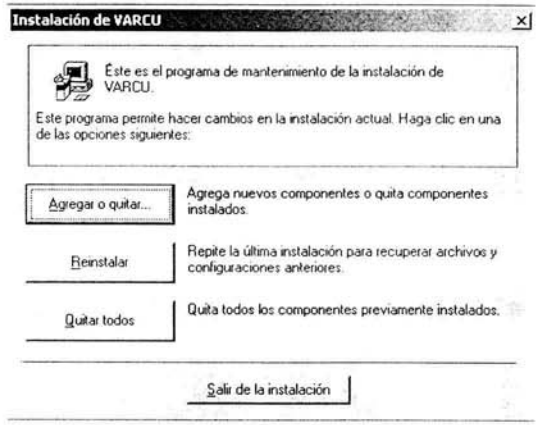
Se inicia la instalación mostrando la siguiente pantalla:



Al terminar, quedarán en la carpeta elegida los archivos necesarios para el buen funcionamiento del programa.

Desinstalación del programa

1. Inicie Microsoft Windows en modo estándar o mejorado.
2. Inserte el disquete 1 de instalación en la unidad A (unidad de inicio).
3. Elija **ejecutar** del menú inicio.
4. Escriba **a:\setup** y luego presione ENTRAR.
5. Elija la opción de **Quitar todos**, lo cual eliminará los archivos que previamente se copiaron en la carpeta creada.



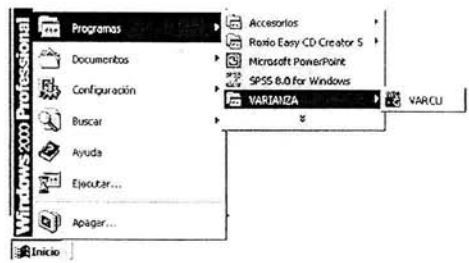
3.3 Empleo del programa VARCU

El programa requiere de una base de datos previamente preparada con las siguientes características:

- Ser compatible con Fox Pro.
- El nombre debe empezar con la letra "Z" y no deberá exceder de 8 caracteres.
- Contar con un número mínimo de 4 campos o un máximo de 26 (el programa trabajará con los campos especificados, aunque la base de datos contenga más).
- Todos los campos deberán ser de tipo Numérico.
- Los nombres de los campos deberán iniciar con la letra "Z"
- Los tres primeros campos deben corresponder al Estrato, UPM y Factor de Expansión, invariablemente en ese orden.
- Deberá estar ordenada por Estrato y UPM.
- Cada estrato deberá contener al menos 2 UPM.

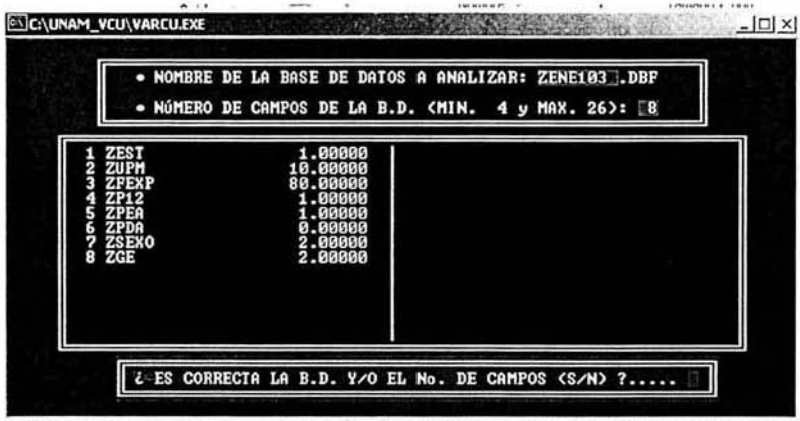
Ejecución del programa

1. Haga clic en **Inicio**, seleccione **Programas** y, a continuación, haga clic en **VARIANZA**.
2. Haga clic en **VARCU**

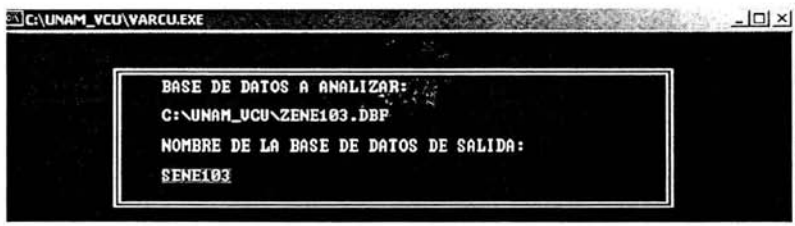


Cuando se presiona la tecla solicitada, se nos pedirá que introduzcamos el nombre de la base de datos que contiene la información para el cálculo de la varianza. Así como, el número de campos que se van a utilizar. En caso de que se presente algún tipo de error con la base de datos, el programa mostrará un mensaje de error (archivo inexistente, número de campos incorrecto ó campos de tipo distinto al numérico).

Si la base de datos no presenta problema alguno, se desplegarán en pantalla los valores de los campos correspondientes al primer registro. Así como la pregunta, "¿ ES CORRECTA LA B.D. Y/O EL No. DE CAMPOS (S/N) ?". En caso negativo teclee "N" y si a continuación desea continuar con la sesión, vuelva a introducir el nombre de la base de datos y el número de campos. Sí estos son correctos, teclee "S" a la pregunta "¿ ES CORRECTA LA B.D. Y/O EL No. DE CAMPOS (S/N) ?".



A continuación nos pedirá el nombre del archivo de salida con un máximo de 7 caracteres, automáticamente se le agrega el carácter "Z". Se recomienda ponerle un nombre parecido al nombre de la base de datos de entrada, para poder identificar de cual base de datos proviene. Si el archivo de salida existe, el programa preguntará si lo quiere reemplazar o no.



El formato del archivo de salida es una base de datos, que nos permite manipularla ya sea en un manejador de base de datos o en Microsoft Excel, conteniendo los siguientes campos:

Campo	Nombre de campo	Tipo	Ancho	Dec
1	VARIABLE	Carácter	43	No
2	ESTIMACION	Numérico	20	6
3	ERROR_EST	Numérico	20	10
4	COEF_VAR	Numérico	10	8
5	EFECTO_DIS	Numérico	15	9

Enseguida, el programa iniciará la verificación del ordenamiento de los estratos y del número de UPM en ellos (el programa soporta hasta 2000 estratos). En el caso de encontrar algún problema, no podrá llevar a cabo análisis alguno e indicará en que consiste el problema, para a continuación abandonar la sesión. De lo contrario, pasará a la pantalla del menú principal, donde se visualizan los análisis que VARCU permite realizar, a los cuales se accede de acuerdo al número de identificación.

Antes de continuar, es importante señalar que para el procesamiento de la información se deben distinguir las variables del tipo dependiente y categórico. Ejemplos de la primera son, la población: menor de 10 años, de mujeres casadas, asalariada, con cierta enfermedad, etc. De la segunda, el sexo, estado civil, entidad, municipio, grupo de edad, etc.

Para explicar la forma de emplear VARCU se consideraran las siguientes variables dependientes provenientes de la Encuesta Nacional de Empleo Urbano (ENEU):

- Población de 12 años y más (ZP12).
- Población económicamente activa (ZPEA).
- Población desocupada abierta (ZPDA).

Y las variables categóricas:

- Sexo (ZSEXO; 1 hombre y 2 mujer).
- Grupo de edad (ZGE: 1 de 12-39 años, 2 de 40 años y más y 3 no especificado).

Procesamiento de Totales

Suponga que se quieren obtener estimaciones de la varianza para las variables dependientes antes citadas. De la pantalla principal se elige el análisis "(1) Total". Posteriormente se muestra la lista de campos (variables), de la cual se seleccionan las variables 1, 2 y 3. En caso de que la captura de los números de las variables sea errónea, indique que la selección no es la correcta y nuevamente capture. El programa permite introducir por vez un máximo de 15 variables dependientes.

Mientras el software esta procesando la información, muestra un mensaje indicándonos la hora de inicio y el mensaje "PROCESANDO". Al término de este, mostrará la hora de finalización.

El resultado del proceso queda guardado en la base de datos de salida que anteriormente se definió. Ahí se almacena el nombre de la base de datos de entrada y salida, el día y la hora de proceso, el nombre del análisis que se realiza, las estimaciones puntuales de las variables

dependientes y sus correspondientes estimaciones del error estándar, coeficiente de variación y efecto de diseño.

Si desea revisar los resultados del análisis, puede elegir la opción "(8) Visualizar Resultados", la cual, estará disponible al término de cualquier análisis. Para este ejemplo mostrará lo siguiente:

Variable	Estimacion	Error_est	Coef_var	Efecto_dis
11/03/03				
17:04:08				
C:\UNAM_UCU\ZENE103.DBF				
C:\UNAM_UCU\ZSENE103.DBF				
**********				
TOTAL (ES)				
Observaciones:	8009.000000			
ZP12	557376.000000	13194.8701536902	0.02367319	0.000000000
ZPEA	295037.000000	9210.6290662665	0.03121856	6.814282127
ZPDA	14226.000000	1680.9846483621	0.11816285	2.273556617

Procesamiento de Totales por Estrato

Considere que se requieren las precisiones estadísticas de las variables dependientes del proceso anterior, pero esta vez por estrato. Seleccione de la pantalla principal la opción "(3) Totales por Estrato". La manera en que se capturan las variables dependientes, es igual que en el proceso anterior. La diferencia radica, en que las estimaciones se generan para cada uno de los estratos.

El resultado correspondiente al procesamiento, se presenta a continuación:

Variable	Estimacion	Error_est	Coef_var	Efecto_dis
**********				
TOTAL (ES) POR ESTRATO				
Estrato:	1.000000			
Observaciones:	1177.000000			
ZP12	110160.000000	5637.0813650141	0.05117176	0.000000000
ZPEA	60773.000000	4827.6894931328	0.07943806	9.131931798
ZPDA	3887.000000	1276.9950541882	0.32852973	4.642461075
Estrato:	2.000000			
Observaciones:	1403.000000			
ZP12	120645.000000	4466.6153213111	0.03702280	0.000000000
ZPEA	64681.000000	3811.3921278689	0.05892599	5.626389647
ZPDA	3702.000000	665.9056132173	0.17987726	1.436027378
Estrato:	3.000000			
Observaciones:	680.000000			
ZP12	64142.000000	3833.9656869733	0.05977309	0.000000000
ZPEA	36272.000000	2798.8167341059	0.07716191	5.261507897
ZPDA	1920.000000	466.8904225476	0.24317210	1.238952081
Estrato:	4.000000			
Observaciones:	1755.000000			
ZP12	165783.000000	9725.0176790144	0.05866113	0.000000000
ZPEA	81643.000000	5791.7589482199	0.07094006	8.565033843
ZPDA	2351.000000	473.3389000343	0.20133513	1.022784785
Estrato:	5.000000			
Observaciones:	1203.000000			
ZP12	96646.000000	3619.6682005887	0.03745285	0.000000000
ZPEA	51668.000000	2371.6123579689	0.04590099	2.909176603
ZPDA	2366.000000	556.3509260720	0.23514409	1.667891457

Procesamiento de Totales por Subpoblación

Para este caso, considere que se necesitan las precisiones estadísticas de:

- La población de 12 años y más (ZP12) por sexo (ZSEXO).
- La población de 12 años y más (ZP12) por sexo (ZSEXO) y grupo de edad (ZGE).
- La población económicamente activa (ZPEA) por sexo (ZSEXO).
- La población económicamente activa (ZPEA) por sexo (ZSEXO) y grupo de edad (ZGE).
- La población desocupada abierta (ZPDA) por sexo (ZSEXO).
- La población desocupada abierta (ZPDA) por sexo (ZSEXO) y grupo de edad (ZGE).

El procedimiento para la captura de las variables es el siguiente:

1. Seleccionar de la pantalla principal, la opción "(5) Totales por Subpoblación".
2. Indique el número de cruces o variables categóricas que van a intervenir en el proceso, máximo 3. Para este ejemplo es 1 (sólo la variable ZSEXO).
3. Elegir la (s) variable (s) categórica (s) que interviene (n) y su correspondiente número de categorías (máximo 99). En este caso es la variable categórica 4 (ZSEXO), con 2 categorías.
4. Dado que se necesita realizar otro análisis subpoblacional, cuando se muestre el mensaje "¿DESEA HACER OTRO ANÁLISIS SUBPOB. (s/n)?" teclee "S".
5. Indique el número de cruces o variables categóricas que van a intervenir en el proceso. En este caso **2** (las variables ZSEXO y ZGE).
6. Seleccione la primera variable categórica y a continuación el número de categorías que contiene, es decir, la número 4 (ZSEXO) con 2 categorías. A continuación seleccione la variable categórica 5 (ZGE) con 3 categorías.
7. Finalmente, capture las variables dependientes ZP12, ZPEA y ZPDA en forma análoga a los procesos anteriores. Como máximo se pueden introducir 10.

El resultado parcial de este proceso es:

Variable	Estimacion	Error_est	Coef_var	Efecto_dis
**********				
TOTAL (ES) POR SUBPOBLACIÓN (ES)				
VARIABLE DEPENDIENTE: ZP12				
VARIABLE (S) CATEGÓRICA (S):				
ZSEXO				
2	292704.000000	6523.5704248453	0.02228726	3.415190401
1	264672.000000	8144.1433499504	0.03077070	5.322737264
ZSEXO ZGE				
2 2	113736.000000	3734.5593052768	0.03283533	1.718426421
2 1	178889.000000	6004.0304584530	0.03356288	3.310030818
1 2	95550.000000	3884.5030925124	0.04065414	2.125899549
1 1	169054.000000	6024.0103115714	0.03563365	3.436646275
2 3	79.000000	79.0000000000	1.00000000	0.881294893
1 3	68.000000	68.0000000000	1.00000000	0.758567973
VARIABLE DEPENDIENTE: ZPEA				
VARIABLE (S) CATEGÓRICA (S):				
ZSEXO				
2	112775.000000	4369.0672694246	0.03874145	2.366873566
1	182262.000000	6158.8599755864	0.03379125	3.449229750
ZSEXO ZGE				
2 2	39651.000000	2205.5510631388	0.05562410	1.473200372
2 1	73124.000000	3700.4384436200	0.05060498	2.404118512
1 1	108023.000000	4324.6363130453	0.04003440	2.395389651
1 2	74171.000000	3750.2077529589	0.05056164	2.439641251
1 3	68.000000	68.0000000000	1.00000000	0.758567973

Procesamiento de Razones

Considere en esta ocasión, que interesa obtener las precisiones estadísticas de las razones simples ZPDA/ZPEA y ZPEA/ZP12. Para esto, las variables son introducidas de la siguiente manera:

1. Seleccione de la pantalla principal, la opción "(2) Razón".
2. Capture el número de la variable dependiente del numerador y a continuación el del denominador. Para el ejemplo 3 y 2.
3. Al mensaje "¿ Otra razón (s/n) ?" responda "S".
4. En forma análoga al paso 2, capture 2 y 1, para indicar la otra razón.

Esta opción permite procesar como máximo 12 razones.

En la pantalla siguiente, se pueden observar los resultados:

Variable	Estimacion	Error_est	Coef_var	Efecto_dis
**********				
R A Z Ó N (ES)				
O b s e r v a c i o n e s:	8009.000000			
ZPDA/ZPEA	0.048218	0.0053065992	0.11005505	2.019285434
ZPEA/ZP12	0.529332	0.0089039857	0.01682117	1.978367463

Procesamiento de Razones por Estrato

Esta opción, al igual que la anterior nos permite obtener precisiones estadísticas de las razones de interés, pero ahora en cada uno de los estratos.

Elija la opción "(3) Razones por Estrato", de la pantalla principal y capture en forma análoga las variables dependientes de las razones del ejemplo anterior.

El resultado correspondiente a este procesamiento, se presenta a continuación:

Variable	Estimacion	Error_est	Coef_var	Efecto_dis
**********				
R A Z Ó N (ES) P O R E S T R A T O				
Estrato:	1.000000			
O b s e r v a c i o n e s:	1177.000000			
ZPDA/ZPEA	0.063959	0.0179000010	0.27986539	3.472174913
ZPEA/ZP12	0.551679	0.0249953490	0.04530775	2.970641653
Estrato:	2.000000			
O b s e r v a c i o n e s:	1403.000000			
ZPDA/ZPEA	0.057235	0.0100125750	0.17493878	1.396508921
ZPEA/ZP12	0.536127	0.0199938980	0.03729324	2.253596408
Estrato:	3.000000			
O b s e r v a c i o n e s:	680.000000			
ZPDA/ZPEA	0.052933	0.0117299196	0.22159723	1.053849508
ZPEA/ZP12	0.565495	0.0184080424	0.03255207	0.936400224

Estrato:	4.000000			
Observaciones:	1755.000000			
ZPDA/ZPEA	0.028796	0.0058717624	0.20390825	1.064881798
ZPEA/ZP12	0.492469	0.0135022696	0.02741750	1.279386207
Estrato:	5.000000			
Observaciones:	1203.000000			
ZPDA/ZPEA	0.045792	0.0107798271	0.23540664	1.708951941
ZPEA/ZP12	0.534611	0.0224628109	0.04201713	2.437690952

Procesamiento de Razones por Subpoblación

Este procedimiento es análogo al de totales por subpoblación, en cuanto a la especificación de las variables categóricas y al de razones, para las variables dependientes.

Considere que se está interesado en obtener las siguientes precisiones estadísticas:

- De ZPDA/ZPEA por sexo (ZSEXO).
- De ZPDA/ZPEA por sexo (ZSEXO) y grupo de edad (ZGE).
- De ZPEA/ZP12 por sexo (ZSEXO).
- De ZPEA/ZP12 por sexo (ZSEXO) y grupo de edad (ZGE).

El procedimiento para la captura de las variables es el siguiente:

1. Seleccionar de la pantalla principal, la opción "(6) Razones por Subpoblación".
2. Indicar cuantas variables categóricas intervienen en el análisis subpoblacional, máximo 3. En este caso 1 (ZSEXO).
3. Capturar el número de la variable categórica y de sus categorías. Para este ejemplo 4 y 2 respectivamente.
4. Dado que se necesita realizar otro análisis subpoblacional, cuando se muestre el mensaje "¿DESEA HACER OTRO ANÁLISIS SUBPOB. (s/n)?" teclee "S".
5. Indique el número de cruces o variables categóricas que van a intervenir en el proceso. En este caso 2 (las variables ZSEXO y ZGE).
6. Seleccione la primera variable categórica y a continuación el número de categorías que contiene, es decir, la número 4 (ZSEXO) con 2 categorías. A continuación seleccione la variable categórica 5 (ZGE) con 3 categorías.
7. Finalmente, capture las variables dependientes de las razones de interés, en forma análoga a los últimos dos procedimientos.

Los resultados de este procesamiento se presentan a continuación:

Variable	Estimacion	Error_est	Coef_var	Efecto_dis
**********				
RAZÓN (ES) POR SUBPOBLACIÓN (ES)				
RAZÓN DEPENDIENTE: ZPDA/ZPEA				
VARIABLE (S) CATEGÓRICA (S):				
ZSEXO				
2	0.040319	0.0080331059	0.19923763	2.097845525
1	0.053105	0.0067892726	0.12784651	1.863541123
ZSEXO ZGE				
2 2	0.016771	0.0076588849	0.45666533	1.573242964
2 1	0.053088	0.0116204830	0.21889134	2.190966055
1 2	0.057138	0.0141876835	0.24830455	3.091112963
1 1	0.050369	0.0071596989	0.14214522	1.291281534
2 3	0.000000	0.0000000000	0.00000000	0.000000000
1 3	0.000000	0.0000000000	0.00000000	0.000000000
RAZÓN DEPENDIENTE: ZPEA/ZP12				
VARIABLE (S) CATEGÓRICA (S):				
ZSEXO				
2	0.385287	0.0121212375	0.03146029	2.025342586
1	0.688633	0.0122858288	0.01784088	2.078210219
ZSEXO ZGE				
2 2	0.348623	0.0179415372	0.05146399	1.798296600
2 1	0.408767	0.0150948082	0.03692762	1.881207913
1 2	0.776253	0.0185643436	0.02391532	2.114766153
1 1	0.638985	0.0179451604	0.02808384	2.632309342
2 3	0.000000	0.0000000000	0.00000000	0.000000000
1 3	1.000000	0.0000000000	0.00000000	0.000000000

4 RESULTADOS

4.1 La ENEU

La Encuesta Nacional de Empleo Urbano se levanta en las Áreas Metropolitanas más importantes del país. El diseño de muestreo vigente hasta el primer trimestre del 2003, es establecido como polietápico, estratificado y por conglomerados, donde la unidad última de selección es la vivienda y la unidad de observación, la persona residente habitual de la vivienda. Para el Área Metropolitana de Chihuahua, en la primera etapa 100 UPM´s son seleccionadas a través del muestreo sistemático con probabilidad desigual; en la segunda etapa 6 USM se seleccionan con muestreo sistemático y probabilidad desigual de cada UPM en muestra; para la tercera etapa se seleccionan k USM de las antes seleccionadas con muestreo sistemático, dando un total de 420 USM; finalmente, en la cuarta etapa son seleccionadas 6 viviendas con muestreo sistemático de cada USM en muestra, dando un total de 2,520 viviendas en muestra.

El método de los conglomerados últimos para el cálculo de la varianza se aplicó en el Área Metropolitana de Chihuahua cuya información corresponde al primer trimestre del 2003. Los registros de la base de datos analizada contienen información correspondiente únicamente a la población de 12 años y más.

4.1 Precisiones estadísticas de totales

En el Cuadro 1 se presentan las estimaciones del error estándar (*e.e.*), coeficiente de variación (*c.v.*) y efecto de diseño (*deff*) para las estimaciones de la población de 12 años y más (ZP12), población económicamente activa (ZPEA) y población desocupada abierta (ZPDA) a nivel general, por sexo, sexo y grupo de edad y estrato, así como los correspondientes intervalos de confianza al 90%.

Con base en dichos resultados y en particular del coeficiente de variación estimado, el usuario de la información puede decidir que estimación es más precisa y por consecuencia confiable. De esta manera, las estimaciones con mejor precisión son las de ZP12 y ZPEA en todas sus categorías, debido a que resultaron con coeficientes de variación bajos. Situación contraria para las estimaciones de ZPDA. Esto es visualizado una vez más, cuando el usuario observa los intervalos de confianza para cada una de las estimaciones. Es decir, se obtienen para la ZPDA intervalos de confianza con una amplitud más grande, que los obtenidos en ZP12 y ZPDA en cada una de sus categorías.

En cuanto a las estimaciones por estrato, los coeficientes de variación indican que las estimaciones para cada uno de los estratos son menos precisas que la general.

Con respecto a las estimaciones del *deff* sólo se mencionara una breve interpretación de su valor. Para el cálculo del *deff* cada registro de la base de datos contiene información referente a la vivienda en muestra. De esta manera, el valor resultante del *deff* estará en función de la vivienda. Siendo de utilidad por ejemplo, para determinar el tamaño de muestra en viviendas de una

encuesta por muestreo probabilística que tiene también a la vivienda como unidad de muestreo última de selección. Es decir, si se esta determinando el tamaño de muestra n para una encuesta por muestreo probabilístico del tipo complejo equivalente a la ENEU en el Área Metropolitana

Cuadro 1. Precisiones estadísticas de totales estimados.

Variable	Estimación	e.e.	c.v.	deff	Intervalo de Confianza al 90% L.I.I.C.	L.S.I.C.
Población de 12 años y más	**557,376**	**13,195**	**0.0237**	**6.23443**	**535,670**	**579,082**
Hombres	**264,672**	**8,144**	**0.0308**	**4.74515**	**251,275**	**278,069**
12 a 39 años	**169,054**	6,024	0.0356	3.15729	**159,145**	**178,963**
40 años y más	**95,550**	3,885	0.0407	2.88000	**89,159**	**101,941**
No especificado	**68**	68	1.0000			
Mujeres	**292,704**	**6,524**	**0.0223**	**3.37986**	**281,972**	**303,436**
12 a 39 años	**178,889**	6,004	0.0336	3.06689	**169,012**	**188,766**
40 años y más	**113,736**	3,735	0.0328	2.21415	**107,592**	**119,880**
No especificado	**79**	79	1.0000			
Población económicamente activa	**295,037**	**9,211**	**0.0312**	**5.38596**	**279,885**	**310,189**
Hombres	**182,262**	**6,159**	**0.0338**	**4.65691**	**172,130**	**192,394**
12 a 39 años	**108,023**	4,325	0.0400	2.73249	**100,908**	**115,138**
40 años y más	**74,171**	3,750	0.0506	3.15687	**68,002**	**80,340**
No especificado	**68**	68	1.0000	0.75465		
Mujeres	**112,775**	**4,369**	**0.0387**	**2.28999**	**105,588**	**119,962**
12 a 39 años	**73,124**	3,700	0.0506	2.33006	**67,038**	**79,211**
40 años y más	**39,651**	2,206	0.0556	1.63649	**36,022**	**43,280**
No especificado	**0**					
Población desocupada abierta	**14,226**	**1,681**	**0.1182**	**2.08239**	**11,461**	**16,991**
Hombres	**9,679**	**1,318**	**0.1362**	1.92763	**7,511**	**11,847**
12 a 39 años	**5,441**	763	0.1402	1.14868	**4,186**	**6,696**
40 años y más	**4,238**	1,107	0.2613	3.27990	**2,417**	**6,059**
No especificado	**0**					
Mujeres	**4,547**	**947**	**0.2083**	**2.03447**	**2,989**	**6,105**
12 a 39 años	**3,882**	899	0.2315	2.10591	**2,403**	**5,361**
40 años y más	**665**	299	0.4501	1.49980	**173**	**1,157**
No especificado	**0**					
Población de 12 años y más	**557,376**	**13,195**	**0.0237**	**6.23443**	**535,670**	**579,082**
Estrato 1	**110,160**	5,637	0.0512	6.33780	**100,887**	**119,433**
Estrato 2	**120,645**	4,467	0.0370	2.89193	**113,297**	**127,993**
Estrato 3	**64,142**	3,834	0.0598	4.44711	**57,835**	**70,449**
Estrato 4	**165,783**	9,725	0.0587	11.12787	**149,785**	**181,781**
Estrato 5	**96,646**	3,620	0.0375	3.13502	**90,691**	**102,601**
Población económicamente activa	**295,037**	**9,211**	**0.0312**	**5.38596**	**279,885**	**310,189**
Estrato 1	**60,773**	4,828	0.0794	7.99992	**52,831**	**68,715**
Estrato 2	**64,681**	3,811	0.0589	4.04327	**58,412**	**70,950**
Estrato 3	**36,272**	2,799	0.0772	3.74934	**31,668**	**40,876**
Estrato 4	**81,643**	5,792	0.0709	6.56505	**72,115**	**91,171**
Estrato 5	**51,668**	2,372	0.0459	2.73633	**47,766**	**55,570**
Población desocupada abierta	**14,226**	**1,681**	**0.1182**	**2.08239**	**11,461**	**16,991**
Estrato 1	**3,887**	1,277	0.3285	4.98694	**1,786**	**5,988**
Estrato 2	**3,702**	666	0.1799	1.02762	**2,606**	**4,798**
Estrato 3	**1,920**	467	0.2432	1.21551	**1,152**	**2,688**
Estrato 4	**2,351**	473	0.2013	0.97923	**1,573**	**3,129**
Estrato 5	**2,366**	556	0.2351	1.74259	**1,451**	**3,281**

de Chihuahua, y n_o es el tamaño de muestra en viviendas bajo muestreo aleatorio simple con reemplazo para el estimador de un total o una razón simple. Entonces al tamaño de muestra para la encuesta que esta siendo planeada será:

$$n = (n_o)\ deff.$$

Además, en el Cuadro 1 también se puede observar que casi en su totalidad las estimaciones del *deff* son mayores a la unidad, lo que era de esperarse debido a que se utilizaron conglomerados como unidades de muestreo en las distintas etapas de muestreo. De manera general se puede decir que, si el *deff* < 1 el diseño de muestreo empleado es mas eficiente que el muestreo aleatorio simple; si el *deff* ≈ 1 entonces ambos diseños son igualmente de eficientes; si el *deff* > 1 el diseño de muestreo empleado es menos eficiente que el muestreo aleatorio simple.

4.2 Precisiones estadísticas de razones simples

En el Cuadro 2 se presentan las estimaciones del error estándar (*e.e.*), coeficiente de variación (*c.v.*) y efecto de diseño (*deff*) para las estimaciones de la tasa de desempleo abierto y tasa neta de participación a nivel general, por sexo, sexo y grupo de edad y estrato, así como los correspondientes intervalos de confianza al 90%.

Cuadro 2. Precisiones estadísticas de las razones simples estimadas.

Variable	Estimación	e.e.	c.v.	deff	Intervalo de Confianza al 90% L.I.I.C.	L.S.I.C.
Tasa de desempleo abierto (ZPDA/ZPEA)	**0.04822**	**0.00531**	**0.11006**	**1.88654**	**0.03949**	**0.05695**
Hombres	**0.05311**	**0.00679**	**0.12785**	1.78195	**0.04194**	**0.06428**
12 a 39 años	**0.05037**	0.00716	0.14215	1.24697	**0.03859**	**0.06215**
40 años y más	**0.05714**	0.01419	0.24830	3.08670	**0.03380**	**0.08048**
No especificado	**0.00000**					
Mujeres	**0.04032**	**0.00803**	**0.19924**	1.96806	**0.02711**	**0.05353**
12 a 39 años	**0.05309**	0.01162	0.21889	2.00933	**0.03398**	**0.07220**
40 años y más	**0.01677**	0.00766	0.45667	1.56398	**0.00417**	**0.02937**
No especificado	**0.00000**					
Tasa neta de participación (ZPEA/ZP12)	**0.52933**	**0.00890**	**0.01682**	**2.34149**	**0.51469**	**0.54397**
Hombres	**0.68863**	**0.01229**	**0.01784**	2.18480	**0.66841**	**0.70885**
12 a 39 años	**0.63899**	0.01795	0.02808	2.31650	**0.60946**	**0.66852**
40 años y más	**0.77625**	0.01856	0.02392	2.07756	**0.74572**	**0.80678**
No especificado	**1.00000**					
Mujeres	**0.38529**	**0.01212**	**0.03146**	1.94663	**0.36535**	**0.40523**
12 a 39 años	**0.40877**	0.01509	0.03693	1.75493	**0.38395**	**0.43359**
40 años y más	**0.34862**	0.01794	0.05146	1.78176	**0.31911**	**0.37813**
No especificado	**0.00000**					
Tasa de desempleo abierto (ZPDA/ZPEA)	**0.04822**	**0.00531**	**0.11006**	**1.88654**	**0.03949**	**0.05695**
Estrato 1	**0.06396**	0.01790	0.27987	3.78886	**0.03451**	**0.09341**
Estrato 2	**0.05724**	0.01001	0.17494	1.03253	**0.04077**	**0.07371**
Estrato 3	**0.05293**	0.01173	0.22160	1.04872	**0.03363**	**0.07223**
Estrato 4	**0.02880**	0.00587	0.20391	1.03534	**0.01914**	**0.03846**
Estrato 5	**0.04579**	0.01078	0.23541	1.79950	**0.02806**	**0.06352**
Tasa neta de Participación (ZPEA/ZP12)	**0.52933**	**0.00890**	**0.01682**	**2.34149**	**0.51469**	**0.54397**
Estrato 1	**0.55168**	0.02500	0.04531	3.48328	**0.51056**	**0.59281**
Estrato 2	**0.53613**	0.01999	0.03729	2.73864	**0.50325**	**0.56901**
Estrato 3	**0.56550**	0.01841	0.03255	1.16184	**0.53522**	**0.59578**
Estrato 4	**0.49247**	0.01350	0.02742	1.41456	**0.47026**	**0.51468**
Estrato 5	**0.53461**	0.02246	0.04202	3.21823	**0.49766**	**0.57156**

De los resultados que aparecen en el Cuadro 2, se desprende que con base en el coeficiente de variación estimado, las estimaciones de la tasa neta de participación son por mucho, más precisas que las de la tasa de desempleo abierto, siendo la estimación de la TDA general, la que resultó con un coeficiente de variación más bajo, del 11.06 % y por tanto confiable. Para las estimaciones por estrato, se observa el mismo comportamiento antes visto, es decir, las estimaciones para cada uno de los estratos son menos precisas que la general.

En cuanto a los efectos de diseño estimados, se les aplica el mismo comentario de la sección anterior.

CONCLUSIONES

La necesidad de obtener para las estimaciones de los parámetros poblacionales que se generan a través de una encuesta por muestreo compleja, de estimaciones de sus errores de muestreo confiables y rápidas, es satisfecha con el método de los conglomerados últimos. Dicho método proporciona estimaciones insesgadas de la varianza para las estimaciones de totales, siempre y cuando se cumplan los supuestos establecidos en el subcapítulo 2.1.1. En el caso de que no se respeten supuestos como por ejemplo, el de utilizar muestreo sin reemplazo con una fracción de muestreo pequeña, en lugar de muestreo con reemplazo en la primera etapa de muestreo, o tener tan solo una UPM en algún estrato, en lugar de al menos dos (teniéndose que colapsar estratos). En ambos casos se obtendrá una estimación de varianza conservadora (véase subcapítulo 2.1.1).En cuanto, a la estimación de la varianza para el estimador de razón simple (linealizado), su sesgo se vuelve insignificante a medida de que el tamaño de muestra es grande (véase pie de pagina 68 en la página 20). Aunque no es el objetivo de este trabajo comparar técnicas de estimación de varianza, cabe hacer notar, que varios autores han evaluado el comportamiento del estimador de varianza del estimador de razón linealizado, con respecto a los que emplean técnicas de remuestreo, concluyendo, que prácticamente tienen el mismo comportamiento. Por otra parte como el lector lo habrá notado, el método de los conglomerados últimos es muy conveniente cuando no se necesitan estimaciones separadas de las contribuciones a la varianza debidas a las distintas etapas de muestreo, lo que significa obtener de manera rápida la estimación de la varianza, ya que esta se calcula con base a las UPM´s de la primera etapa de muestreo.

Una vez obtenida para la estimación de determinado parámetro poblacional de la estimación del error estándar y coeficiente de variación, el usuario de la información esta en posibilidades de decidir si la estimación es confiable o no. Además de poder generar diferentes intervalos de confianza de la estimación para diferentes niveles de confianza.

En cuanto al *deff*, su estimación permite básicamente comparar la eficiencia del diseño de muestreo empleado, contra el muestreo aleatorio simple. Además, de que su valor puede ser utilizado en la determinación del tamaño de muestra de una encuesta del mismo tipo que este siendo planeada.

Por otra parte, ya sea empleando el método de los conglomerados últimos o cualquier otro, es prácticamente imposible conocer las estimaciones de los errores estándar para la totalidad de estimaciones que genera una encuesta probabilística del tipo complejo, debido entre otras causas, al excesivo tiempo de procesamiento y alto costo económico. Afortunadamente en los últimos años se ha desarrollado un método indirecto para el cálculo del error estándar llamado, el método de la función generalizadora de varianza. De manera breve se puede decir, que este último método se aplica a encuestas complejas continuas y que busca adecuar un modelo matemático para la estimación de errores estándar. Esto requiere obtener estimaciones de los errores estándar y efectos de diseño, ya sea con el método de los conglomerados últimos o con técnicas de remuestreo, para distintos periodos de tiempo.

Sin duda el conocimiento que proporcionan materias impartidas a lo largo de la carrera de Actuaría, permiten abordar el tema del cálculo de la varianza en una encuesta compleja, el cual es considerado generalmente en estudios de postgrado bajo bibliografía escasa y difícil de conseguir.

Finalmente, el objetivo implícito en esta tesina de explicar la importancia del cálculo de la varianza para las estimaciones obtenidas a través de una encuesta por muestreo probabilística, compleja o no, se busco cumplir con las definiciones e interpretaciones presentadas en los Capítulos 1 y 4. De esta manera bajo la perspectiva de este trabajo, el usuario de la información debe considerar a la estimación del parámetro poblacional, como un valor sujeto al error de muestreo y no como un valor censal sin error alguno.

BIBLIOGRAFÍA

(1) Azorín, Francisco. Métodos y Aplicaciones de Muestreo. Alianza Editorial. Madrid, 1986.

(2) Raj, Des. Teoría del Muestreo. Fondo de Cultura Económica. México, 1992.

(3) Kish, Leslie. Muestreo de Encuestas. Editorial Trillas. México, 1982.

(4) Cochran, William G. Técnicas de Muestreo. CECSA. México, 1992.

(5) Särndal, Carl-Erick, Swensson, Bengt y Wretman, Jan. Model Assisted Survey Sampling. Springer-Verlag. New York, 1993.

(6) Lehtonen, Risto y Pahkinen, Erkki J. Practical Methods for Design of Complex Surveys. John Wiley & Sons. England, 1995.

(7) Wolter, Kirk M. Introduction to Variance Estimation. Springer-Verlag. New York, 1985.

(8) Hansen, Morris H., Hurwitz, William N. y Madow, William G. Sample Survey Methods and Theory Vol. I. John Wiley & Sons. New York, 1953.

(9) Statistical Laboratory, Iowa State University. PC CARP. 1986.

(10) Skinner, C. J. y Smith, T. M. F. Analysis of Complex Surveys. John Wiley & Sons. England, 1989.

(11) Lessler, J. T. A Taxonomy of Error Sources and Error Maesures for Surveys, Final Report. Research Triangle Park, NC: Research Triangle Institute.

(12) K. R. W. Brewer and Muhammad Hanif. Sampling with Unequal Probabilities. Springer-Verlag. New York Heidelberg Berlin.

(13) Durbin, J.Some Results in Sampling Theory When the Units Are Selected with Unequal Probabilities. Journal of the Royal Statistical Society, B 15, 262-269. 1953.

Sitios web

www.inegi.gob.mx