

Informe sobre el Cálculo de Errores de Muestreo

Encuesta sobre Capital Social (ECS)



INDICE

1. Introducción.....	3
2. Método de expansión de Taylor.....	3
3. Cálculo de errores. ECS.....	4
3.1 Diseño Muestral.....	4
3.2 Procedimiento de cálculo.....	5
3.3 Estadísticos y dominios para el cálculo de errores en la ECS.....	5
3.4 Resultados e Interpretación.....	7
Bibliografía.....	8

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:

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

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 sobre Capital Social (en adelante ECS).

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:

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

Donde:

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

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

y

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

Notación:

$h = 1, 2, \dots, H$ indica el estrato con un total de H estratos.

$i = 1, 2, \dots, n_h$ indica el número de clusters en el estrato h , con un total de n_h clusters.

$j = 1, 2, \dots, m_{hi}$ indica el número de unidad dentro del cluster i del estrato h , con un total de m_{hi} unidades

$n = \sum_{h=1}^H \sum_{i=1}^{n_h} m_{hi}$ 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

$y_{hij} = (y_{hij}(1), y_{hij}(2), \dots, y_{hij}(P))$ 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. ECS.

3.1 Diseño Muestral [1].

Se trata de una muestra estratificada de una sola etapa y equilibrada tanto por sexo y edad en los Territorios Históricos como por número de individuos en las comarcas de la C.A.E.

Las principales características de su diseño muestral son las siguientes:

Unidad de muestreo: Población de 15 y más años residentes en viviendas y establecimientos colectivos.

Tamaño muestral real: 7.000

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

Diseño muestral: Muestreo aleatorio estratificado

VARIABLES DE ESTRATIFICACIÓN:

- Territorio Histórico
- Tamaño de municipio: capitales (mayores de 100.000 habitantes), medianos (de 100.000 a 20.000) y pequeños (20.000 y menos)
- Nacionalidad (española o extranjera)

Etapas del muestreo: Unietápico con selección directa del individuo

Extracción: Aleatoria equilibrada

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 estadificado*/  
  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 ECS serán los siguientes:

CLASS = Variables de capital social categorizadas (categorías: alto, medio, bajo).
VAR = Variables cuantitativas de capital social (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 ECS

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 15 y más años de la C.A. de Euskadi por indicadores de tamaño de red amplia y próxima y relaciones personales en la red según características sociodemográficas (%). Coeficientes de variación.
- Población de 15 y más años de la C.A. de Euskadi por indicadores de homogeneidad de la red de amigos, ayuda prestada y recibida según características sociodemográficas (%). Coeficientes de variación.
- Población de 15 y más años de la C.A. de Euskadi por indicadores de acceso a ayuda financiera, por problemas de salud y emocional según características sociodemográficas (%). Coeficientes de variación.
- Población de 15 y más años de la C.A. de Euskadi por indicadores de confianza general, en profesiones y en redes según características sociodemográficas (%). Coeficientes de variación.
- Población de 15 y más años de la C.A. de Euskadi por indicadores de reciprocidad y confianza en instituciones según características sociodemográficas (%). Coeficientes de variación.
- Población de 15 y más años de la C.A. de Euskadi por indicadores de interés en temas socio-políticos y participación electoral según características sociodemográficas (%). Coeficientes de variación.
- Población de 15 y más años de la C.A. de Euskadi por indicadores de cooperación, influencia e independencia personales según características sociodemográficas (%). Coeficientes de variación.
- Población de 15 y más años de la C.A. de Euskadi por indicadores de participación social y política y en asociaciones según características sociodemográficas (%). Coeficientes de variación.
- Población de 15 y más años de la C.A. de Euskadi por indicadores de información en temas socio-políticos y uso de Internet y redes virtuales según características sociodemográficas (%). Coeficientes de variación.
- Población de 15 y más años de la C.A. de Euskadi por indicadores de acceso a medios de comunicación y variedad de fuentes de información según características sociodemográficas (%). Coeficientes de variación.
- Población de 15 y más años de la C.A. de Euskadi por indicadores de cohesión social, seguridad y corrupción según características sociodemográficas (%). Coeficientes de variación.
- Población de 15 y más años de la C.A. de Euskadi por indicadores de felicidad y salud según características sociodemográficas (%). 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:

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

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 Normal¹ para los estadísticos más comunes, por lo que la construcción de este intervalo vendrá dada por la siguiente expresión:

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

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

Población de 15 y más años de la C.A. de Euskadi por indicadores de tamaño de red amplia y próxima y relaciones personales en la red según TH. (%). Coeficientes de variación. 2012.

Territorio de residencia	Tamaño de red amplia				Tamaño de red próxima				Relaciones personales en la red			
	Bajo	Medio	Alto	Media	Bajo	Medio	Alto	Media	Bajo	Medio	Alto	Media
C.A. de Euskadi	2,9	1,6	3,0	0,7	2,1	1,7	4,1	1,0	1,8	1,8	5,4	0,7
Alava	5,4	3,2	6,4	1,5	4,0	3,5	8,7	2,2	3,8	3,5	10,6	1,3
Bizkaia	4,2	2,3	4,4	1,1	3,2	2,5	6,0	1,5	2,6	2,8	7,8	1,0
Gipuzkoa	5,1	2,6	4,9	1,2	3,5	2,8	6,8	1,7	2,9	3,1	9,6	1,1

Para la tabla anterior, el error relativo al 95% para el tamaño medio de red amplia en la C.A. de Euskadi es del 3,1 % (es decir, $1,96 \times 1,6$). O lo que es lo mismo, a un nivel de confianza del 95% podemos afirmar que el verdadero valor del tamaño medio de red amplia en la C. A. de Euskadi oscila en un intervalo del $\pm 3,1$ % de la estimación dada. Es decir:

$$[52,3 \pm (0,031 \times 52,3)] = [50,66, 53,94]$$

Es importante señalar aquellas estimaciones que sobrepasen un determinado porcentaje del error relativo al 95%, para que el usuario tome las debidas precauciones 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.).

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.

Bibliografía

[1] EUSTAT. "Encuesta sobre Capital Social. Ficha metodológica."

<http://www.eustat.es/document/en%5Fcapital%5Fsocial%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 Guía 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.