# *Informe sobre el Cálculo de Errores de Muestreo*

Encuesta de Población en Relación con la Actividad (PRA)

# 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].

A continuación se introducen estos métodos y su aplicación concreta en el caso de la Encuesta de la Población en Relación con la Actividad (en adelante PRA) desde el periodo 2005.

## *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:

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

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
$n = \sum_{h=1}^{H} \sum_{i=1}^{n_h} m_{hi}$ 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 estadística que nos ocupa.

## *3. Cálculo de errores PRA*

### *3.1 Diseño Muestral [1]*

En 2005, la muestra de la PRA sufre un cambio que supone, no sólo un incremento de unidades con respecto a periodos anteriores, sino una menor complejidad en cuanto a diseño [2]. La actual muestra se basa en una muestra probabilística continua, es decir, un panel de viviendas que se va renovando continuamente. La muestra tiene un tamaño aproximado de 5.000 viviendas al trimestre (un total aproximado de 13.500 individuos) y una rotación de un octavo de un trimestre a otro, de modo que cada vivienda permanece en la muestra durante dos años.

La muestra de viviendas se extrae aleatoriamente del Directorio de Viviendas de modo estratificado a nivel de Territorio Histórico. Dentro de cada estrato, se muestrean viviendas de forma sistemática

(con la misma probabilidad). A través de un informante se recoge la información de todos los individuos de la vivienda por lo que, a efectos de diseño, las viviendas se comportan como pequeños conglomerados. Las unidades seleccionadas son viviendas y los resultados se refieren a los individuos.
Este diseño 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 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 nueva PRA serán los siguientes:

CLUSTER = Identificador de vivienda.
STRATA = Territorio Histórico.
WEIGHT = Elevador trimestral de personas /Elevador trimestral de familias
Elevador anual de personas /Elevador anual de familias.
RATIO = Tasas de paro, actividad y ocupación.
VAR = Totales de población parada, ocupada, activa,…
DOMAIN = Cruces por variables socio-demográficas y económicas. (Ver apartado 3.3)

## *3.3 Estadísticos y dominios para el cálculo de errores en la PRA*

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

**Trimestrales**

- Tasa de actividad y paro de la población de 16 y más años según el territorio histórico (%).
- Tasa de actividad de la población de 16 y más años según el sexo y la edad (%).
- Tasa de paro de la población de 16 y más años según el sexo y la edad (%).
- Tasa de ocupación de la población de 16 a 64 años según el sexo, la edad y el territorio histórico (%).

- Población de 16 y más años según el territorio histórico y el sexo (en miles).
- Población de 16 y más años activa según el territorio histórico y el sexo (en miles).
- Población de 16 y más años ocupada según el territorio histórico y el sexo (en miles).
- Población de 16 y más años ocupada por el territorio histórico según el sector económico (en miles).
- Población de 16 y más años parada según el territorio histórico y el sexo (en miles).

**Anuales**

- Población de 16 y más años por su relación con la actividad según el territorio histórico y el sexo (en miles).
- Población de 16 y más años ocupada por situación profesional y tipo de contrato (en miles).
- Familias por la relación con la actividad de sus miembros según el territorio histórico (en miles).

Podemos resumir lo anterior en las siguientes tablas según estadístico y variable de cruce:

**Errores trimestrales**

| Estadístico\Variable de cruce | Territorio Histórico | Sexo | Edad (3 grupos) | Sector económico |
|---|---|---|---|---|
| Tasa de actividad | X | X | X | |
| Tasa de paro | X | X | X | |
| Tasa de ocupación | X | X | X | |
| Población activa | X | X | | |
| Población ocupada | X | X | | X |
| Población parada | X | X | | |
| Población de 16 y más años | X | X | | |

**Nota:** Los cruces sombreados se dan de forma simultánea

**Errores anuales**

| Estadístico\Variable de cruce | Territorio histórico | Sexo | Relación con la actividad | Situación profesional | Tipo de contrato |
|---|---|---|---|---|---|
| Población de 16 y más años | X | X | X | | |
| Población ocupada | | | | X | X |
| Familias | X | | X | | |

**Nota:** Los cruces sombreados se dan de forma simultánea.

## *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:

**T.1 Coeficientes de variación e intervalos de confianza para la tasa de actividad y paro de la población de 16 y más años según el territorio histórico (%). IV-2004**

Fuente: EUSTAT. Encuesta de Población en Relación con la Actividad.

| | C.A. de Euskadi | | Araba / Alava | | Bizkaia | | Gipuzkoa | |
|---|---|---|---|---|---|---|---|---|
| | Tasa de actividad | Tasa de paro | Tasa de actividad | Tasa de paro | Tasa de actividad | Tasa de paro | Tasa de actividad | Tasa de paro |
| Estimación | 55,0 | 7,0 | 57,7 | 5,1 | 53,3 | 7,7 | 56,5 | 6,7 |
| L. Inferior 95% | 53,8 | 6,2 | 55,1 | 3,7 | 51,6 | 6,4 | 54,7 | 5,5 |
| L. Superior 95% | 56,2 | 7,8 | 60,2 | 6,4 | 55,1 | 9,0 | 58,4 | 8,0 |
| CV(%) | **1,1** | **5,9** | **2,3** | **13,5** | **1,7** | **8,4** | **1,7** | **9,3** |

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 Tasa de Actividad de la C.A. de Euskadi es del 2,1% (1,96*1,1). O lo que es lo mismo, a un nivel de confianza del 95% podemos afirmar que el verdadero valor de la Tasa de Actividad de la C.A. de Euskadi oscila en un intervalo del ±2,1% de la estimación dada:

$$(55{,}0 \pm 0{,}021*55{,}0) = (53{,}8,\ 56{,}2)$$

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 (2005), *"Encuesta de Población en Relación con la Actividad. Ficha metodológica."* http://www.eustat.es/document/poblact_c.html

[2] EUSTAT (2005), *"Encuesta de Población en Relación con la Actividad. Nota metodológica.2005."* http://www.eustat.es/document/datos/notamet_nuevaPRA_c.pdf

[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. "User's Guide".* 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.