Informe sobre el Cálculo de Errores de Muestreo

Encuesta de la Sociedad de la Información
(ESI- Familias)

INDICE

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. Además, las incidencias que se producen durante la recogida de información, el ajuste a determinadas características de la población (post-estratificación) y otros factores a lo largo del desarrollo de una encuesta, implican variaciones en el cálculo de los elevadores o pesos finales.

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 [3], [5].

En lo que sigue introduciremos estos métodos y su aplicación concreta en el caso de la Encuesta de la Sociedad de la Información desde el periodo 2000.

2. Método de expansión de Taylor **[3], [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 obtiene aproximaciones lineales del estimador y calcula su varianza utilizando ésta como estimación del error muestral.

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

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

Donde:

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

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

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 en la operación que nos ocupa.

3. Cálculo de errores ESI - Familias.

3.1 Diseño Muestral [1]

La Encuesta sobre la Sociedad de la Información a Familias (ESI-Familias) es una encuesta por muestreo sobre la población de la C.A. de Euskadi de 6 y más años. Esta encuesta toma como base del muestreo el panel de viviendas familiares seleccionadas para la Encuesta de la Población en Relación con la Actividad (PRA) en el mismo trimestre de referencia. A partir de 2004, esta muestra consta de 5.088 viviendas y se extrae aleatoriamente del Directorio de Viviendas de modo estratificado a nivel de Territorio Histórico. Dentro de cada estrato, se muestrean viviendas de forma sistemática (con la misma probabilidad) [2].

Dentro de cada vivienda la selección de la primera persona se realiza de forma aleatoria mediante una tabla de Kish, y además, cuando hay ocupados o estudiantes, uno de cada es seleccionado por el mismo procedimiento. Desde el año 2003 se completa la muestra con todos los menores de 6 a 14 años hasta llegar a una muestra cercana a los 7.500 individuos.

La encuesta que se explota anualmente (1er trimestre del año de referencia) y los resultados hacen referencia tanto a individuos como a familias, mereciendo un tratamiento especial los usuarios de Internet.

El diseño descrito se adapta perfectamente a las especificaciones del método heurístico expuesto en el apartado anterior. Sólo habrá que indicar los parámetros requeridos por el procedimiento de SAS para la correcta estimación de la varianza.

3.2 Procedimiento de cálculo.

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*/
CLUSTER *variables* ; */variable que indica el clúster en el muestreo por conglomerados*/
DOMAIN *variables* ; /*variables que delimitan el dominio/cruce para el que se calculan los errores*/
RATIO *variable/variable* ; /*variables ratio para las cuales se quiere calcular el error muestral*/
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 ESI Familias serán los siguientes:

CLUSTER = Identificador de vivienda.
STRATA = Territorio Histórico.
WEIGHT = Elevador anual de personas /Elevador anual de familias.
VAR = Variables de equipamiento y uso de las Tecnologías de la Información.
DOMAIN = Cruces por variables socio-demográficas y económicas.

3.3 Estadísticos y dominios para el cálculo de errores en la ESI - Familias.

Se estimarán errores de muestreo para los siguientes cruces y estadísticos:

Familias

- Familias de la C.A. de Euskadi por equipamientos TIC, según el Territorio Histórico (%) 2011. Errores de muestreo.

- Familias de la C.A. de Euskadi por equipamientos TIC, según el tipo de familia (%) 2011. Errores de muestreo.

- Familias de la C.A. de Euskadi por equipamientos televisivos en el hogar, según Territorio Histórico (%) 2011. Errores de muestreo.

- Familias de la C.A. de Euskadi por equipamientos televisivos en el hogar, según tipo de familia (%) 2011. Errores de muestreo.

Población

- Población de 15 y más años de la C.A. de Euskadi por equipamientos TIC y televisivos en el hogar, según Territorio Histórico (%) 2011. Errores de muestreo.
- Población de 15 y más años de la C.A. de Euskadi con ordenador en el hogar, por sexo y edad, según Territorio Histórico (%) 2011. Errores de muestreo.
- Población de 15 y más años de la C.A. de Euskadi con ordenador en el hogar, por nivel de instrucción y relación con la actividad, según Territorio Histórico (%) 2011. Errores de muestreo.
- Población de 15 y más años de la C.A. de Euskadi con internet en el hogar, por sexo y edad, según Territorio Histórico (%) 2011. Errores de muestreo.
- Población de 15 y más años de la C.A. de Euskadi con internet en el hogar, por nivel de instrucción y relación con la actividad, según Territorio Histórico (%) 2011. Errores de muestreo.
- Población de 15 y más años de la C.A. de Euskadi por Territorio Histórico y características sociodemográficas, según equipamientos TIC del hogar(%) 2011. Errores de muestreo.
- Población de 15 y más años de la C.A. de Euskadi por posibilidad de acceso a internet y Territorio Histórico (%) 2011. Errores de muestreo.

Uso de Internet

- Población de 15 y más años de la C.A. de Euskadi usuaria de Internet por sexo y edad, según Territorio Histórico (%) 2011. Errores de muestreo.
- Población de 15 y más años de la C.A. de Euskadi usuaria de Internet por nivel de instrucción y relación con la actividad, según Territorio Histórico (%) 2011. Errores de muestreo.
- Población de 15 y más años de la C.A. de Euskadi usuaria de Internet por servicios utilizados y duración media de la última conexión, según Territorio Histórico (%) 2011. Errores de muestreo.
- Población de 15 y más años de la C.A. de Euskadi usuaria de Internet por lugar de acceso e idiomas utilizados, según Territorio Histórico (%) 2011. Errores de muestreo.

- Población de 15 y más años de la C.A. de Euskadi que ha comprado por Internet por bienes adquiridos, forma de pago y opinión sobre la seguridad de pago, según Territorio Histórico (%) 2011. Errores de muestreo.
- Población de 15 y más años de la C.A. de Euskadi por el fin de la conexión, frecuencia de acceso, duración de la conexión y territorio histórico (%) 2011. Errores de muestreo.
- Población de 15 y más años de la C.A. de Euskadi por los servicios utilizados, tipos de web visitadas y territorio histórico (%) 2011. Errores de muestreo.

Podemos resumir lo anterior en las siguientes tablas según estadístico y variable de cruce:

Estadístico	Equipamientos TIC	Territorio Histórico	Sexo	Edad	Nivel de instrucción	Relación con la actividad	Tipo de familia	Relación con Internet	Comercio electrónico
Porcentaje población 15 y más años	X	X	X	X	X	X	X		
Total población de 15 y más años (miles)	X	X	X	X	X	X	X		
Porcentaje población 15 y más años usuaria de Internet		X	X	X	X	X		X	X
Total población de 15 y más años usuaria de Internet (miles)		X	X	X	X	X	X	X	X
Porcentaje de familias (%)	X	X					X		
Total familias (en miles)	X	X					X		

3.4 Resultados e Interpretación.

Aparte de la estimación del error de muestreo (2), SAS proporciona 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:

$$CV = \frac{\sqrt{Var(\hat{\theta})}}{\hat{\theta}} \qquad (3)$$

- **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:

$$\left[\hat{\theta} - 1{,}96\sqrt{Var(\hat{\theta})},\quad \hat{\theta} + 1{,}96\sqrt{Var(\hat{\theta})}\right] \qquad (4)$$

  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 $\hat{\theta}$ contiene al verdadero valor del parámetro poblacional en el 95% de los casos (posibles muestras).

[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.

Con la información proporcionada por SAS, 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:

Familias de la C.A. de Euskadi por equipamientos TIC en el hogar (%) 2011. Errores de muestreo

	Total (miles)	Ordenador	Internet	Telefono m	E-mail
C.A. de Euskadi					
Estimación	846,3	62,4	57,0	90,1	55,1
L. Inferior 95%	842,8	61,0	55,5	89,2	53,7
L. Superior 95%	849,8	63,8	58,4	90,9	56,6
CV(%)	0,2	1,1	1,3	0,5	1,3

Fuente: EUSTAT. Encuesta de la Sociedad de la Información - ESIF.

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 de la C.A. de Euskadi con ordenador en el hogar es del 2,2 % (1,96*1,1). O lo que es lo mismo, a un nivel de confianza del 95% podemos afirmar que el verdadero valor del porcentaje de familias con ordenador en la C.A. de Euskadi oscila en un intervalo del ±2,2% de la estimación dada. Es decir,

$$(62{,}4 \pm 0{,}02156*62{,}4) = \text{entre } 61{,}0\% \text{ y } 63{,}8\%$$

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 (2005), *"Encuesta sobre la Sociedad de la Información-ESI-Familias. Ficha metodológica."*. http://www.eustat.es/document/esi_c.html

[2] EUSTAT (2005), *"Encuesta de Población en Relación con la Actividad. Nota metodológica.2005."* http://www.eustat.es/document/datos/notamet_nuevaPRA_c.pdf

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

[4] Sas Institute Inc. (2004), *"SAS/STAT® 9.1 Guia 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.