Informe sobre el Cálculo de Errores de Muestreo

Encuesta de Medio Ambiente - Familias
EMAF 2020

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:

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

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 [2], [4].

En lo que sigue introduciremos estos métodos y su aplicación concreta en el caso de la Encuesta de Medio Ambiente - Familias (en adelante EMAF).

2. Método de expansión de Taylor.

Este método [4] 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:

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

Donde:

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

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

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 [3], 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].

Marco muestral: Se trata de una muestra en dos etapas estratificada por zonas.
Además la muestra está equilibrada por características de sus ocupantes en los Territorios Históricos y por el número de personas en las zonas.

Las principales características de su diseño muestral son las siguientes:

Tamaño muestral: 7.000

Marco muestral: Registro de Población y Directorio de Viviendas de Eustat

Diseño muestral: Muestreo aleatorio estratificado

Variables de estratificación:
- Territorio Histórico
- Zonas

Etapas del muestreo: Bietápico. En una primera etapa se seleccionan viviendas y en una segunda las personas de las viviendas que van a responder al cuestionario individual.

Extracción: Aleatoria equilibrada

3.2 Procedimiento de cálculo.

La sintaxis básica del procedimiento de SAS implementado para el cálculo de errores de esta encuesta es la siguiente [3]:

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 utilizados para el caso concreto de la EMAF serán los siguientes:

STRATA = Variable estrato formada por la zona geográfica y las tipologías de sección.
CLUSTER = Variable que identifica a la Unidad Primaria del Muestreo. En este caso será la variable sección censal.
DOMAIN = Variables de clasificación sociodemográfica.
VAR = Variables cuantitativas y cualitativas de medio ambiente familiar.
WEIGHT = Elevador de vivienda o de persona según estimaciones calculadas.

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

Se difunden tablas de coeficientes de variación para todas las estimaciones (porcentajes, medias, índices, etc.) publicadas en el apartado de tablas estadísticas de la Web para esta operación. Las tablas de errores publicadas son:

Tablas de coeficientes de variación para viviendas por las características sociodemográficas de la persona de referencia.

- Viviendas de la C.A. de Euskadi por hábitos y dispositivos de ahorro de agua (%) . Coeficientes de variación.

- Viviendas de la C.A. de Euskadi por tipo de energía utilizada (%). Coeficientes de variación.

- Viviendas de la C.A.de Euskadi por tipo de calefacción utilizada.(%). Coeficientes de variación.
- Viviendas de la C.A.de Euskadi por tipo de aislamiento e iluminación utilizada (%).Coeficientes de variación.
- Viviendas de la C.A.de Euskadi por los grados de temperatura diurna.(%). Coeficientes de variación.
- Viviendas de la C.A.de Euskadi por el tratamiento dado a sus residuos (%). Coeficientes de variación.
- Viviendas de la C.A.de Euskadi por grandes electrodomésticos (%).Coeficientes de variación.
- Viviendas de la C.A.de Euskadi por equipamiento audiovisual (%).Coeficientes de variación.
- Viviendas de la C.A.de Euskadi por pequeños electrodomésticos (%).Coeficientes de variación.
- Viviendas de la C.A.de Euskadi con problemas de ruidos y medidas tomadas (%).Coeficientes de variación.
- Viviendas de la C.A.de Euskadi con problemas de malos olores y medidas tomadas (%).Coeficientes de variación.
- Viviendas de la C.A.de Euskadi por número de vehículos para uso personal (%).Coeficientes de variación.
- Viviendas de la C.A.de Euskadi por el uso de ciertos productos (%).Coeficientes de variación.
- Viviendas de la C.A.de Euskadi por la importancia de ciertos factores al comprar (%).Coeficientes de variación.

Tablas de coeficientes de variación para personas de 16 y más años por características sociodemográficas.

- Personas de 16 y más años de la C.A. de Euskadi por tipo de transporte que utilizan según el sexo y por grupos de edad (%). Coeficientes de variación.
- Personas de 16 y más años de la C.A. de Euskadi, por opiniones y actitudes medioambientales, según sexo y por grupos de edad (%). Coeficientes de variación.
- Personas de 16 y más años de la C.A. de Euskadi por medio de transporte utilizado (%).Coeficientes de variación.
- Personas de 16 y más años de la C.A. de Euskadi que usan transporte público (%). Coeficientes de variación.
- Personas de 16 y más años de la C.A. de Euskadi que van caminando o en bici (%).Coeficientes de variación.
- Personas de 16 y más años de la C.A. de Euskadi por opiniones medioambientales (%).Coeficientes de variación.

- Personas de 16 y más años de la C.A. de Euskadi con actividades medioambientales (%).Coeficientes de variación.

- Personas de 16 y más años de la C.A. de Euskadi favorables a medidas medioambientales (%).Coeficientes de variación.

Tablas de coeficientes de variación para indicadores de medio ambiente por características sociodemográficas de la persona de referencia.

- Indicadores de medio ambiente de vivinedas de la C.A. de Euskadi (%). Coeficientes de variación.

- Viviendas de la C.A. de Euskadi por nivel de indicadores de medio ambiente (%).Coeficientes de variación.

3.4 Resultados e Interpretación.

A partir del coeficiente de variación, se pueden calcular otras medidas del error que son de utilidad y ayudan a la interpretación del mismo. Entre éstas, las más interesantes son:

- **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:

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

Siendo $\hat{\theta}$ el valor del estadístico de interés (media, total, proporción,..).

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

- Error relativo al 95% de confianza: 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.

Por ejemplo, si el porcentaje de familias en la C.A. de Euskadi que reciclan papel y cartón es del 92,3% con un coeficiente de variación del 0,5%, el correspondiente error relativo al 95% será del 0,98% (es decir, 1,96 x 0,5). O lo que es lo mismo, a un nivel de confianza del 95% podemos afirmar que el

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

verdadero valor del porcentaje de familias en la C.A. de Euskadi que reciclan papel y cartón oscila dentro de un intervalo del ±0,98 % de la estimación dada. Es decir:

$$[92,3 \pm (0,0098 \times 92,3)] = [91,39\%,\ 93,2\%]$$

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. *"Encuesta de Medio Ambiente - Familias. Ficha metodológica."* http://www.eustat.es/document/EMAF2015%5Fc.asp

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

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

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