EUSKAL ESTATISTIKA ERAKUNDA
INSTITUTO VASCO DE ESTADÍSTICA

Informe sobre el Cálculo de Errores de Muestreo

Encuesta de Bienestar Personal (EBP)

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:

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

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 Bienestar Personal (en adelante EBP).

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:

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

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

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 [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. EBP.

3.1 Diseño Muestral [1].

El diseño muestral es el correspondiente a las encuestas de origen: Encuesta de condiciones de vida (ECV), Encuesta de capital social (ECS), Encuesta de presupuestos de tiempo (EPT), Encuesta de medio ambiente a familias (EMAF), etc. alternando sucesivamente.
El universo poblacional es la población de 16 y más años de edad que reside en viviendas familiares de la C. A. de Euskadi, población objeto del total de encuestas origen, excepto en el caso de la ECS, que contempla también a personas residentes en establecimientos colectivos.

Este diseño muestral se adapta perfectamente a las especificaciones del método heurístico descrito 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 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 EBP serán los siguientes:

CLASS = Variables de bienestar personal categorizadas (categorías: alto, medio, bajo).
VAR = Variables cuantitativas de bienestar personal (medias).
STRATA = Territorio Histórico y Tamaño de municipio.
DOMAIN = Variables de clasificación sociodemográfica. (Ver apartado 3.3)
WEIGHT = Elevador de persona.

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

Se difunden tablas de coeficientes de variación para todas las estimaciones (porcentajes y medias) publicadas en el apartado de tablas estadísticas de la Web Estas tablas son:
Tablas de coeficientes de variación para porcentajes y medias según características sociodemográficas

- Población de 16 y más años de la C.A. de Euskadi por indicadores de satisfacción y valor de la vida y satisfacción con el tiempo de ocio según características sociodemográficas (% y media). Coeficientes de variación.

- Población de 16 y más años de la C.A. de Euskadi por indicadores de satisfacción con la vivienda y el entorno y con la economía doméstica según características sociodemográficas (% y media). Coeficientes de variación.

- Población de 16 y más años de la C.A. de Euskadi por indicadores de estado de ánimo y relaciones personales según características sociodemográficas (% y media). Coeficientes de variación.

- Población de 16 y más años de la C.A. de Euskadi por indicadores de confianza en las personas y en los poderes públicos según características sociodemográficas (% y media). Coeficientes de variación.

3.4 Resultados e Interpretación.

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:

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

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

Otra forma de interpretar esta información es calcular **el 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.

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:

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

A continuación se presenta un modelo de tabla de difusión de errores:

Población de 16 y más años de la C.A. de Euskadi por indicadores de satisfacción y valor de la vida y satisfacción con el tiempo de ocio según características sociodemográficas (% y media) (*). Coeficientes de variación. 2014

	Satisfacción con la vida				Valor de la vida				Satisfacción con el tiempo de ocio			
	Bajo	Medio	Alto	MEDIA	Bajo	Medio	Alto	MEDIA	Bajo	Medio	Alto	MEDIA
TOTAL	5,9	1,6	1,7	0,4	7,7	1,9	1,4	0,3	3,6	1,6	2,1	0,5
Territorio de residencia												
Araba/Álava	5,9	1,6	1,7	0,7	16,2	3,3	2,6	0,6	7,1	3,0	3,7	1,0
Bizkaia	11,8	2,8	3,2	0,6	10,9	2,9	2,0	0,5	5,2	2,4	3,2	0,8
Gipuzkoa	8,1	2,4	2,6	0,6	13,0	3,1	2,4	0,6	6,3	2,8	3,4	0,9

Para la tabla anterior, el error relativo al 95% para la valoración de medio respecto a la satisfacción con la vida de la población de 16 y más años en Euskadi es del 3,1 % (es decir, 1,96 x 1,6). O lo que es lo mismo, a un nivel de confianza del 95% podemos afirmar que la valoración de medio respecto a la satisfacción con la vida oscila en un intervalo del ±3,1 % de la estimación dada. Es decir:

[47,1 ± (0,031 x 47,1)] = [45,64, 48,56]

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

[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.

(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 Bienestar Personal"*
http://es.eustat.eus/document/ebp_c.html#axzz3uUPxBwpe

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