Informe sobre el Cálculo de Errores de Muestreo

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

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 en las Empresas desde el periodo 2005.

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:

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

Donde:

$$e_{hi\cdot} = \left(\sum_{j=1}^{m_{hi}} w_{hij} \left(y_{hij} - \hat{\bar{Y}} \right) \right) / w_{\cdots}$$

$$e_{h\cdot\cdot} = \left(\sum_{i=1}^{n_h} e_{hi\cdot} \right) / n_h$$

y

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

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
$n = \sum_{h=1}^{H} \sum_{i=1}^{n_h} m_{hi}$ 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
$y_{hij} = (y_{hij}(1), y_{hij}(2), \ldots, y_{hij}(P))$ 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 - Empresas.

3.1 Diseño Muestral [1]

La ESI -Empresas- es una encuesta constituida por una muestra probabilística continua, es decir, un panel sobre los establecimientos económicos de la C.A. de Euskadi de todos los sectores de actividad, salvo el sector primario y el servicio doméstico. Este panel es censal en los establecimientos de 100 ó más empleados y muestral, en los menores de 100 empleados.

La estratificación se realiza por Territorio Histórico, por tamaño del establecimiento (agrupado en 5 modalidades) y por rama de actividad, clasificada según la sectorización normalizada A38, que

corresponde a agrupaciones de actividades de la Clasificación Nacional de Actividades Económicas [2]. La afijación es proporcional a la raíz cuadrada del tamaño de cada estrato, excepto en aquellos sectores de actividad que están infra-representados por este reparto, en los cuáles se utiliza la afijación proporcional al tamaño. Se completa una muestra cercana a la muestra teórica de unos 7.200 establecimientos.

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 - Empresas serán los siguientes:

STRATA = Territorio Histórico x Actividad (A31) x Estrato de empleo
WEIGHT= Elevador anual de establecimientos.
VAR = Variables de equipamiento y uso de las Tecnologías de la Información, Internet y comercio electrónico.
DOMAIN = Cruces por variables geográficas, actividad, estrato de empleo y titularidad de la empresa.

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

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

Equipamientos TIC en empresas y establecimientos

- Equipamientos TIC en los establecimientos de la C.A. de Euskadi por territorio histórico, sector de actividad y estrato de empleo (%). Errores de muestreo.
- Equipamientos TIC en los establecimientos de 10 y más empleados de la C.A. de Euskadi por territorio histórico y sector de actividad (%). Errores de muestreo.

- Equipamientos de redes e intercambios en los establecimientos de la C. A. de Euskadi por territorio histórico, sector de actividad y estrato de empleo (%). Errores de muestreo.
- Equipamientos de redes e intercambios electrónicos de 10 y más empleados de la C.A. de Euskadi por territorio histórico y sector de actividad (%). Errores de muestreo.
- Medidas de seguridad informática en los establecimientos de la C.A. de Euskadi según estrato de empleo y sector de actividad (%). Errores de muestreo.
- Establecimientos de la C.A. de Euskadi que disponen de sistemas de gestión de información por estrato de empleo y sector de actividad (%). Errores de muestreo.

Empresas y establecimientos usuarios de Internet

- Tipo de conexión en los establecimientos con acceso a Internet de la C.A. de Euskadi por territorio histórico, sector de actividad y estrato de empleo (%). Errores de muestreo.
- Tipo de conexión en los establecimientos de 10 y más empleados con acceso a Internet de la C. A. de Euskadi por territorio histórico y sector de actividad (%). Errores de muestreo.
- Prestaciones de la web en los establecimientos con sitio web de la C. A. de Euskadi según estrato de empleo y sectores de actividad (%). Errores de muestreo.
- Establecimientos con sitio web de la C.A. de Euskadi según prestaciones ofertadas (%).Errores de muestreo.
- Idiomas disponibles en la web en los establecimientos con sitio web de la C. A. de Euskadi por territorio histórico, sectores de actividad y estrato de empleo (%). Errores de muestreo.
- Idiomas disponibles en la web en los establecimientos de 10 y más empleados con sitio web de la C. A. de Euskadi por territorio histórico y sector de actividad (%). Errores de muestreo.
- Trámites electrónicos con la Administración Pública en los establecimientos con acceso a internet de la C. A. de Euskadi según estrato de empleo y sector de actividad (%). Errores de muestreo.

Comercio electrónico

- Comercio electrónico en los establecimientos de la C. A. de Euskadi según territorio histórico, por estrato de empleo (%). Errores de muestreo.
- Compras y ventas por comercio electrónico en la C.A. de Euskadi según territorio histórico, sector de actividad y estrato de empleo (millones €). Errores de muestreo.

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).

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:

Establecimientos de 10 y más empleados y empleo de la C.A. de Euskadi por equipamientos de tecnologías de la información y territorio histórico.Errores de muestreo. 2012

	Ordenador personal		Correo electrónico		Internet		Tfno. móvil	Página web
	% s/establ.	% s/empleo	% s/establ.	% s/empleo	% s/establ.	% s/empleo	% s/establ.	% s/establ.
Total	98.6	65.8	97.2	58.0	96.7	55.3	94.5	81.6
L. Inferior 95%	98.0	64.1	96.3	56.2	95.8	53.3	93.3	79.4
L. Superior 95%	99.3	67.5	98.1	59.8	97.7	57.2	95.6	83.7
CV(%)	**0.3**	**1.3**	**0.4**	**1.6**	**0.5**	**1.8**	**0.6**	**1.3**
Territorio Histórico								
Araba/Alava	97.6	66.4	95.6	55.6	95.8	51.5	93.9	81.0
L. Inferior 95%	95.2	62.5	92.9	51.9	93.2	47.0	91.6	76.7
L. Superior 95%	99.9	70.4	98.2	59.2	98.4	56.0	96.1	85.4
CV(%)	**1.2**	**3.0**	**1.4**	**3.3**	**1.4**	**4.5**	**1.2**	**2.7**
Bizkaia	98.9	67.3	97.3	60.0	96.6	57.7	94.8	81.5
L. Inferior 95%	98.1	64.8	96.0	57.3	95.1	54.9	93.1	78.1
L. Superior 95%	99.7	69.7	98.6	62.7	98.1	60.5	96.5	84.6
CV(%)	**0.4**	**1.9**	**0.7**	**2.3**	**0.8**	**2.5**	**0.9**	**2.1**
Gipuzkoa	98.7	63.2	97.8	56.1	97.4	53.4	94.2	82.2
L. Inferior 95%	97.8	60.1	96.7	53.0	96.2	50.3	92.4	78.6
L. Superior 95%	99.6	66.3	98.9	59.2	98.6	56.6	96.0	85.7
CV(%)	**0.5**	**2.5**	**0.6**	**2.8**	**0.6**	**3.0**	**1.0**	**2.2**

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.

[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.

Para la tabla anterior, el error relativo al 95% para el porcentaje de establecimientos con ordenador en la C.A. de Euskadi es del 0,588 % (1,96*0,3). O lo que es lo mismo, a un nivel de confianza del 95% podemos afirmar que el verdadero valor del porcentaje de establecimientos con ordenador en la C.A. de Euskadi oscila en un intervalo del ±0,588 % de la estimación dada. Es decir:

$$(98{,}6 \pm 0{,}00588*98{,}6) = \text{entre } 98{,}0\ \%\ \text{y}\ 99{,}3\ \%$$

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 al 95% (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 (2006), *"Encuesta sobre la Sociedad de la Información - ESI-Empresas. Ficha metodológica."*. http://www.eustat.es/document/esie_c.html

[2] Clasificación Nacional de Actividades Económicas (CNAE 1993) – Rev1 http://www.eustat.es/document/datos/CNAE93REV1.xls

[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 Guía 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.