Informe sobre el Cálculo de Errores de Muestreo

Encuesta sobre la conciliación de la vida laboral, familiar y personal (CVL)

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 sobre la Conciliación de la Vida Laboral, Familiar y Personal (CVL).

2. Método de expansión de Taylor [3], [5].

Este método permite calcular estimaciones del error muestral para totales, medias y ratios 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:

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

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

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 en la operación que nos ocupa.

3. Cálculo de errores.

3.1 Diseño Muestral [1]

La Encuesta sobre la conciliación de la vida laboral, familiar y personal (CVL) es una encuesta por muestreo sobre la población residente en la C.A. de Euskadi.

Se toma como base del muestreo las viviendas familiares seleccionadas para la Encuesta de la población en relación con la actividad (PRA) en el mismo trimestre de referencia. Se seleccionan todas las personas ocupadas de 16 y más años que residen en estas viviendas. El conjunto de

individuos seleccionados constituye la muestra de la Encuesta sobre la conciliación de la vida laboral, familiar y personal, cuya entrevista se realiza mediante el cuestionario específico adjunto.

El cuestionario no contiene las características sociodemográficas de las personas encuestadas, dado que esta información se encuentra recogida ya en la base de datos de la PRA -referida al mismo período- y se integra previamente a la explotación de los datos.

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 serán los siguientes:

CLUSTER = NUMC.
STRATA = Territorio histórico.
CLASS= Variables por columna (categóricas).
WEIGHT = Elevador de personas.
VAR = Variables por columna numéricas y categóricas.
DOMAIN = Cruces por variables socio-demográficas y económicas. (Ver apartado 3.3)

3.3 Estadísticos y dominios para el cálculo de errores.

Siguiendo el criterio adoptado por otras encuestas de Eustat para la publicación de los errores muestrales, se difundirán una relación de tablas de errores que se intentará ajustar en la medida de lo posible a la que se publica 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 sobre la conciliación de la vida laboral, familiar y personal. CVL

- Población ocupada de la C.A. de Euskadi por horas diarias dedicadas a actividades del trabajo doméstico, según características sociodemográficas (Media). Errores de muestreo.

- Población ocupada de la C.A. de Euskadi por grado de dificultad para solicitar permisos, según características sociodemográficas (Media). Errores de muestreo.

- Población ocupada de la C.A. de Euskadi por grado de satisfacción con el tiempo dedicado a aspectos de conciliación, según características sociodemográficas (Media) (1). Errores de muestreo.

- Población ocupada de la C.A. de Euskadi por grado de dificultad para compaginar aspectos de conciliación, según características sociodemográficas (Media). Errores de muestreo.

- Población ocupada de la C.A. de Euskadi por perjuicios relacionados con otros aspectos relativos a la conciliación (maternidad, paternidad y excedencia), según características sociodemográficas (Media). Errores de muestreo.

- Población ocupada de la C.A. de Euskadi por grado de satisfacción personal, según características sociodemográficas (Media) (1). Errores de muestreo.

- Población ocupada de la C.A. de Euskadi por preferencias laborales, según características sociodemográficas (%). Errores de muestreo.

- Población ocupada de la C.A. de Euskadi por grado de satisfacción con el trabajo, según características sociodemográficas (Media) (1). Errores de muestreo.

- Población ocupada de la C.A. de Euskadi por utilidad y satisfacción con la formación recibida de la empresa, según características sociodemográficas (Media). Errores de muestreo.

- Hijos/as menores de 15 años de la C.A. de Euskadi por cuidado diario durante la jornada laboral, según características sociodemográficas de los menores. Miles y (%). Errores de muestreo.

- Hijos/as menores de 15 años de la C.A. de Euskadi por cuidado diario fuera de la jornada laboral, según características sociodemográficas de los menores. Miles y (%). Errores de muestreo.

- Personas dependientes de la población ocupada de la C.A. de Euskadi por cuidado fuera de la jornada laboral, según características sociodemográficas de las personas dependientes. Miles y (%). Errores de muestreo.

(1) Se ha utilizado una escala de 0 a 10 en la que 0 significa ninguna y 10 máxima

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 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).

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:

Población ocupada de la C.A. de Euskadi por satisfacción con el trabajo, según características sociodemográficas (Media) (1). Errores de muestreo. 2011

Fuente: Encuesta sobre la conciliación de la vida laboral, familiar y personal (CVL)

	Jornada laboral	Flexibilidad de horarios	Descanso durante jornada laboral	Vacaciones y permisos	Estabilidad	Remuneración salarial	Promoción	Satisfacción general
C.A. de Euskadi								
Estimación	6,9	5,9	6,1	7,0	6,7	6,0	3,4	6,6
L. Inferior 95%	6,8	5,8	6,0	6,9	6,6	6,0	3,3	6,6
L. Superior 95%	7,0	6,0	6,2	7,0	6,8	6,1	3,5	6,7
CV(%)	**0,5**	**0,8**	**0,7**	**0,6**	**0,6**	**0,6**	**1,6**	**0,5**

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% de la población ocupada de la C.A.E. por satisfacción con respecto a la jornada laboral es del 0,98 % (1,96*0,5). O lo que es lo mismo, a un nivel de confianza del 95% podemos afirmar que el verdadero valor oscila en un intervalo del ±0,98 % de la estimación dada:

$$(6{,}9 \pm 0{,}0098*6{,}9) = (6{,}83\ ,\ 6{,}97)$$

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 (2010), *"Encuesta de conciliación de la vida laboral, familiar y personal. Ficha metodológica"*. http://www.eustat.es/document/conc_vida_laboral_c.asp#

[2] Eustat (2005), *"Encuesta de población en relación con la actividad. Ficha metodológica."*. http://www.eustat.es/document/poblact_c.html

[3] Fuller, W. A. (1975), *"Regression Analysis for Sample Survey,"* Sankhyā , 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.