Departamento Administrativo Nacional de Estadística

Dirección de Regulación, Planeación, Estandarización y Normalización
-DIRPEN-

Especificaciones de Coeficiente y Varianza
Encuesta de Consumo Cultural
-ECC-

Julio 2008

TABLA DE CONTENIDO

1. ESPECIFICACIONES DE COEFICIENTE Y VARIANZA

1.1 Factores de expansión

De acuerdo a la teoría de muestreo el factor de expansión es la capacidad que tiene cada individuo seleccionado en una muestra probabilística para representar el universo en el cual esta contenido. Es decir, es la magnitud de representación que cada selección posee para describir una parte del universo de estudio. Cuando el diseño es MAS se asume que individuos dentro de una misma unidad de muestreo tienen la misma capacidad de representar al universo en consideración, en tanto que diferentes unidades de muestreo deben reflejar lo mejor posible la densidad y distribución del universo estudiado.

Todas las pautas anteriores se logran con la construcción u obtención de marcos muéstrales que deben proveer información fiel sobre las características demográficas principales del universo que se pretende estudiar; además debe permitir de alguna forma, ubicar a todos y cada uno de los individuos pertenecientes a dicho universo.

Un diseño muestral MAS se traduce como un diseño de muestreo aleatorio simple y se dice que dicho diseño es probabilístico porque asigna una probabilidad de selección a todas y cada una de las unidades del universo o la unidad muestral (cuando el diseño es en etapas); es decir, todos los individuos en consideración de los que se pretende inferir tienen una probabilidad de ser seleccionados; en particular esta probabilidad es la misma con el diseño MAS.

El factor de expansión por teoría para un diseño MAS sobre k unidades de muestreo esta definido por:

$$\pi_{\mathrm{I}k} = \frac{N_{\mathrm{I}k}}{n_{\mathrm{I}k}}$$

Donde

$N_{\mathrm{I}k}$ Denota el tamaño total de elementos en la unidad de muestreo.

$n_{\mathrm{I}k}$ Denota el número de elementos a ser seleccionados dentro de la unidad de muestreo.

Es decir, para el diseño en consideración de la encuesta de consumo cultural $ESTMAS - MAS - MASC$ y de acuerdo a los diagramas presentados anteriormente, se tienen los factores de expansión por etapas de la siguiente manera:

ETAPA 1: Selección de municipios dentro de los estratos

$$f\exp_{I(i)} = \frac{N_{\mathrm{I}(i)}}{n_{\mathrm{I}(i)}} \quad \text{para} \quad i = 1,2,3,4$$

con

$N_{\mathrm{I}(i)}$: Cantidad de municipios en el estrato i

$n_{\mathrm{I}(i)}$: Numero de municipios seleccionados en el estrato i

ETAPA 2: Selección de conglomerados dentro de municipios

$$f\exp_{II(j)} = \frac{N_{II(j)}}{n_{II(j)}} \quad \text{para} \quad j = 1,\ldots,n_{I(i)}$$

con

$N_{II(j)}$: Cantidad de conglomerados en el municipio j

$n_{II(j)}$: Cantidad de conglomerados seleccionados en el municipio j

ETAPA 3: Selección de hogares dentro de conglomerado

$$f\exp_{III(k)} = \frac{N_{III(k)}}{n_{III(k)}} \quad \text{para} \quad k = 1,\ldots,n_{II(j)}$$

con

$N_{III(k)}$: Cantidad de hogares dentro del conglomerado k

$n_{III(k)}$: Cantidad de hogares seleccionados dentro del conglomerado k

Nota: El factor de expansión correspondiente a cada individuo dentro de un hogar seleccionado es igual a la unidad ($f\exp_{individuo} = 1$), puesto que se eligen todas las personas de cinco años y más, es decir, esta parte del diseño es la que se conoce como muestreo por conglomerados.

1.2 Factor de expansión final

El factor de expansión final se define como el producto de los factores de expansión de las tres etapas de diseño:

$$F\exp = \left(f\exp_{I(i)}\right)\left(f\exp_{II(j)}\right)\left(f\exp_{III(k)}\right)$$

Este factor se aplica a cada uno de los individuos pertenecientes a los hogares seleccionados en la última etapa del diseño. La suma de los factores de expansión de todos los individuos de la muestra provee una estimación del tamaño del universo de estudio.

1.3 Estimadores de Totales y Razones

Con los factores de expansión calculados se define el estimador del total de una variable de estudio X (t_x) en el universo, en función de los valores observados con la muestra, de la siguiente manera:

$$\hat{t}_x = \hat{t}_\pi(x) = \sum_{k \in s_\Omega} (F\exp) * x_k$$

Donde s_Ω es una muestra del universo de estudio con valores particulares medidos x_k.

Y el estimador de la varianza para la anterior expresión es:

$$\hat{V}ar(\hat{t}_x) = \frac{N^2(1-f)}{n} S^2{}_{xk} \quad \text{Con} \quad S^2{}_{xk} = \frac{\sum_{s_\Omega}(x_k - \bar{x}_s)^2}{(n-1)} \quad ; \quad f = \frac{n}{N}$$

con N y n tamaños de universo y muestra respectivamente; siempre que el diseño de muestreo sea MAS.

Si Y es otra variable de estudio, en general el estimador de la razón $Z = \frac{t_y}{t_x}$ se define por

$$\hat{Z} = \frac{\hat{t}_\pi(y)}{\hat{t}_\pi(x)}$$

El cual no es un estimador lineal, luego el cálculo de la estimación de su varianza difiere de los métodos convencionales. Aplicando el método de linearizacion de Taylor y construyendo una nueva variable que recoge dicho termino de varianza, se obtiene un estimador de la varianza del estimador.

El estimador de la varianza para el estimador de razón es

$$\hat{V}ar(\hat{z}) = \frac{1}{\hat{t}^2{}_\pi}\left[\hat{V}(\hat{t}_y) + \hat{z}^2\hat{V}(\hat{t}_x) - 2\hat{z}Cov(\hat{t}_y;\hat{t}_x)\right]$$

o calculándose directamente con la fórmula común de varianza en el diseño MAS, a través de la construcción de una nueva variable U para cada uno de los elementos de la muestra

$$u_k = \frac{1}{\hat{t}_x} y_k - \frac{\hat{t}_y}{\hat{t}^2{}_x} x_k$$

Los promedios y las proporciones se trabajan como casos particulares del estimador de razón cuando los numeradores y los denominadores cambian en concepto, y estos pueden pasar de ser variables continuas a categóricas.
Una forma de generar estimaciones para diferentes proporciones o promedios (dependiendo del requerimiento de investigación) es mediante la formación de variables indicadoras de dominio que toman el valor de 1 si un individuo de la muestra pertenece a un dominio de interés y 0 en caso contrario.
Entiéndase al dominio de un universo como un conjunto de individuos contenido en éste, que presentan una característica discriminatoria Z que los ubica dentro de una categoría particular observada.

La variable indicadora se establece como

$$I^z{}_k = \begin{cases} 1 & \text{si el individuo k tiene la caracteristica Z} \\ 0 & \text{en caso contrario} \end{cases}$$

Es así que el estimador del total para la variable $I^z{}_k$ se convierte en el estimador del total de individuos del universo pertenecientes a un dominio Z; y en particular, si $I^z{}_k$ siempre es 1 para cualquier individuo de la muestra, es simplemente el estimador del total de individuos en el universo.

Las proporciones y los promedios solicitados en cuadros de salida, se construyen formando cocientes de diversos estimadores de dominios y de totales. Las formulas no varían.

Para el diseño multietápico $ESTMAS - MAS - MASC$ las formulas de estimación por etapas están definidas como sigue

- *Etapa 1:* $\hat{t}_{I(i)x} = \sum_{j \in sI} \frac{N_{I(i)}}{n_{I(i)}} (\hat{t}_{II(i)(j)x})$ con estimador de varianza

$$\hat{V}_I(\hat{t}_{I(i)x}) = \left(\frac{N_{I(i)}{}^2}{n_{I(i)}}\right)\left(1 - \frac{n_{I(i)}}{N_{I(i)}{}^2}\right) S^2_{sI}(\hat{t}_{II(i)(j)x}) + \sum_{j \in sI} \frac{N_{I(i)}}{n_{I(i)}} \hat{V}_{II}(\hat{t}_{II(i)(j)x})$$

donde $$S_{sI}^2(\hat{t}_{II(i)(j)x}) = \frac{\sum_{sI}(\hat{t}_{II(i)(j)x} - \bar{\hat{t}}_{II(i)x})^2}{\left(n_{I(i)} - 1\right)}$$

- *Etapa 2:* $\hat{t}_{II(i)(j)x} = \sum_{k \in sII} \frac{N_{II(j)}}{n_{II(j)}}(\hat{t}_{III(i)(j)(k)x})$ con estimador de varianza

$$\hat{V}_{II}(\hat{t}_{II(i)(j)x}) = \left(\frac{N_{II(j)}}{n_{II(j)}}^2\right)\left(1 - \frac{n_{II(j)}}{N_{II(j)}}^2\right) S_{sII}^2(\hat{t}_{III(i)(j)(k)x}) + \sum_{k \in sII} \frac{N_{II(j)}}{n_{II(j)}} \hat{V}_{III}(\hat{t}_{III(i)(j)(k)x})$$

donde $$S_{sII}^2(\hat{t}_{II(i)(j)(k)x}) = \frac{\sum_{sII}(\hat{t}_{III(i)(j)(k)x} - \bar{\hat{t}}_{I(i)(j)x})^2}{\left(n_{II(j)} - 1\right)}$$

- *Etapa 3:* $\hat{t}_{III(i)(j)(k)x} = \sum_{l \in sIII} \frac{N_{III(k)}}{n_{III(k)}}(\hat{t}_{IV(i)(j)(k)(l)x})$ con estimador de varianza

$$\hat{V}_{III}(\hat{t}_{II(i)(j)(k)x}) = \left(\frac{N_{III(k)}}{n_{III(k)}}^2\right)\left(1 - \frac{n_{III(k)}}{N_{III(k)}}^2\right) S_{sIII}^2(\hat{t}_{IV(i)(j)(k)(l)x}) + \sum_{k \in sIII} \frac{N_{III(k)}}{n_{III(k)}} \hat{V}_{IV}(\hat{t}_{IV(i)(j)(k)(l)x})$$

donde $$S_{sIII}^2(\hat{t}_{IV(i)(j)(k)(l)x}) = \frac{\sum_{sIII}(\hat{t}_{IV(i)(j)(k)(l)x} - \bar{\hat{t}}_{I(i)(j)(k)x})^2}{\left(n_{III(k)} - 1\right)}$$

Finalmente, para obtener la estimación general a la que se quiere llegar, se tiene:

$$\hat{t}_x = \sum_{i=1}^{3}(\hat{t}_{I(i)x})$$

Con estimador de varianza

$$\hat{V}_I(\hat{t}_{I(i)x}) = \sum_{I=1}^{3} \hat{V}_I(\hat{t}_{I(i)x})$$

Si el estimador resulta ser de razones, las formulas para el calculo de la varianza no se alteran si se emplean en lugar de x_l los valores de u_l .

1.4 Programa de Estimación

El programa de estimación es una rutina en SAS que ayuda a realizar los cálculos anteriores, con base en las observaciones de la base de datos de la encuesta.

El programa está compuesto por una macro principal que se encarga de estimar una sola etapa agrupando por llaves especiales a los individuos sobre los cuales deben realizarse las sumas y las varianzas de la respectiva etapa. Es decir, la base de datos debe llevar una llave que le permita al programa discriminar los individuos que van en la etapa a trabajar; y estas llaves se construyen haciendo mención a la unidad muestral que se debe trabajar en cada etapa.

Finalmente, la salida es otra base de datos que contiene los estimativos de total y de varianza de la etapa trabajada, que a su vez retroalimenta la misma macro para trabajar la etapa siguiente con otra llave. La macro se itera tres veces con tres llaves diferentes (una vez por cada etapa), hasta obtener los resultados de la primera etapa en donde mediante el empleo de consultas y pasos data, se recalculan las fórmulas para obtener la estimación final, teniendo en cuenta la estratificación.

En la generación de cuadros de salida se emplean tres macro que a su vez utilizan la macro principal de estimación por etapas, para generar cuadros de salida cruzados en dos variables y calcular la proporción por dominio respecto al total estimado. Los cuadros salen del programa con todos los dominios de una variable enfrentados con todos los dominios de la otra.