Informe sobre el Cálculo de Errores de Muestreo

Encuesta sobre Condiciones de Vida - ECV

INDICE

1. Introducción ***3***

2. Método de expansión de Taylor ***3***

3. Cálculo de errores. ***4***

3.1 Diseño Muestral *4*
3.2 Procedimiento de cálculo *5*
3.3 Estadísticos y dominios para el cálculo de errores en la ECV *6*
3.4 Resultados e Interpretación *7*

Bibliografía ***9***

1. Introducción.

Podemos definir error de muestreo como la imprecisión que se comete al estimar una característica de la población de estudio (parámetro) mediante el valor obtenido a partir de una parte o muestra de esa población (estadístico).

Este error depende de muchos factores, entre ellos, del procedimiento de extracción de esa parte de la población (diseño muestral), del número de unidades que se extraen (tamaño de la muestra), de la naturaleza de la característica a estimar, etc. Una expresión generalizada del error de muestreo sería la siguiente:

$$\text{Error de muestreo} = \sqrt{Var(\hat{\theta})} \qquad (1)$$

Siendo $\hat{\theta}$ el estadístico de interés (media, total, proporción,..). Este estadístico tomará valores distintos dependiendo de la muestra extraída. La variabilidad del estadístico en el muestreo determinará el error muestral.

La expresión de este error cambiará dependiendo de la técnica de muestreo utilizada, haciéndose más complejo su cálculo conforme más complicado sea el diseño muestral. La mayoría de las encuestas de EUSTAT tienen un diseño muestral complejo que incluye estratificación, probabilidades de selección desiguales y varias etapas de muestreo. Estos diseños se aplican con el fin de producir estimadores puntuales lo más buenos posibles, pero en la práctica complican sobremanera la estimación de los errores de muestreo.

La literatura ha sugerido algunas alternativas a los métodos convencionales de cálculo de errores muestrales. Estas técnicas heurísticas proporcionan una buena estimación del error muestral a partir de los pesos finales y las características del diseño muestral [1], [5].

En lo que sigue introduciremos estos métodos y su aplicación concreta en el caso de la Encuesta sobre Condiciones de Vida.

2. Método de expansión de Taylor [2], [5].

Este método permite calcular estimaciones del error muestral para totales, medias y proporciones en muestras con estratificación, clústers y probabilidades desiguales, como es el caso de muchas operaciones estadísticas en EUSTAT. El método admite en su formulación varias etapas de muestreo pero sólo tiene en consideración la variabilidad de las unidades de primera etapa para la estimación del error muestral. Se obtienen aproximaciones lineales del estimador y se calcula la varianza utilizando ésta como estimación del error muestral.

La expresión para el cálculo de la varianza estimada para la media poblacional es la siguiente:

$$\hat{V}(\bar{\hat{Y}}) = \sum_{h=1}^{H} \frac{n_h(1-f_{h)}}{n_h - 1} \sum_{i=1}^{n_h} (e_{hi.} - \bar{e}_{h..})^2 \qquad (2)$$

Donde:

$$e_{hi.} = \frac{\sum_{j=1}^{m_{hi}} w_{hij} (y_{hij} - \hat{\bar{Y}})}{w...}$$

$$\bar{e}_{h..} = \frac{\sum_{j=1}^{n_h} e_{hi.}}{n_h}$$

$$y$$

$$w... = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}$$

Notación:

$h = 1, 2, \ldots, H$ indica el estrato con un total de H estratos.
$i = 1, 2, \ldots, n_h$ indica el número de clusters en el estrato h, con un total de n_h clusters.
$j = 1, 2, \ldots, m_{hi}$ indica el número de unidad dentro del cluster i del estrato h, con un total de m_{hi} unidades
$n = \sum_{h=1}^{H} \sum_{i=1}^{n_n} m_{hi}$ es el número total de observaciones en la muestra.

w_{hij} indica el elevador de la observación j en el cluster i del estrato h
$y_{hij} = (y_{hij}(1), y_{hij}(2), \ldots, y_{hij}(P))$ son los valores observados de la variable Y en la observación j del cluster i del estrato h. (variables numéricas y categóricas).

El procedimiento PROC SURVEYMEANS del paquete estadístico SAS [4], implementa este método de estimación de errores muestrales y será la herramienta que se utilice para el cálculo de los errores muestrales del fichero de familias de la operación que nos ocupa.

3. Cálculo de errores

3.1 Diseño Muestral.

La ECV es una encuesta por muestreo sobre la población de la CAE de 6 y más años. La encuesta va dirigida a una muestra de individuos seleccionados aleatoriamente en dos etapas. En primer lugar se seleccionan las viviendas, y dentro de cada vivienda se sortea un individuo de forma aleatoria. El método de muestreo de las viviendas es el estratificado (primera etapa) y dentro de las viviendas el muestreo, para seleccionar las personas, es aleatorio (segunda etapa).

A partir del año 2009 los estratos de las viviendas son zonas geográficas o agrupaciones de comarcas, y además, las capitales de los Territorios Históricos. En total, se trata de 12 estratos o zonas, que pasamos a detallar a continuación.

- Zona 1: Llanada Alavesa y Cantábrica Alavesa, excepto Vitoria-Gasteiz.
- Zona 2: Resto de Alava, que comprende el resto de las comarcas: Valles Alaveses, Montaña Alavesa, Rioja Alavesa y Estribaciones del Gorbea.
- Zona 3: Gran Bilbao, excepto Bilbao.
- Zona 4: Duranguesado.
- Zona 5: Gernika-Bermeo, Plentzia-Mungia y Markina-Ondarroa.
- Zona 6: Arratia-Nervión y Encartaciones.
- Zona 7: Donostialdea y Bajo Bidasoa, excepto Donostia-San Sebastián
- Zona 8: Alto Deba, Bajo Deba y Urola Costa.
- Zona 9: Tolosaldea y Goierri.
- Zona 10: Vitoria-Gasteiz
- Zona 11: Bilbao
- Zona 12: Donostia-San Sebastián

En la segunda etapa, ya se ha dicho que una vez seleccionada la vivienda, el muestreo es aleatorio entre todos los individuos de la vivienda.

3.2 Procedimiento de cálculo.

La encuesta se explota a nivel de familias (viviendas) e individuos por lo que se dispondrá de dos ficheros: fichero de familias y fichero de individuos.

La sintaxis básica del procedimiento de SAS implementado para el cálculo de errores es la siguiente [4]:

PROC SURVEYMEANS < *nombre_fichero* > < *opciones de salida* >;
BY *variables* ; /*cálculo de errores por subpoblaciones independientes*/
CLASS *variables* ; /*cálculo de errores para variables cualitativas*/
DOMAIN *variables* ; /*variables que delimitan el dominio/cruce para el que se calculan los errores*/
STRATA *variables* < / *option* > ; /*variable que indica el estrato en el muestreo estatificado*/
VAR *variables* ; /* variables cuantitativas y cualitativas para las que se pretende calcular los errores muestrales*/
WEIGHT *variable* ; /* variable peso pre-calculada (opcional)*/

Los parámetros generales de esta sintaxis para el caso concreto de la ECV serán los siguientes:

STRATA = zonas+capitales
WEIGHT = Elevador anual de familias o de individuos según corresponda.
VAR = Variables propias de la encuesta.
DOMAIN = Cruces por variables geográficas y socio-demográficas.

3.3 Estadísticos y dominios para el cálculo de errores en la ECV

Siguiendo el criterio adoptado por otras encuestas de EUSTAT para la publicación de los errores muestrales, se difundirán tantas tablas de errores como cruces se publiquen en el apartado de tablas estadísticas de la Web para la operación dada. En este caso estas tablas son:

Tablas de Coeficientes de Variación e Intervalos de Confianza. Encuesta de Condiciones de Vida. ECV

Familias
Familias de la C.A. de Euskadi por el grado de relaciones que mantienen con familiares, amigos/as y vecinos/as, según el territorio histórico (%). Coeficientes de variación.

Familias de la C. A. de Euskadi por el estado del medio ambiente y el grado de seguridad ciudadana, según el territorio histórico (%). Coeficientes de variación.

Familias de la C. A. de Euskadi por los servicios del edificio y del entorno, según el territorio histórico (%). Coeficientes de variación.

Familias de la C. A. de Euskadi por la situación económica objetiva y la apreciación subjetiva, según el territorio histórico (%). Coeficientes de variación.

Individuos
Población de 6 y más años de la C.A. de Euskadi, por el grado de dependencia, según el territorio histórico (%). Coeficientes de variación.

Población de 16 y más años ocupada, de la C.A. de Euskadi por las condiciones del centro de trabajo, según territorio histórico (%). Coeficientes de variación.

Población de 6 y más años de la C. A. de Euskadi por la frecuencia con que sale a comer o cenar los fines de semana según territorio histórico y sexo (%). Coeficientes de variación.

Población de 6 y más años de la C. A. de Euskadi por la existencia y tipo de relaciones con amigos/as y vecinos/as, según territorio histórico (%). Coeficientes de variación.

Población de 6 y más años de la C.A. de Euskadi por la periodicidad de las revisiones médicas y sexo (miles). Coeficientes de variación.

Población de la C.A. de Euskadi que ha consultado al médico especialista por número de consultas y sexo (miles) Coeficientes de variación.

Población de la C.A. de Euskadi que ha consultado al médico de atencion primaria por número de consultas y sexo (miles). Coeficientes de variación.

3.4 Resultados e Interpretación.

Junto con el estadístico de interés se proporcionan otras medidas del error que son de utilidad y ayudan a la interpretación del mismo. Entre éstas, las más interesantes son:

- **El Coeficiente de Variación.** Es una medida relativa del error que permite comparar precisiones entre distintos grupos o poblaciones. Se trata de una magnitud adimensional muy utilizada como medida del error muestral y su expresión es:

$$CV = \frac{\sqrt{Var(\hat{\theta})}}{\hat{\theta}}$$

- Intervalo de Confianza al 95%. Este intervalo de confianza se basa en la distribución en el muestreo del estadístico (proporción, media, tasa,…). Por el Teorema Central del Límite, la mayor parte de las veces podemos asumir una ley Normal[1] para los estadísticos más comunes, por lo que la construcción de este intervalo vendrá dada por la siguiente expresión:

$$(\hat{\theta} - 1{,}96\sqrt{Var(\hat{\theta})}, \quad \hat{\theta} + 1{,}96\sqrt{Var(\hat{\theta})})$$

El valor 1,96 es el percentil de una distribución Normal con media 0 y desviación típica 1 que encierra una probabilidad del 95%. Esto permite afirmar que el intervalo calculado para el estadístico θˆ contiene al verdadero valor del parámetro poblacional en el 95% de los casos (posibles muestras).

Con esta información proporcionada, se construirán las tablas definitivas de errores que contendrán la estimación del estadístico, el límite inferior y superior del intervalo de confianza al 95% y el coeficiente de variación en porcentaje. A continuación se presenta un modelo de tabla de difusión de errores:

F5. Coeficientes de Variación e Intervalos de Confianza para el porcentaje de familias por la situación económica objetiva y la apreciación subjetiva (%).

	C.A. de Euskadi	Araba / Alava	Bizkaia	Gipuzkoa
Situación económica objetiva				
Mala	12,5	14,6	13,2	10,3
L. Inferior 95%	11,5	12,5	11,8	8,9
L. Superior 95%	13,4	16,6	14,6	11,6
CV(%)	**3,8**	**7,3**	**5,5**	**6,7**
Normal	42,4	44,1	41,5	42,9
L. Inferior 95%	40,9	40,9	39,3	40,6
L. Superior 95%	43,8	47,3	43,8	45,3
CV(%)	**1,8**	**3,7**	**2,7**	**2,8**
Buena	45,2	41,4	45,3	46,8
L. Inferior 95%	43,7	38,2	43,1	44,5
L. Superior 95%	46,6	44,5	47,4	49,0
CV(%)	**1,6**	**3,9**	**2,4**	**2,5**

Fuente: EUSTAT. Encuesta de Condiciones de Vida. ECV - 2009

Otra forma de interpretar esta información consiste en calcular el **error relativo** al 95% de confianza, que se obtiene al multiplicar el percentil 1,96 por el Coeficiente de Variación. Este error relativo nos permite hablar en términos de puntos porcentuales del valor de la estimación.

Para la tabla anterior, el error relativo al 95% para el porcentaje de familias con situación económica normal es del 3,53 % (1,96*1,8). O lo que es lo mismo, a un nivel de confianza del 95%

[1] Se asume un tamaño muestral suficientemente 'grande' (n >30). Cuando no podemos realizar esta asunción, el intervalo de confianza se calculará con el correspondiente percentil al 95% de la distribución t-Student con n-1 grados de libertad.

podemos afirmar que el verdadero valor del porcentaje de familias con situación económica normal oscila en un intervalo del ±3,14 % de la estimación dada. Es decir:

(42,4± 0,0353*42,4) = entre 40,9% y 43,8%

Es importante señalar aquellas estimaciones que sobrepasen un determinado porcentaje del error relativo al 95%, para que el usuario tome las debidas cautelas a la hora de interpretar la información dada. Un umbral razonable estaría en aquellas estimaciones que sobrepasen el 20% de error relativo (C.V. > 10% aprox.), señalando de forma especial aquellas casillas donde este error sea mayor que el 30% (C.V. > 15% aprox.).

Bibliografía

[1] EUSTAT. 1998 "*El método de replicación para la estimación de errores de muestreo*". D. Morganstein, "Seminario Internacional de Estadística, 37".. **http://www.eustat.es/prodserv/vol37_c.html**

[2] Fuller, W. A. (1975), *"Regression Analysis for Sample Survey,"* Sankhy , 37, Series C, Pt. 3, 117 - 132.

[3] Kalton, G. (1979) *"Ultimate Cluster Sampling"* Journal of the Royal Statistical Society. Series A, Vol.142, No. 2, pp. 210-222.

[4] Sas Institute Inc. (2004), *"SAS/STAT® 9.1 Guía de Usuario".* Copyright © 2004, Cary, NC, USA. ISBN 1-59047-243-8

[5] Woodruff, R. S. (1971), "*A Simple Method for Approximating the Variance of a Complicated Estimate"* Journal of the American Statistical Association, 66, 411 -414.