EL MÉTODO DEL CUBO:
APLICACIONES DEL MUESTREO EQUILIBRADO
EN LA ORGANIZACIÓN ESTADISTICA VASCA

Aritz Adin Urtasun

EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA

Donostia-San Sebastián, 1
01010 VITORIA-GASTEIZ
Tel.: 945 01 75 00
Fax.: 945 01 75 01
E-mail: eustat@eustat.es
www.eustat.es

Presentación

Eustat, consciente de la creciente demanda de estadísticas de calidad cada vez más desagregadas, organizó en 2010 el XXIII Seminario Internacional de Estadística con el título “Muestreo equilibrado y eficiente: el Método del Cubo”.

El objetivo de Eustat es redefinir los diseños actuales, para que con el mismo o similar coste se puedan obtener muestras que proporcionen estimadores de calidad para ámbitos o dominios mas desagregados. Con este mismo objetivo se convocó una beca de dos años de duración de formación e investigación en el campo de las metodologías estadístico-matemáticas, mas concretamente enfocada hacia la optimización de muestras.

Los resultados de esta investigación han sido aplicados en diferentes operaciones estadísticas dentro del Plan Vasco de Estadísticas 2010-2012: Estudio del “bullying” en el alumnado de centros de Educación Primaria y Educación Secundaria Obligatoria, Encuesta sobre la Sociedad de la Información Familias, Encuesta de Innovación Tecnológica, Encuesta de Pobreza y Desigualdades Sociales y Estudio de las Mujeres en el Ámbito Rural Vasco.

El objetivo de esta publicación es difundir la investigación realizada durante la beca y aportar material útil a todos los usuarios interesados en el conocimiento y utilización de muestreos eficientes y equilibrados.

Este documento tiene dos partes diferenciadas. En la primera, se encuentran los conceptos y definiciones correspondientes a la teoría de muestreo; así como los planes de muestreo probabilísticas simples y complejos. En la segunda, la descripción del Método del Cubo y su aplicación a diferentes encuestas-tipo de la Organización Estadística Vasca.

Vitoria-Gasteiz, Diciembre de 2012

Javier Forcada Sainz

Director General de EUSTAT

Índice

PRESENTACIÓN 1

ÍNDICE 2

1. INTRODUCCIÓN 4

2. INTRODUCCIÓN A LA TEORÍA DE MUESTREO 5

- DEFINICIONES Y NOTACIÓN BÁSICA 5
- PROPORCIONES MUESTRALES 6
- ESTIMADOR DE HORVITZ-THOMPSON 6

3. PLANES DE MUESTREO PROBABILÍSTICOS 7

- MUESTREO ALEATORIO SIMPLE 7
- MUESTREO ESTRATIFICADO 8
- MUESTREO POR CONGLOMERADOS O CLUSTERS 10
- RESUMEN DE LOS MÉTODOS PRESENTADOS 11

4. PLANES DE MUESTREO COMPLEJOS 13

- MUESTREO BIETÁPICO (O DE DOS ETAPAS) 13
- SELECCIÓN DE LAS UP-S CON PROBABILIDADES IGUALES 14
- PLAN BIETÁPICO AUTOPONDERADO 15

5. MÉTODO DEL CUBO: MUESTRE EQUILIBRADO 16

- REPRESENTACIÓN POR UN CUBO 16
- MUESTRAS EQUILIBRADAS 16
- DESCRIPCIÓN DEL MÉTODO 18

6. MACROS DE SAS PARA SELECCIONAR MUESTRAS EQUILIBRADAS 19

- MACRO *EXE_CUBE* 19
- MACRO *ECHANT_STRAT* 20
- MACRO AUXILIAR *DISJUNCTIVE* 21
- MACRO AUXILIAR *CREAR_ESTRATO* 21
- EJEMPLO DE USO DE LAS MACROS 22

7. MUESTRAS EQUILIBRADAS EN EUSTAT CON EL MÉTODO DEL CUBO 26

- MUESTRA DE CENTROS DE ESO PARA EL ESTUDIO DEL "BULLYING" EN LA COMUNIDAD AUTÓNOMA DE EUSKADI 26
- MUESTRA PARA LA ENCUESTA DE LA SOCIEDAD DE LA INFORMACIÓN (ESI-EMPRESAS) 30
- MUESTRA PARA LA ENCUESTA DE CAPITAL SOCIAL (ECS) 33

MUESTRA PARA LA ENCUESTA DE INNOVACIÓN TECNOLÓGICA (EIT) 38
MUESTRA PARA LA ENCUESTA DE POBREZA Y DESIGUALDADES SOCIALES (EPDS) 42
MUESTRA PARA EL ESTUDIO DE LAS MUJERES EN EL ÁMBITO RURAL VASCO 47
MUESTRA PARA LA ENCUESTA DE EUSKADI Y DROGAS 52

8. CONCLUSIONES 56

EQUILIBRIO Y ESTRATIFICACIÓN 56
ELECCIÓN DE LAS VARIABLES DE EQUILIBRIO 56
EQUILIBRIO Y CALIBRACIÓN 57
Análisis de los resultados 57
1. Calibración de la encuesta de Euskadi y Drogas 2012 57
2. Calibración de la Encuesta de Capital Social 2012 58
INTERÉS DEL MUESTREO EQUILIBRADO 60

9. BIBLIOGRAFÍA 61

1. Introducción

El contenido recogido en este Cuaderno Técnico, es fruto del trabajo realizado durante el disfrute de la beca de formación e investigación en metodologías estadístico-matemáticas, para el tema de optimización de muestras, concedida en el año 2010 por el Instituto Vasco de Estadística / Euskal Estatistika Erakundea.

El presente documento está dividido en los siguientes capítulos:

En el primer capítulo se realiza una introducción y se mencionan los objetivos que han marcado la elaboración de este cuaderno técnico.

En segundo capítulo, se expone una introducción a la teoría de muestreo, con las definiciones y notación básica del diseño de muestreo, proporciones muestrales y definición del estimador de Horvitz-Thompson y su varianza.

En los siguientes dos capítulos, se desarrollan los conceptos de planes de muestreo probabilísticos y planes de muestreo complejos, presentando la mayoría de los métodos utilizados en la estadística oficial.

En el quinto capítulo, se aborda el concepto de muestreo equilibrado y se presenta el Método del Cubo para seleccionar muestras equilibradas.

El objetivo del sexto capítulo es detallar las macros de SAS que permiten seleccionar muestras equilibradas.

En el séptimo capítulo, se presentan las distintas muestras equilibradas en Eustat con el Método del Cubo.

Finalmente, se muestran algunas conclusiones relacionadas con el equilibrio, la estratificación y la calibración.

Quiero agradecer el apoyo a todos los componentes del Área de Metodología, Innovación e I+D y, en general, la amabilidad de todo el personal de Eustat.

PALABRAS CLAVE: Diseños muestrales, Probabilidades de inclusión, Estimador de Horvitz-Thompson, Muestras equilibradas, Método del Cubo, Variables de equilibrio, Estratificación, Calibración.

2. Introducción a la teoría de muestreo

Antes de poder presentar el Método del Cubo para seleccionar muestras equilibradas y de mostrar el interés del método, debemos empezar por una presentación general de la teoría de muestreo.

Definiciones y notación básica

El objetivo es estudiar una población finita $U = \{1,\ldots, N\}$ de tamaño N.

Definimos la variable de interés y que toma valores y_k, $k \in U$; cuyo total y media son:

$$Y = \sum_{k \in U} y_k \qquad \text{y} \qquad \bar{Y} = \frac{1}{N} \sum_{k \in U} y_k$$

Una muestra s es un subconjunto de la población $s \subset U$.

Un diseño muestral o plan de muestreo $p(s)$ es una distribución de probabilidad sobre todas las muestras posibles en donde $\sum_{s \subset U} p(s) = 1$.

La muestra aleatoria S toma el valor s con la probabilidad $\Pr(S = s) = p(s)$.

Definimos la probabilidad de inclusión, como la probabilidad de que la unidad k sea seleccionada en la muestra aleatoria S:

$$\pi_k = E(I_k) = \Pr(k \in S) = \sum_{k \in s} p(s) \qquad \text{donde} \qquad I_k = \begin{cases} 1 & \text{si} \quad k \in S \\ 0 & \text{si} \quad k \notin S \end{cases}$$

De igual modo, se define la probabilidad de inclusión de 2.orden como:

$$\pi_{kl} = E(I_k I_l) = \Pr(k \text{ y } l \in S) = \sum_{k,l \in s} p(s)$$

Si el diseño muestral es de tamaño fijo, entonces $\sum_{k \in U} \pi_k = n$.

Proporciones muestrales

Supongamos que la variable de interés definida sobre la población *U* es una variable cualitativa. En este caso, la variable de interés nos da información acerca de alguna cualidad de las unidades de la población o la pertenencia o no a una determinada clase.

Supongamos que nuestra variable de interés clasifica las unidades de la población en dos clases C y C'.

Para cada unidad de la población, definimos la característica y_k como:

$$y_k = \begin{cases} 1 \text{ si } k \in C \\ 0 \text{ si } k \notin C \end{cases} \qquad \forall k \in U$$

Definimos el total de elementos de la población (total de la clase) y la proporción de elementos de la población (proporción de la clase) que pertenecen a C como:

$$Y = \sum_{k \in U} y_k = A \qquad \text{y} \qquad \bar{Y} = \frac{1}{N} \sum_{k \in U} y_k = \frac{A}{N} = P$$

Podemos considerar el problema de estimar A y P como si estimásemos el total y la media poblacional en donde cada y_k toma los valores 0 o 1.

Si escribimos la cuasivarianza S^2 en función de *P* y *Q* = 1-*P*

$$S^2 = \frac{\sum_{k \in U} (y_k - \bar{Y})^2}{N-1} = \frac{\sum_{k \in U} y_k^2 - N\bar{Y}^2}{N-1} = \frac{1}{N-1}(NP - NP^2) = \frac{N}{N-1} PQ$$

Cuyo estimador insesgado es:

$$s^2 = \frac{n}{n-1} pq \qquad \text{donde} \qquad p = \frac{\sum_{k \in S} y_k}{n} = \frac{a}{n}$$

Estimador de Horvitz-Thompson

Se definen el estimador de Horvitz-Thompson del total y de la media poblacional de la variable de interés y como:

$$\hat{Y}_\pi = \sum_{k \in S} \frac{y_k}{\pi_k} \qquad \text{y} \qquad \hat{\bar{Y}}_\pi = \frac{1}{N} \sum_{k \in S} \frac{y_k}{\pi_k}$$

El estimador de Horvitz-Thompson es insesgado si $\pi_k > 0$, $k \in U$.

Para diseños de tamaño fijo, se puede estimar la varianza por:

$$\hat{Var}(\hat{\bar{Y}}_\pi) = -\frac{1}{2} \sum_{k \in S} \sum_{\substack{l \in S \\ l \neq k}} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}}.$$

3. Planes de muestreo probabilísticos

Forman parte de este tipo de muestreo todos aquellos métodos para los que puede calcularse la probabilidad de extracción o selección de cualquiera de las muestras posibles.

Tal y como se explicará más adelante, el Método del Cubo parte de las probabilidades de inclusión definidas por el diseño para seleccionar una muestra equilibrada; es decir, en realidad este método optimiza los métodos de muestreo probabilísticos.

A continuación se definirán los tres principales tipos de muestreo probabilístico.

Muestreo aleatorio simple

El muestreo aleatorio simple (m.a.s.) es un método de muestreo en donde se selecciona una muestra de tamaño *n* de una población de tamaño *N* de tal manera que todas las muestras del mismo tamaño tienen la misma probabilidad de ser seleccionadas.

El diseño muestral para un m.a.s. de tamaño fijo *n* es:

$$p(s)=\begin{cases}\binom{N}{n}^{-1} & \text{si} \quad \text{card(s)} = \text{n} \\ 0 & \text{en caso contrario}\end{cases}$$

Por lo tanto, la probabilidad de inclusión de la unidad k es:

$$\pi_k=\sum_{k\in s} p(s)=\sum_{k\in s}\binom{N}{n}^{-1}=\binom{N-1}{n-1}\binom{N}{n}^{-1}=\frac{n}{N}\ , \quad \forall k\in U$$

Es decir, todos los individuos de *U* tienen la misma probabilidad de ser seleccionados.

El estimador de H-T para la media poblacional en un m.a.s. es

$$\hat{\bar{Y}}_\pi=\frac{1}{N}\sum_{k\in S}\frac{y_k}{\pi_k}=\frac{1}{N}\sum_{k\in S}y_k\frac{N}{n}=\frac{1}{n}\sum_{k\in S}y_k$$

El estimador insesgado de la varianza de $\hat{\bar{Y}}_\pi$ es:

$$\widehat{Var}(\hat{\bar{Y}}_\pi)=(1-f)\frac{s_y^2}{n}$$

donde

$$s_y^2=\frac{1}{n-1}\sum_{k\in S}(y_k-\hat{\bar{Y}}_\pi)^2$$

y $f=\frac{n}{N}$ es definida como la fracción de muestreo

Muestreo estratificado

Supongamos que la población *U* está dividida en subpoblaciones o estratos U_h, $h = 1,...,H$; donde los estratos cumplen las siguientes propiedades:

(i) $\bigcup_{h=1}^{H} U_h = U$

(ii) $U_h \cap U_i = \phi$, $h \neq i$

(iii) Si N_h es el tamaño de U_h, entonces $\sum_{h=1}^{H} N_h = N$

Un diseño muestral es estratificado si en cada estrato se selecciona una muestra aleatoria simple de tamaño fijo n_h, donde $\sum_{h=1}^{H} n_h = n$ es el tamaño de la muestra.

Esta técnica de muestreo se utiliza cuando la población de estudio es muy heterogénea, pudiendo dividirla en estratos internamente homogéneos. De esta manera, podemos lograr estimadores más precisos en cada estrato, combinándolos para obtener un estimador de la población total más preciso.

Como en cada estrato las unidades son seleccionadas siguiendo un m.a.s. la probabilidad de inclusión de la unidad *k* es:

$$\pi_k = \frac{n_h}{N_h}, \quad \forall k \in U .$$

El estimador de Horvitz-Thompson de la media para un muestreo estratificado:

$$\hat{\bar{Y}}_{st} = \frac{1}{N}\sum_{k \in S}\frac{y_k}{\pi_k} = \frac{1}{N}\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{k \in S_h} y_k = \frac{1}{N}\sum_{h=1}^{H} N_h \hat{\bar{Y}}_h$$

La varianza del estimador puede estimarse sin sesgo por:

$$\hat{Var}(\hat{\bar{Y}}_{st}) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 (1 - f_h)\frac{s_{yh}^2}{n_h}$$

donde $s_{yh}^2 = \frac{1}{n_h - 1}\sum_{h=1}^{H}(y_k - \hat{\bar{Y}}_h)^2$ es la cuasivarianza muestral del estrato *h*.

Afijaciones en muestreos estratificados

Existen distintos criterios a la hora de repartir el tamaño de la muestra entre los estratos. Vamos a presentar los más utilizados.

1. Afijación proporcional

Consiste en asignar a cada estrato un número de unidades muestrales proporcionales a su tamaño.

Por lo tanto, diremos que un plan estratificado tiene una afijación proporcional si:

$$\frac{n_h}{N_h} = \frac{n}{N}, \quad \text{para} \quad h = 1,\ldots,H$$

Si suponemos que $n_h = \frac{nN_h}{N}$ es entero, el estimador de la media poblacional es:

$$\hat{\bar{Y}}_{prop} = \frac{1}{N}\sum_{h=1}^{H} N_h \hat{\bar{Y}}_h = \frac{1}{n}\sum_{k\in S} y_k$$

De la misma manera se pueden realizar afijaciones proporcionales a la raíz, al cubo o a cualquier potencia menor que 1.

2. Afijación de mínima varianza

La afijación de mínima varianza o afijación de Neyman consiste en determinar los valores de n_h de forma que para un tamaño de muestra fijo igual a *n*, la varianza de los estimadores sea mínima.

Utilizando los multiplicadores de Lagrange, se obtienen los valores de n_h necesarios

$$n_h = n\frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h} \quad \text{para} \quad h = 1,\ldots,H$$

3. Afijación de tamaño de muestra mínimo

En este caso el problema consiste en buscar la afijación que da el tamaño de muestra mínimo n*, para una varianza fijada *V*.

Una vez más, gracias a los multiplicadores de Lagrange, se tiene que:

$$n^* = \frac{\left(\sum_{h=1}^{H} N_h S_{yh}\right)^2}{V + \sum_{h=1}^{H} N_h S_{yh}^2}$$

Muestreo por conglomerados o clusters

Supongamos que la población U está dividida en M subconjuntos U_i, $i=1,...,M$ llamados conglomerados; los cuales cumplen las siguientes propiedades:

(i) $\bigcup_{i=1}^{M} U_i = U$

(ii) $U_i \bigcap U_j = \phi$, $i \neq j$

(iii) $\sum_{i=1}^{M} N_i = N$ donde N_i es el número de elementos del conglomerado U_i.

Un diseño muestral es por conglomerados si se selecciona una muestra de conglomerados de tamaño m, que denotaremos s_I, con un plan $p_I(s_I)$ en donde todas las unidades de los conglomerados seleccionados son observadas.

La muestra aleatoria completa viene dada por $S = \bigcup_{i \in S_I} U_i$ cuyo tamaño es $n = \sum_{i \in S_I} N_i$. El tamaño de la muestra es generalmente aleatorio.

Esta técnica de muestreo se utiliza cuando la población se encuentra dividida de manera natural en grupos que se supone que contienen toda la variabilidad de la población; es decir, cada conglomerado representa fielmente la característica poblacional a estudiar (simplificando la recogida de información muestral).

Selección de conglomerados con probabilidades iguales

Si suponemos que todos los conglomerados tienen la misma probabilidad de ser seleccionados, entonces el plan de muestreo consiste en seleccionar los conglomerados siguiendo un m.a.s de tamaño *m*.

En este caso, la probabilidad de seleccionar un conglomerado es $\pi_{Ii} = \frac{m}{M}$, obteniendo la siguiente expresión simplificada para el estimador de Horvitz-Thompson de la media:

$$\hat{\bar{Y}}_\pi = \frac{1}{N}\sum_{k \in S}\frac{y_k}{\pi_k} = \frac{1}{N}\sum_{i \in S_i}\frac{N_i \bar{Y}_i}{\pi_{Ii}} = \frac{M}{Nm}\sum_{i \in S_i} N_i \bar{Y}_i$$

donde $\bar{Y}_i = \frac{1}{N_i}\sum_{k \in U_i} y_k$ es la media para el conglomerado U_i, $i=1,...,M$

La varianza del estimador puede estimarse sin sesgo por:

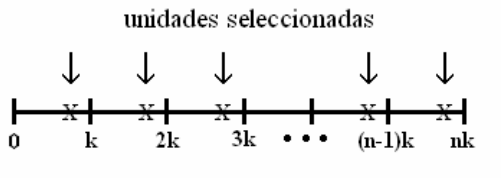
$$\hat{Var}(\hat{\bar{Y}}_\pi) = \frac{M-m}{N^2 m}\frac{M}{m-1}\sum_{i \in S_I}\left(Y_i - \frac{\hat{Y}_\pi}{M}\right)^2$$

Muestreo sistemático con probabilidades iguales

Supongamos que las *N* unidades de la población *U* son numeradas de 1 a *N* en algún orden (aleatorio o siguiendo algún criterio de orden).

Si *n* es el número de unidades a seleccionar en la muestra, definimos $k = N/n$ como el intervalo de muestreo.

Seleccionamos un número aleatorio $r \in \{1,...,k\}$ como unidad de inicio. A partir de r, las unidades que se encuentran a una distancia lk para $l = 1,2,...,n-1$ son seleccionadas en la muestra.



El muestreo sistemático puede verse como un muestreo por conglomerados donde el problema consiste en escoger un único cluster de los *k* posibles.

Composición de las k posibles muestras sistemáticas

1	2	...	i	...	k
y_1	y_2		y_i		y_k
y_{k+1}	y_{k+2}		y_{k+i}		y_{2k}
...	...		...		...
$y_{(n-1)k+1}$	$y_{(n-1)k+2}$		$y_{(n-k)k+3}$		y_{nk}

Resumen de los métodos presentados

Se define el coeficiente de variación del estimador $\hat{\theta}$, como el cociente entre la desviación estándar y su valor real θ, $CV(\hat{\theta}) = \frac{\sqrt{Var(\hat{\theta})}}{\theta}$.

Por lo tanto, el estimador del coeficiente de variación de $\hat{\theta}$, es $cv(\hat{\theta}) = \frac{\sqrt{\hat{V}ar(\hat{\theta})}}{\hat{\theta}}$

A continuación se muestra una tabla con las formulas del estimador, varianza y coeficientes de variación tanto para la media poblacional como para las proporciones de los distintos métodos presentados.

		Muestreo aleatorio simple	Muestreo estratificado	Muestreo por conglomerados
Media poblacional $\bar{Y}$	Estimador $\hat{\bar{Y}}$	$\hat{\bar{Y}}_\pi = \frac{1}{n}\sum_{k\in S} y_k$	$\hat{\bar{Y}}_{st} = \frac{1}{N}\sum_{h=1}^{H} N_h \hat{\bar{Y}}_h$	$\hat{\bar{Y}}_\pi = \frac{M}{Nm}\sum_{i\in S_I} N_i \bar{Y}_i$
	Varianza $\hat{Var}(\hat{\bar{Y}})$	$Var(\hat{\bar{Y}}_\pi) = (1-f)\frac{s_y^2}{n}$	$\hat{Var}(\hat{\bar{Y}}_{st}) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 (1-f_h)\frac{s_{yh}^2}{n_h}$	$\hat{Var}(\hat{\bar{Y}}_\pi) = \frac{M-m}{N^2 m}\frac{M}{m-1}\sum_{i\in S_I}\left(Y_i - \frac{\hat{Y}_\pi}{M}\right)^2$
	Coef. de variación $cv(\hat{\bar{Y}})$	$cv(\hat{\bar{Y}}_\pi) = \sqrt{\frac{(1-f)}{n}}cv(\hat{\bar{Y}})$	$cv(\hat{\bar{Y}}_{st}) = \frac{\sqrt{\sum_{h=1}^{H} N_h^2 (1-f_h)\frac{s_{yh}^2}{n_h}}}{\sum_{h=1}^{H} N_h \hat{\bar{Y}}_h}$	$cv(\hat{\bar{Y}}_\pi) = \frac{\sqrt{\left(1-\frac{m}{M}\right)\frac{m}{m-1}\sum_{i\in S_I}\left(Y_i - \frac{\hat{Y}_\pi}{M}\right)}}{\sum_{i\in S_I} N_i \bar{Y}_i}$
Proporciones P	Estimador $\hat{P}$	$\hat{P} = \frac{1}{n}\sum_{k\in S} y_k$	$\hat{P}_{st} = \frac{1}{N}\sum_{h=1}^{H} N_h p_h$	$\hat{P} = \frac{\sum_{i\in S_I} a_i}{\sum_{i\in S_I} N_i}$; donde $a_i = p_i N_i$
	Varianza $\hat{Var}(\hat{P})$	$\hat{Var}(\hat{P}) = (1-f)\frac{p(1-p)}{n-1}$	$\hat{Var}(\hat{P}_{st}) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 (1-f_h)\frac{p_h q_h}{n_h - 1}$	$\hat{Var}(p) = \frac{M-m}{M}\frac{m}{m-1}\frac{\sum_{i\in S_I} a_i^2 - 2p\sum_{i\in S_I} a_i N_i + p^2\sum_{i\in S_I} N_i^2}{\sum_{i\in S_I} N_i}$
	Coef. de variación $cv(\hat{P})$	$cv(\hat{P}) = \sqrt{(1-f)\frac{(1-p)}{p(n-1)}}$	$cv(\hat{P}_{st}) = \frac{\sqrt{\sum_{h=1}^{H} N_h^2 (1-f_h)\frac{p_h(1-p_h)}{n_h - 1}}}{\sum_{h=1}^{H} N_h p_h}$	$cv(\hat{P})^2 = \frac{M-m}{M}\frac{m}{m-1}\frac{\sum_{i\in S_I} a_i^2 - 2p\sum_{i\in S_I} a_i N_i + p^2\sum_{i\in S_I} N_i^2}{(\sum_{i\in S_I} a_i)^2}$

4. Planes de muestreo complejos

Pese a que los métodos presentados forman los tres principales tipos de muestreos probabilísticos, a la hora de definir los diseños de las encuestas elaboradas por EUSTAT o por los distintos órganos estadísticos, estos diseños suelen ser un poco más complejos.

Muestreo bietápico (o de dos etapas)

Supongamos que la población $U = \{1,...,k,...,N\}$ está compuesta de *M* subpoblaciones U_i, $i = 1,...,M$ llamadas unidades primarias.

Al mismo tiempo, cada unidad primaria U_i se compone de N_i unidades secundarias donde $\sum_{i=1}^{M} N_i = N$.

De manera general, un muestreo bietápico se define de la siguiente manera:

- Se selecciona una muestra S_I de unidades primarias de tamaño *m.*
- Si una unidad primaria es seleccionada en la primera etapa, se selecciona una muestra S_i de tamaño n_i de unidades secundarias.
- Los planes bietápicos tienen que cumplir las propiedades de invarianza e independencia.

La muestra aleatoria completa viene dada por $S = \bigcup_{i \in S_I} S_i$ cuyo tamaño es $n = \sum_{i \in S_I} n_i$.

Podemos definir:

- $\pi_{I,i}$ como la probabilidad de seleccionar la unidad primaria U_i
- $\pi_{k|i}$ como la probabilidad de seleccionar la unidad k dado que U_i ha sido seleccionada.

Por lo tanto, la probabilidad de inclusión de la unidad *k* es:

$$\pi_k = \pi_{I,i}\pi_{k|i}, \qquad k \in U_i$$

El estimador de H-T de la media en un muestreo bietápico es:

$$\hat{\bar{Y}}_\pi = \frac{1}{N}\sum_{k \in S}\frac{y_k}{\pi_k} = \frac{1}{N}\sum_{i \in S_I}\sum_{k \in S_i}\frac{y_k}{\pi_{I,i}\pi_{k|i}} = \frac{1}{N}\sum_{i \in S_I}\frac{N_i\hat{\bar{Y}}_i}{\pi_{I,i}}$$

donde $\hat{\bar{Y}}_i = \frac{1}{N_i}\sum_{k \in S_i}\frac{y_k}{\pi_{k|i}}$ es el estimador de H-T de la media de la unidad primaria U_i

Además, en un plan bietápico se tiene que $\hat{V}ar(\hat{Y}_\pi) = \hat{V}ar_{UP} + \hat{V}ar_{US}$,

donde $\hat{V}ar_{UP}$ es la parte de la varianza que se refiere a las unidades primarias y $\hat{V}ar_{US}$ a las unidades secundarias.

Por lo tanto, en un muestreo bietápico, podemos combinar los principales planes de muestreo probabilísticos presentados (muestreo aleatorio simple, estratificado y por conglomerados) tanto en la selección de las unidades primarias como secundarias.

Selección de las UP-s con probabilidades iguales

Supongamos que en las dos etapas del muestreo se usa un muestreo aleatorio simple.

Entonces, las probabilidades antes definidas toman la siguiente forma:

$$\pi_{I,i} = \frac{m}{M}, \quad i = 1,...,M$$

$$\pi_{k|i} = \frac{n_i}{N_i}, \quad i = 1,...,M, \; k \in S_i$$

En este caso, la probabilidad de inclusión de la unidad *k* es:

$$\pi_k = \frac{mn_i}{MN_i}, \quad k \in U_i$$

Si modificamos la formula del estimador H-T para muestreos bietápicos tenemos que:

$$\hat{\bar{Y}}_\pi = \frac{1}{N}\sum_{k\in S}\frac{y_k}{\pi_k} = \frac{M}{Nm}\sum_{i\in S_I}\sum_{k\in S_i}\frac{N_i y_k}{n_i}$$

Y su estimador de la varianza se simplifica

$$\hat{V}ar(\hat{\bar{Y}}_\pi) = \frac{M-m}{N^2 m} M s_I^2 + \frac{M}{N^2 m}\sum_{k\in S_i} N_i \; \frac{N_i - n_i}{n_i} s_i^2$$

donde

$$s_I^2 = \frac{1}{m-1}\sum_{i\in S_I}\left(\hat{Y}_i - \frac{\hat{Y}_\pi}{M}\right)^2 \qquad \text{y} \qquad s_i^2 = \frac{1}{n_i - 1}\sum_{k\in S_I}\left(y_k - \frac{\hat{Y}_\pi}{N_i}\right)^2$$

Plan bietápico autoponderado

Supongamos que en la primera etapa, las unidades primarias son seleccionadas con probabilidad de inclusión proporcionales al tamaño (PPT); es decir,

$$\pi_{I,i} = \frac{N_i}{N} m$$

En la segunda etapa, se seleccionan las unidades secundarias según un muestreo aleatorio simple de tamaño fijo $n_i = n_0$ (en cada unidad primaria); es decir,

$$\pi_{k|i} = \frac{n_0}{N_i}$$

Por lo tanto, las probabilidades de inclusión de la unidad *k* son iguales para todas las unidades de la población *U*:

$$\pi_k = \pi_{I,i}\pi_{k|i} = \frac{N_i}{N} m \frac{n_0}{N_i} = \frac{mn_0}{N}$$

5. Método del Cubo: Muestre Equilibrado

El Método del Cubo (Deville and Tillé, 2004), es un método que permite seleccionar muestras equilibradas con probabilidades de inclusión iguales o desiguales, optimizando los métodos de muestreo probabilísticos.

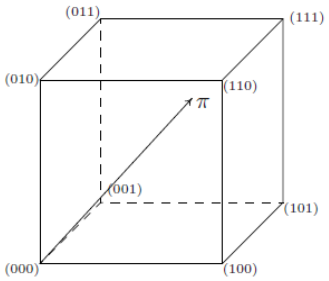
Intuitivamente, este método permite mantener las proporciones de la población original en la muestra sobre ciertas variables de equilibrio (variables cualitativas), teniendo siempre en cuenta las probabilidades de inclusión del diseño. Estas variables de equilibrio, deben estar fuertemente correlacionadas con las variables de interés.

Representación por un cubo

Consideremos una población finita *U* = {1,…, *N*} de tamaño *N*, donde el objetivo es estimar el total (o media) de ciertas variables de interés.

Para poder entender el funcionamiento del Método del Cubo, supongamos que una muestra es en realidad un vector $\mathbf{s} = (s_1 \ldots s_k \ldots s_N)^t$ donde s_k toma el valor 1 si la unidad *k* esta en la muestra y 0 en caso contrario.

Geométricamente, cada vector *s* es un vértice de un N-cubo.



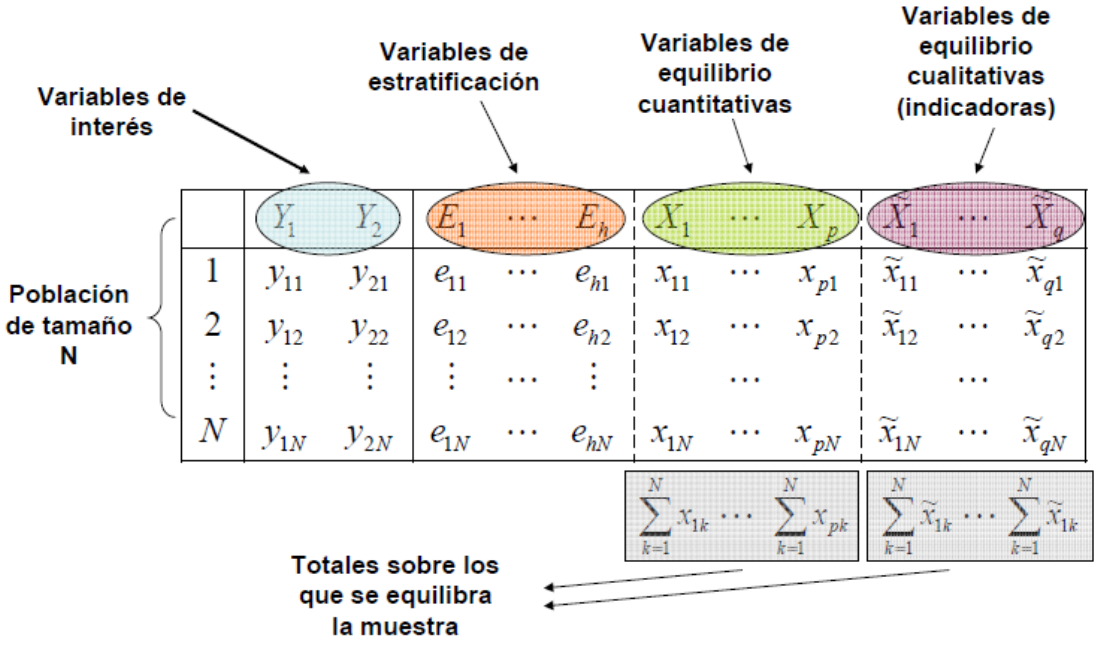
Muestras posibles en una población de tamaño N=3

Por lo tanto, un diseño muestral *p(.)* se trataría de una distribución de probabilidad de todas las posibles muestras sobre el conjunto $S = \{0,1\}^N$; definiendo la probabilidad de inclusión de la unidad k como $\pi_k = \Pr(\mathbf{S}_k = 1)$.

Muestras equilibradas

Supongamos que disponemos de ciertas variables auxiliares con valores conocidos para todas las unidades de la población, $k \in U$.

Estas variables auxiliares podrían ser utilizadas bien como variables de estratificación (cualitativas), o bien como variables de equilibrio (cualitativas o cuantitativas).



Por lo tanto, se dice que una **muestra *s* es equilibrada sobre las variables** $x_1,x_2,...,x_p$ si se verifican las ecuaciones de equilibrio:

$$\hat{\mathbf{X}}_{\pi} = \mathbf{X} \Leftrightarrow \sum_{k\in s} \frac{x_{kj}}{\pi_k} = \sum_{k\in U} x_{kj} \qquad \begin{array}{l} \forall s \in S \ \text{con} \ \ p(s) > 0 \\ j = 1,..., p \end{array}$$

Es decir, que los estimadores de Horvitz-Thompson de las variables $x_1,x_2,...,x_p$ en la muestra son iguales a los totales de estas variables en la población.

El vector de probabilidades de inclusión $\boldsymbol{\pi}$ estará siempre predeterminado por el propio diseño muestral.

Las ecuaciones que derivan de estas restricciones de equilibrio, definen un subespacio (Q) de dimensión N – p en R^N . Por lo tanto, el problema se traduce en elegir un vértice (una muestra) del N-cubo que quede dentro del subespacio Q.

Dado que no es posible seleccionar una muestra exactamente equilibrada, el Método del Cubo implementa un método que selecciona muestras aproximadamente equilibradas.

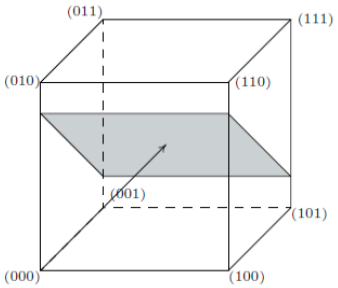
Descripción del método

El método del cubo propuesto por Deville y Tillé (2004), consta de dos fases:

1. Fase de vuelo

Es una generalización del método de escisión (Ver *"Teoría de Muestreo"*).

Se trata de un camino aleatorio que comienza con el vector de probabilidad de inclusión $\boldsymbol{\pi}$ y que permanece en la intersección del cubo y el subespacio definido por las ecuaciones de equilibrio (Q).



2. Fase de aterrizaje

Si al final de la fase de vuelo una muestra (un vértice) no ha sido seleccionada, se deberá aplicar la fase de aterrizaje.

Existen tres posibles soluciones para esta fase:

- Eliminar progresivamente las variables de equilibrio y volver a aplicar la fase de vuelo (es necesario suprimir las variables en orden de menor a mayor importancia).
- Usar la programación lineal para calcular la mejor muestra aproximadamente equilibrada (minimizando la diferencia en equilibrio).
- Escoger el vértice más cercano al vector de probabilidades que se obtiene en la fase de vuelo, redondeando las probabilidades de inclusión que todavía no son iguales a 0 o 1.

Deville y Tille programaron una implementación mucho más rápida de la fase de vuelo (Ver "*Fast SAS Macros for balancing simples user´s guide*"), la cual consume la mayor parte del tiempo de ejecución, obteniendo las siguientes ventajas:

o No hay restricciones en el tamaño de la población.

o El tiempo de ejecución depende linealmente del tamaño de la población.

6. Macros de SAS para seleccionar muestras equilibradas

A continuación, se van a presentar las macros de SAS que nos permiten seleccionar muestras equilibradas.

Las dos principales macros (*exe_cube* y *echant_estrat*) han sido desarrolladas por Guillaume Chauvet e Yves Tillé; mientras que las macros auxiliares *disjunctive* y *crear_estrato* han sido elaboradas en Eustat con el objetivo de agilizar el manejo de las anteriores.

A pesar de que en Eustat se ha optado por trabajar con las macros de SAS que implementan el Método del Cubo, también están disponibles las funciones que seleccionan muestras equilibradas en R (ver paquete *sampling:* *http://cran.r-project.org/web/packages/sampling/index.html*).

Macro *exe_cube*

La macro de SAS *exe_cube,* permite seleccionar muestras equilibradas utilizando el Método del Cubo (Fast Cube Method).

Datos de entrada

Se trata de una tabla de SAS con todas las unidades de la población sobre la que se va a seleccionar la muestra.

Debe contener al menos:

- Una variable de identificación
- Variable con las probabilidades de inclusión
- Variables sobre las que se quiere equilibrar la muestra

Esta tabla no puede tener valores faltantes en las variables mencionadas.

Sintaxis de la macro

Esta es una breve descripción de los argumentos necesarios:

- BASE = Nombre de la librería SAS que contiene la tabla con los datos de entrada.
- DATA = Nombre de la tabla de SAS con los datos de entrada.
- ID = Variable de identificación de las unidades de la población.
- PI = Variable con las probabilidades de inclusión.

- CONTR = Variables sobre las que se quiere equilibrar la muestra.
- ATTER = Opción seleccionada para la fase de aterrizaje

1. Las variables de equilibrio son eliminadas progresivamente
2. Se consideran todas las posibles muestras para las unidades restantes (valores distintos de 0 o 1), seleccionando aquellas que proporcionan una menor diferencia al equilibrio.
3. Mismo procedimiento que la opción 2 pero considerando únicamente las muestras con tamaño igual a la suma de las probabilidades de inclusión (tamaño muestral fijo).
4. Se redondean las probabilidades de inclusión para las unidades restantes manteniendo el tamaño de la muestra predeterminado.

Para utilizar las opciones 3 o 4, debe introducirse la variable de probabilidades de inclusión en el parámetro *contr*.

- COMPEQ = Igual a 1 para equilibrar también el complementario de la muestra.
- SORT = Nombre de la tabla de SAS con los datos de salida, que se guardara en la librería especificada en el parámetro *base.* Contiene todas las unidades de la población, así como la variable *ech*; igual a 1 si la unidad ha sido seleccionada y 0 en caso contrario.

Macro *echant_strat*

La macro de SAS echant_strat permite seleccionar muestras estratificadas con el Método del Cubo (Fast Cube Method), globalmente equilibradas en la población total y aproximadamente equilibradas en cada estrato.

Los pasos que sigue la macro para seleccionar una muestra equilibrada son:

1. Fase de vuelo independiente en cada uno de los estratos
2. Fase de vuelo conjunta con todas las unidades restantes que no hayan sido seleccionadas en los estratos
3. Fase de aterrizaje con las unidades todavía no seleccionadas.

Datos de entrada

Tiene que haber una tabla de SAS con las unidades de la población para cada una de los estratos definidos para la muestra estratificada.

Cada tabla debe contener al menos, las mismas variables que hemos definido para la macro *exe_cube*.

Sintaxis de la macro

Esta es una breve descripción de los argumentos necesarios:

- DATA = Nombre de las tablas de SAS con los datos de entrada de cada estrato.
- ID = Variable de identificación de las unidades de la población.
- PI = Variable con las probabilidades de inclusión.
- CONTR = Variables sobre las que se quiere equilibrar la muestra.
- SORT = Nombre de la tabla de SAS con los datos de salida.

Macro auxiliar *disjunctive*

La macro de SAS *disjunctive* permite dividir una o más variables de interés en variables disjuntas en función de ciertas categorías. La macro además, permite introducir los nombres de dichas categorías.

Descripción

Supongamos que en una población de tamaño N, dada una variable de interés Y y una variable cualitativa X que toma los valores 1, 2,…, *L*; la macro disjunctive nos devuelve las variables disjuntas $Y^1, Y^2, \ldots, Y^L$ donde:

$$y_i^l = \begin{cases} y_i & \text{si } x_i = l \\ 0 & \text{si } x_i \neq l \end{cases} \quad \text{para} \quad \begin{array}{l} i = 1, \ldots, N \\ l = 1, \ldots, L \end{array}$$

Sintaxis de la macro

Esta es una breve descripción de los argumentos necesarios:

- DATA = Nombre de la tabla de SAS que contiene los datos de la población.
- VAR = Variable(s) de interés.
- CATEG = Variable cualitativa que contiene las categorías para crear las variables disjuntas
- NOMBRES_CATEG (opcional) = Nombres de las categorías de la variable *categ*.

  Por defecto categ1, categ2,…, categL.

Resultados y salidas

La macro *disjunctive* añade a la tabla de entrada las variables disjuntas creadas a partir de la variable de interés *var*.

Los nombres de estas nuevas variables son la unión del nombre de la variable *var* y los nombres definidos por la variable *nombres_categ* (separados por el símbolo "_").

Estos nombres son guardados en la variable local macro *contr_categ*.

Macro auxiliar *crear_estrato*

La macro de SAS *crear_estrato* permite dividir una tabla de SAS en varias tablas en función de una variable de estratificación.

Sintaxis de la macro

Esta es una breve descripción de los argumentos necesarios:

- DATA = Nombre de la tabla de SAS que contiene los datos de la población.
- ID = Variable de identificación

- VAR_ESTRAT = Variable sobre la que se quiere realizar la estratificación

Resultados y salidas

La macro *crear_estrato* devuelve una tabla de SAS para cada uno de los valores de la variable *var_estrat*.

Los nombres de las tablas de salida son por defecto del estilo: $\text{estrato\_}\{var\_estrat\}_j$

donde $\{var\_estrat\}_j$ es el j-ésimo valor de la variable *var_estrat*.

Estos nombres son guardados en la variable local macro *datos_estrat*.

Ejemplo de uso de las macros

Supongamos que queremos seleccionar una muestra estratificada de establecimientos, equilibrando la muestra sobre el número de empleados por Territorio Histórico.

Nuestra tabla de SAS inicial con los datos de la población tendría un aspecto como esta:

datos

id	*estrato*	*pik*	*empleo*	*TH*
1	*A*	π_1	e_1	48
2	*A*	π_2	e_2	20
3	*B*	π_3	e_3	20
4	*B*	π_4	e_4	01
5	*B*	π_5	e_5	48
6	*C*	π_6	e_6	01
7	*C*	π_7	e_7	20

donde

01 = Araba, 20 = Gipuzkoa y 48 = Bizkaia;

π_k es la probabilidad de inclusión del establecimiento *k;*

e_k es el número de empleados en el establecimiento *k.*

- En primer lugar aplicaremos la macro *disjunctive* para calcular las variables de equilibrio disjuntas correspondientes al número de empleados por TH.

```
%global contr_categ;
%disjunctive(
      DATA = datos,
      VAR = empleo,
      CATEG = TH,
      NOMBRES_CATEG = Araba Gipuzkoa Bizkaia
);
```

datos

id	*estrato*	*pik*	*empleo*	*TH*	*empleo _ Araba*	*empleo _ Gipuzkoa*	*empleo _ Bizkaia*
1	*A*	π_1	e_1	48	0	0	e_1
2	*A*	π_2	e_2	20	0	e_2	0
3	*B*	π_3	e_3	20	0	e_3	0
4	*B*	π_4	e_4	01	e_4	0	0
5	*B*	π_5	e_5	48	0	0	e_5
6	*C*	π_6	e_6	01	e_6	0	0
7	*C*	π_7	e_7	20	0	e_7	0

Tal y como hemos mencionado, el objetivo es seleccionar una muestra equilibrada sobre el número de empleados por TH, es decir, sobre los totales:

$$\sum_{k\in N} empleo\_Araba_k \ , \quad \sum_{k\in N} empleo\_Gipuzkoa_k \ \text{ y } \ \sum_{k\in N} empleo\_Bizkaia_k$$

En este caso, la variable macro *contr_categ* guarda los valores:

```
&contr_categ. = empleo_Araba empleo_Gipuzkoa empleo_Bizkaia.
```

- A continuación, aplicaríamos la macro *crear_estrato* para obtener un dataset con los datos correspondientes a cada uno de los estratos.

```
%global datos_estrat;
%crear_estrato(
      DATA = datos,
      ID = id,
      VAR_ESTRAT = estrato
);
```

estrato_A

id	*estrato*	*pik*	*empleo*	*TH*	*empleo _ Araba*	*empleo _ Gipuzkoa*	*empleo _ Bizkaia*
1	*A*	π_1	e_1	48	0	0	e_1
2	*A*	π_2	e_2	20	0	e_2	0

estrato_B

id	*estrato*	*pik*	*empleo*	*TH*	*empleo _ Araba*	*empleo _ Gipuzkoa*	*empleo _ Bizkaia*
3	*B*	π_3	e_3	20	0	e_3	0
4	*B*	π_4	e_4	01	e_4	0	0
5	*B*	π_5	e_5	48	0	0	e_5

estrato_C

id	*estrato*	*pik*	*empleo*	*TH*	*empleo _ Araba*	*empleo _ Gipuzkoa*	*empleo _ Bizkaia*
6	*C*	π_6	e_6	01	e_6	0	0
7	*C*	π_7	e_7	20	0	e_7	0

En este caso, la variable macro *datos_estrat* guarda los valores:

```
&datos_estrat. = estrato_A estrato_B estrato_C
```

- Por último, llamaremos a la macro *echant_strat* que selecciona la muestra equilibrada para muestras estratificadas con el Método del Cubo.

```
%echant_strat(
      DATA = &datos_estrat.,
      ID = id,
      PI = pik,
      CONTR = pik &contr_categ.,
      SORT = muestra
);
```

La salida de la macro tendría un aspecto como este:

muestra

id	*ech*
1	ech_1
2	ech_2
3	ech_3
4	ech_4
5	ech_5
6	ech_6
7	ech_7

donde $ech_k = \begin{cases} 1 & \text{si la unidad } k \text{ ha sido seleccionada} \\ 0 & \text{en caso contrario} \end{cases}$ para todo $k \in \{1,...,7\}$

*** *Observación***:

En algunas ocasiones, el objetivo puede ser equilibrar la muestra sobre totales que hacen referencia a las propias unidades muestrales.

Por ejemplo, en el caso anterior se podría querer equilibrar la muestra sobre el número de establecimientos por Territorio Histórico.

En ese caso, debemos crear una variable que toma el valor 1 para todas las unidades, la cual introduciremos en la macro *%disjunctive* para crear las variables de equilibrio deseadas.

datos

id	*estrato*	*pik*	*empleo*	*TH*	*UNO*
1	*A*	π_1	e_1	48	1
2	*A*	π_2	e_2	20	1
3	*B*	π_3	e_3	20	1
4	*B*	π_4	e_4	01	1
5	*B*	π_5	e_5	48	1
6	*C*	π_6	e_6	01	1
7	*C*	π_7	e_7	20	1

```
%global contr_categ;
%disjunctive(
      DATA = datos,
      VAR = UNO,
      CATEG = TH,
      NOMBRES_CATEG = Araba Gipuzkoa Bizkaia
);
```

datos

id	*estrato*	*pik*	*empleo*	*TH*	*UNO*	*UNO _ Araba*	*UNO _ Gipuzkoa*	*UNO _ Bizkaia*
1	*A*	π_1	e_1	48	1	0	0	1
2	*A*	π_2	e_1	20	1	0	1	0
3	*B*	π_3	e_1	20	1	0	1	0
4	*B*	π_4	e_1	01	1	1	0	0
5	*B*	π_5	e_1	48	1	0	0	1
6	*C*	π_6	e_1	01	1	1	0	0
7	*C*	π_7	e_1	20	1	0	1	0

7. Muestras equilibradas en EUSTAT con el Método del Cubo

A continuación, se van a presentar algunos de los diseños muestrales que han sido equilibrados mediante el Método del Cubo en Eustat.

Para cada uno de los casos, se describirá el diseño metodológico: la ficha técnica, las variables de estratificación, afijaciones y probabilidades de inclusión y las variables sobre las que se ha equilibrado la muestra. También se presentarán algunos de los resultados obtenidos.

Muestra de centros de ESO para el estudio del "bullying" en la Comunidad Autónoma de Euskadi

El Departamento de Educación, Universidades e Investigación, por medio del Instituto Vasco de Evaluación e Investigación (ISEI-IVEI), realiza una encuesta a alumnado de ESO sobre el maltrato escolar en los centros de la Comunidad Autónoma de Euskadi.

Para ello, se debía extraer una muestra de conglomerados (centros) de forma que se evalúe un máximo de 40 alumnos por centro seleccionado.

Ficha Técnica

- Marco

  Lo componen los centros de Secundaria de la CAE que tienen al menos un grupo en los cursos de 1º, 2º, 3º y 4º de la ESO.

- Diseño muestral

  Se trata de una muestra de conglomerados desiguales con submuestreo en la segunda etapa.

  1.a etapa

  Unidades muestrales

  Centros de secundaria de la CAE

  Estratificación

  Para la selección de los centros se realiza un muestreo estratificado por Territorio Histórico y red (pública y privada).

  Afijación

  Proporcional al número de centros en cada estrato.

  Sorteo

  Muestreo probabilístico proporcional al tamaño (PPT) del número de alumnos por centro.

2.a etapa

Unidades muestrales

Alumnos de secundaria de la CAE.

Estratificación

40 alumnas (10 de 1º, 10 de 2º, 10 de 3º y 10 de 4º) por centro seleccionado siempre que sea posible. No hay un mínimo de alumnos por centro.

Sorteo

Muestreo aleatorio simple.

La muestra final es autoponderada por estratos (Territorio y Red).

- Tamaño de la muestra

El tamaño de la muestra óptimo para un muestreo de conglomerados, se calculó a partir de la siguiente fórmula:

$$n_{centros} = n_a \frac{\left[(1+\delta(\overline{M}-1)\right]}{\overline{M}}$$

donde n_a es el tamaño de la muestra para un aleatorio simple y el resto es el denominado efecto de diseño en muestreo de conglomerados.

Con $\overline{M}$ = Número medio de alumnos por centro

δ = Correlación intracentro

$$n_a = \frac{Nz_{\alpha/2}^2 S^2}{Ne^2 + z_{\alpha/2}^2 S^2} = \frac{N}{\left[1+(N-1)\frac{e^2}{z_{\alpha/2}^2 pq}\right]}$$

N = Número total de alumnos (unidades elementales)

e = Error máximo admisible

$z_{\alpha/2}^2$ = Valor critico para el nivel de significación α

- Variables de equilibrio

La muestra ha sido equilibrada sobre las siguientes variables:

- Número de alumnos por curso y número de grupos por curso.

  De este modo, las estimaciones de la media de alumnos por centro y grupo son lo más parecidas a los datos facilitados por Estadística Educativa.

- Número de centros pertenecientes a cada tipo de tamaño.

Codificación del tamaño del centro en 5 grupos, minimizando la inercia intraclase en función del tamaño en alumnos: [0-143], [144-243], [244-361], [362-506] y [507-708].

Resultados

A continuación se muestran los resultados obtenidos con el Método del Cubo para las variables de equilibrio.

Cada una de las tablas compara la distribución poblacional con la obtenida a partir de los elevadores de la muestra. Los porcentajes están dados por columnas.

Distribución del número de alumnos por curso

	Poblacional	Muestral (elevado)
1º ESO	19.664 (27,21%)	19.617 (27,14%)
2º ESO	18.633 (25,78%)	18.649 (25,80%)
3º ESO	17.669 (24,45%)	17.764 (24,58%)
4º ESO	16.306 (22,56%)	16.243 (22,47%)
TOTAL	**72.272**	**72.272**

Distribución del número de grupos por curso

	Poblacional	Muestral (elevado)
1º ESO	870 (25,02%)	869 (24,04%)
2º ESO	852 (24,50%)	849 (24,47%)
3º ESO	896 (25,77%)	896 (25,82%)
4º ESO	859 (24,71%)	856 (24,67%)
TOTAL	**3.477**	**3.470**

Distribución del número de centros por tipo de tamaño

	Poblacional	Muestral (elevado)
Tamaño 1	100 (30,12%)	95 (28,79%)
Tamaño 2	128 (38,55%)	129 (39,09%)
Tamaño 3	61 (18,37%)	63 (19,09%)
Tamaño 4	31 (9,34%)	31 (9,39%)
Tamaño 5	12 (3,61%)	12 (3,64%)
TOTAL	**332**	**330**

Teniendo en cuenta las variables sobre las que ha sido equilibrada la muestra, también se han obtenido muy buenos estimadores de la media de alumnos por centro y grupo para cada uno de los cursos.

CURSO 2011/12	Media alumno por centro		Media alumno por grupo	
	Poblacional	Muestral (elevado)	Poblacional	Muestral (elevado)
1º ESO	59.23	59.44	22.60	22.57
2º ESO	56.21	56.51	21.90	21.97
3º ESO	53.22	53.83	19.72	19.83
4º ESO	49.11	49.22	18.98	18.98
TOTAL	**217.69**	**219.00**	**20.79**	**20.33**

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

El objetivo genérico de la ESI, llevada a cabo por EUSTAT, es proporcionar a los responsables políticos, agentes económicos y sociales, universidad, investigadores privados y ciudadanía en general, información periódica sobre la penetración de las nuevas tecnologías de la información y la de la comunicación (TIC) en las empresas del País Vasco.

La muestra de la ESI-Empresas se caracteriza por ser un panel que cada año incluye a las empresas titulares que han contestado en anteriores repeticiones de la encuesta. Debido a diversas incidencias (bajas, sustituciones, no-respuesta…) el reparto original de la muestra se deteriora, por lo que se tomó la decisión de actualizar la muestra conforme a un nuevo reparto muestral que, respetando el diseño original, recoge la nueva distribución de la población en los estratos.

En el año 2012, se decide renovar el panel en casi un 15%. Además, se introduce el Método del Cubo para seleccionar muestras equilibradas con el objetivo de obtener una distribución equilibrada en las comarcas del País Vasco.

Ficha Técnica

- Marco

  Lo componen los establecimientos de cualquier sector de actividad que ejerza su actividad en el ámbito de la CAE, salvo el sector primario y el servicio doméstico.

- Diseño muestral

  Se trata de una muestra estratificada de una sola etapa.

  Unidades muestrales

  Todos los establecimientos que forman parte del marco mencionado.

  Estratificación

  Se realiza un muestro estratificado por el cruce de las siguientes variables:

  - Territorio Histórico

    1 = Araba; 2 = Bizkaia; 3 = Gipuzkoa

  - Estrato de empleo

    1 = 0-5 empleados; 2 = 6-9 empleados; 3 = 10-19 empleados;
    4 = 20-49 empleados; 5 = 50-99 empleados; 6 = 100 y más empleados

  - Sector de actividad (CNAE09 a 2 dígitos)

  Afijación

  Elementos autorrepresentados: establecimientos con 100 empleados y más (estrato de empleo 6).

  Para el resto de los establecimientos se realizan dos afijaciones diferentes:

1. Partiendo de un tamaño muestral prefijado en el diseño original de n=7000, se realiza un reparto proporcional a la raíz del nº de establecimientos por territorio y directamente proporcional al nº de establecimientos por estrato (territorio, actividad y empleo).

   El tamaño de la muestra en cada estrato es calculado a partir de la siguiente fórmula:

$$n_{TH_i Act_j Emp_k} = n_{TH_i} \frac{estab_{Act_j Emp_k}}{\sum_{j \in Act} \sum_{k=1}^{5} estab_{Act_j Emp_k}}$$

   donde $$n_{TH_i} = (7000 - censales) \frac{\sqrt{estab_{TH_i}}}{\sum_{i=1}^{3} \sqrt{estab_{TH_i}}} \qquad i = 1,2,3$$

   Finalmente se añaden establecimientos hasta obtener un tamaño mínimo de 5 establecimientos en los estratos de empleo agrupados (menos de 10 empleados y más de 10 empleados).

2. Reparto en función del error muestral máximo de un 10% en cada sector de actividad (sin tener en cuenta los estratos censales).

   El tamaño de la muestra en cada sector de actividad es calculado a partir de la fórmula

$$n_h = \frac{N_h z_{\alpha/2}^2 S_h^{\,2}}{N_h e^2 + z_{\alpha/2}^2 S_h^{\,2}} = \frac{N_h}{\left[1 + (N_h - 1)\frac{e^2}{z_{\alpha/2}^2 pq}\right]}$$

   donde N_h = Número de establecimientos en el estrato h
   e = Error máximo admisible
   $z_{\alpha/2}^2$ = valor critico para el nivel de significación α

Una vez realizada ambas afijaciones, se reparten las unidades faltantes hasta obtener el tamaño de muestra necesario para las unidades no censales. Este reparto se realiza de forma proporcional al tamaño del estrato en los sectores infra-representados con respecto a la primera afijación.

Finalmente, estas afijaciones por sector de actividad, se reparten de manera proporcional a la raíz en cada territorio y empleo agrupado.

Sorteo

Se realiza un muestreo aleatorio simple en cada uno de los estratos, dando prioridad a los establecimientos que estén especificados en el marco como altas.

- Variables de equilibrio

  Con el objetivo de obtener mejores estimaciones a nivel comarcal, la muestra ha sido equilibrada sobre el número de establecimientos en cada comarca (20 comarcas).

- Sustitutos

  Para completar la muestra se necesita una bolsa de sustitutos de unos 3.500 establecimientos. El número de suplentes por estrato, es proporcional a la muestra teórica en cada uno de los estratos de empleo y territorio.

  Al igual que en la muestra principal, la muestra de sustitutos se equilibrará con el Método del Cubo sobre el número de establecimientos en cada comarca.

Resultados

A continuación se muestran los resultados obtenidos con el Método del Cubo al equilibrar la el número de establecimientos por comarca.

Distribución del número de establecimientos por comarca

	Poblacional	Muestral (elevado)
Valles Alaveses	405 (0.22 %)	523 (0.29 %)
Llanada Alavesa	18.903 (10.49 %)	19.063 (10.58 %)
Montaña Alavesa	248 (0.14 %)	257 (0.14 %)
Rioja Alavesa	1.311 (0.73 %)	1.135 (0.63 %)
Estribaciones del Gorbea	780 (0.43 %)	749 (0.42 %)
Cantábrica Alavesa	2.180 (1.21 %)	2.099 (1.16 %)
Arratia - Nervión	1.787 (0.99 %)	1.399 (0.78 %)
Gran Bilbao	73.572 (40.82 %)	72.517 (40.24 %)
Durangaldea	7.517 (4.17 %)	7.795 (4.33 %)
Encartaciones	2.356 (1.31 %)	2.364 (1.31 %)
Gernika – Bermeo	3.425 (1.90 %)	3.364 (1.87 %)
Markina – Ondarroa	1.828 (1.01 %)	2.446 (1.36 %)
Plentzia – Mungia	4.008 (2.22 %)	4.609 (2.56 %)
Bajo Bidasoa	7.169 (3.98 %)	8.343 (4.63 %)

Bajo Deba	4.191 (2.33 %)	4.989 (2.77 %)
Alto Deba	4.197 (2.33%)	4.742 (2.63 %)
Donostialdea	31.422 (17.44 %)	28.724 (15.94 %)
Goierri	4.929 (2.73 %)	5.192 (2.88 %)
Tolosaldea	4.029 (2.24 %)	4.105 (2.28 %)
Urola Costa	5.966 (3.31 %)	5.809 (3.22 %)
TOTAL	180.223	180.223

Los porcentajes están dados por columnas

Muestra para la Encuesta de Capital Social (ECS)

El capital social es entendido como un recurso al que se accede cuando se dispone de redes personales amplias con las que se participa activamente en los distintos ámbitos económicos y sociales, en un ambiente de confianza y que puede facilitar el desarrollo personal y social, así como el desarrollo económico de una sociedad.

En concreto, en la Encuesta de Capital Social, realizada por Eustat, el capital social está concebido como un conjunto de dimensiones de relación y participación, entre las que se encuentran: las redes sociales de familiares y amigos, la confianza en las personas y las instituciones, la participación social y la cooperación, la información y la comunicación, la cohesión y la inclusión social y la felicidad y la salud.

En el año 2012, se decide seleccionar la muestra para la ECS utilizando el Método del Cubo. De esta manera, hemos logrado obtener una muestra equilibrada por sexo y edad en cada uno de los Territorios Históricos, además de ayudar a obtener mejores estimaciones a nivel comarcal.

Ficha Técnica

- Marco

  El marco de la muestra de la Encuesta sobre Capital Social lo compone la población de 15 años y más residente en viviendas y establecimientos colectivos de la Comunidad Autónoma de Euskadi.

- Diseño muestral

  Se trata de una muestra estratificada de una sola etapa.

Unidades muestrales

Población de 15 años y más residentes en viviendas y establecimientos colectivos de la CAE

Tamaño de la muestra

Se seleccionan n = 7000 individuos.

Estratificación

Se realiza un muestro estratificado por el cruce de las siguientes variables:

- Territorio Histórico

  01 = Araba; 20 = Gipuzkoa; 48 = Bizkaia

- Tamaño del municipio

  Capitales, Medianos (20.000-100.000) y Pequeños (20.000 y menos)

- Nacionalidad

  0 = Nacionales; 1 = Extranjeros

Afijación

Se ha establecido un criterio para cada uno de los niveles de estratificación:

1. Reparto proporcional a la raíz cuadrada del nº de individuos por Territorio.
2. Reparto proporcional al nº de individuos por tamaño de municipio.
3. Reparto proporcional a la potencia 2/3 del nº de individuos por nacionalidad.

Para escoger la afijación más conveniente en el tercer nivel, se han tenido en cuenta las tasas de no respuesta de la anterior encuesta realizada (ECS 2007). Dado que los métodos de recogida de la información de la encuesta son los mismos, podemos suponer que las tasas de respuesta para la encuesta actual van a ser similares.

Por lo tanto, se ha buscado la afijación que permite conseguir el tamaño de muestra mínimo necesario (unas 400 unidades) para poder dar estimaciones a nivel de capitales y población extranjera, teniendo en cuenta estas tasas de respuesta.

El tamaño de la muestra en cada estrato viene especificado por la siguiente fórmula:

$$n_{TH_i TMUN_j NACi_k} = n_{TH_i TMUN_j} \frac{\sqrt[3]{(N_{TH_i TMUN_j NACI_k})^2}}{\sum_k \sqrt[3]{(N_{TH_i TMUN_j NACI_k})^2}}$$

donde $$n_{TH_i TMUN_j} = 7000 \frac{\sqrt{N_{TH_i}}}{\sum_i \sqrt{N_{TH_i}}} \frac{N_{TH_i TMUN_j}}{\sum_j N_{TH_i TMUN_j}}$$

para $$i \in \{Araba, Gipuzkoa, Bizkaia\}$$
$$j \in \{Capitales, Medianos, Pequeños\}$$
$$k \in \{Nacional, Extranjero\}$$

Sorteo

Se realiza un muestreo aleatorio simple en cada uno de los estratos.

- Variables de equilibrio

  La muestra ha sido equilibrada sobre las siguientes variables:

  - Número de individuos en el cruce de Territorio (Araba, Gipuzkoa, Bizkaia), Sexo (Hombres y Mujeres) y Edad (15-24, 24-34, 35-44, 45-54, 55-64 y más de 65 años).

  - Número de individuos en cada una de las 20 comarcas de la CAE.

- Sustitutos

  Para completar la muestra se necesita una bolsa de sustitutos de otros 7.000 individuos. Estos sustitutos han sido extraídos respetando el mismo reparto muestral por estratos que en la muestra original, equilibrando la muestra sobre las mismas variables que los titulares.

Resultados

A continuación se muestran los resultados obtenidos con el Método del Cubo para las variables de equilibrio.

Cada una de las tablas compara la distribución poblacional con la obtenida a partir de los elevadores de la muestra. Los porcentajes están dados por columnas

Distribución por Territorio, Sexo y Edad

TH = ARABA (01)

	Hombres		Mujeres		TOTAL	
	Poblacional	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**
15-24 años	13.818 *(10,06%)*	13.729 *(10,02%)*	12.831 (9,24%)	12.762 (9,17%)	26.649 *(9,65%)*	26.491 *(9,59%)*
25-34 años	23.028 *(16,77%)*	22.923 *(16,73%)*	21.541 (15,51%)	21.725 (15,60%)	44.569 *(16,13%)*	44.648 *(16,16%)*
35-44 años	28.954 *(21,08%)*	28.948 *(21,13%)*	26.298 (18,93%)	26.278 (18,87%)	55.252 *(20,0%)*	55.226 *(19,99%)*
45-54 años	24.889 *(18,12%)*	24.895 *(18,17%)*	24.891 (17,92%)	25.039 (17,98%)	49.780 *(18,02%)*	49.934 *(18,08%)*
55-64 años	20.051 *(14,60%)*	19.942 *(14,55%)*	20.355 (14,65%)	20.332 (14,60%)	40.406 *(14,63%)*	40.274 *(14,58%)*
Más de 65 años	26.584 *(19,36%)*	26.590 *(19,40%)*	33.009 (23,76%)	33.086 (23,76%)	59.593 *(21,57%)*	59.676 *(21,60%)*
TOTAL	**137.324** *(100 %)*	**137.027** (100 %)	138.925 (100 %)	139.222 (100 %)	**276.249** *(100 %)*	**276.249** *(100%)*

TH = GIPUZKOA (20)

	Hombres		Mujeres		TOTAL	
	Poblacional	Muestral (elevado)	Poblacional	Muestral (elevado)	Poblacional	Muestral (elevado)
15-24 años	30.206 *(10,18%)*	30.273 *(10,22%)*	28.416 *(9,09%)*	28.371 *(9,07%)*	58.622 *(9,62%)*	58.644 *(9,63%)*
25-34 años	45.461 *(15,32%)*	45.452 *(15.34%)*	43.313 *(13.86%)*	43.517 *(13,91%)*	88.774 *(14,57%)*	88.968 *(14,60%)*
35-44 años	60.481 *(20,39%)*	60.491 *(20,41%)*	56.318 *(18,02%)*	56.361 *(18,01%)*	116.799 *(19,17%)*	116.852 *(19,18%)*
45-54 años	54.351 *(18,32%)*	54.228 *(18,30%)*	54.409 *(17,41%)*	54.480 *(17,41%)*	108.760 *(17,85%)*	108.707 *(17,84%)*
55-64 años	45.126 *(15,21%)*	44.881 *(15,14%)*	46.428 *(14,85%)*	46.525 *(14,87%)*	91.554 *(15,03%)*	91.406 *(15,0%)*
Más de 65 años	61.051 *(20,58%)*	61.021 *(20,59%)*	83.677 *(26,77%)*	83.638 *(26,73%)*	144.728 *(23,76%)*	144.659 *(23,74%)*
TOTAL	**296.676** *(100 %)*	**296.346** *(100 %)*	**312.561** *(100 %)*	**312.891** *(100 %)*	**609.237** *(100 %)*	**609.237** *(100 %)*

TH = BIZKAIA (48)

	Hombres		Mujeres		TOTAL	
	Poblacional	Muestral (elevado)	Poblacional	Muestral (elevado)	Poblacional	Muestral (elevado)
15-24 años	47.497 *(9.80%)*	47.673 *(9,83%)*	45.007 *(8,59%)*	45.152 *(8,62%)*	92.504 *(9,17%)*	92.825 *(9,20%)*
25-34 años	76.941 *(15,87%)*	76.969 *(15,88%)*	73.755 *(14,07%)*	73.658 *(14,06%)*	150.696 *(14,94%)*	150.627 *(14,93%)*
35-44 años	97.104 *(20,03%)*	97.136 *(20,04%)*	93.542 *(17,85%)*	93.318 *(17,81%)*	190.646 *(18,90%)*	190.454 *(18,88%)*
45-54 años	90.348 *(18,64%)*	90.178 *(18,60%)*	93.048 *(17,75%)*	92.807 *(17,71%)*	183.396 *(18,18%)*	182.985 *(18,14%)*
55-64 años	72.330 *(14,92%)*	72.308 *(14,91%)*	77.119 *(14,71%)*	77.329 *(14,76%)*	149.449 *(14,81%)*	149.637 *(14,83%)*
Más de 65 años	100.487 *(20,73%)*	100.558 *(20,74%)*	141.669 *(27,03%)*	141.762 *(27,05%)*	242.156 *(24,0%)*	242.320 *(24,02%)*
TOTAL	**484.707** *(100 %)*	**484.821** *(100 %)*	**524.140** *(100 %)*	**524.026** *(100 %)*	**1.008.847** *(100 %)*	**1.008.847** *(100 %)*

Distribución del número de individuos por comarca

	Poblacional	Muestral elevado
Valles Alaveses	5.107 (0,27%)	5.051 (0,27%)
Llanada Alavesa	221.595 (11,69%)	221.680 (11,69%)
Montaña Alavesa	2.855 (0,15%)	2.886 (0,15%)
Rioja Alavesa	9.852 (0,52%)	9.835 (0,52%)
Estribaciones del Gorbea	7.296 (0,38%)	7.292 (0,38%)
Cantabrica Alavesa	30.043 (1,58%)	30.004 (1,58%)
Arratia-Nervión	20.289 (1,07%)	20.386 (1,08%)
Gran Bilbao	768.311 (40,53%)	767.962 (40,51%)
Durangaldea	83.470 (4,40%)	83.513 (4,41%)
Encartaciones	27.787 (1,47%)	27.742 (1,46%)
Gernika-Bermeo	40.183 (2,12%)	40.331 (2,13%)
Markina-Ondarroa	23.128 (1,22%)	23.333 (1,23%)
Plentzia-Mungia	46.202 (2,44%)	46.104 (2,43%)
Bajo Bidasoa	66.403 (3,50%)	66.418 (3,50%)
Bajo Deba	47.748 (2,52%)	47.664 (2,51%)
Alto Deba	53.540 (2,82%)	53.584 (2,83%)
Donostialdea	282.424 (14,90%)	282.508 (14,90%)
Goierri	57.859 (3,05%)	57.781 (3,05%)
Tolosaldea	40.147 (2,12%)	40.193 (2,12%)
Urola Costa	61.490 (3,24%)	61.462 (3,24%)
TOTAL	**1.895.729**	**1.895.729**

Muestra para la Encuesta de Innovación Tecnológica (EIT)

El principal objetivo de la EIT, llevada a cabo por EUSTAT, es el conocimiento del esfuerzo que se realiza desde los distintos sectores de la economía en innovación, así como la obtención de una serie de indicadores que nos permitan comparar el nivel alcanzado en la Comunidad Autónoma de Euskadi (CAE) con el resto de países de su entorno.

La muestra de la EIT se caracteriza por ser un panel que cada año incluye a las empresas titulares que han contestado en anteriores repeticiones de la encuesta. Al igual que en el caso de la ESIE, el reparto original de la muestra se deteriora por diversas incidencias (altas, bajas, modificaciones,…), por lo que se actualiza la muestra conforme a un nuevo reparto muestral que, respetando el diseño original, recoge la nueva distribución de la población en los estratos.

En el año 2012, se decide renovar el panel en casi un 7%. Además, se introduce el Método del Cubo para seleccionar muestras equilibradas con el objetivo de obtener una distribución equilibrada en las comarcas de la CAE y sus capitales.

Ficha Técnica

- Marco

  Lo componen todos los establecimientos de cualquier sector de actividad que ejerza su actividad en el ámbito de la CAE, salvo el sector primario, la administración pública, las actividades asociativas, las actividades de los hogares y las actividades de organización y organismos extraterritoriales

- Diseño muestral

  Se trata de una muestra estratificada de una sola etapa.

  Unidades muestrales

  Todos los establecimientos que forman parte del marco mencionado.

  Estratificación

  Se realiza un muestro estratificado por el cruce de las siguientes variables:

  - Territorio Histórico

    1 = Araba; 2 = Bizkaia; 3 = Gipuzkoa

  - Estrato de empleo

    1 = 0-9 empleados; 2 = 10-49 empleados;

    3 = 50-249 empleados; 4 = 250 y más empleados

  - Sector de actividad (CNAE09 a 2 dígitos)

  Afijación

  Elementos autorrepresentados: establecimientos con 250 empleados y más (estrato de empleo 4) o establecimientos que correspondan a la actividad 46 en los estratos de empleo 2 y 3.

  Para el resto de los establecimientos se realiza la siguiente afijación teórica:

- Se reparten 2400 establecimientos para los estratos de 10 y más empleados y 750 establecimientos para los estratos de menos de 10 empleados.

- El reparto se realiza de manera proporcional a la raíz del nº de establecimientos por territorio y estrato de empleo, realizándose después otra afijación proporcional a la raíz del nº de establecimientos por estrato de actividad.

  Es decir, el tamaño de la muestra en cada estrato viene especificado por la siguiente fórmula:

$$n_{TH_i Emp_j Act_k} = n_{TH_i Emp_j} \frac{\sqrt{estab_{TH_i Emp_j Act_k}}}{\sum_{k \in Act} \sqrt{estab_{TH_i Emp_j Act_k}}} \qquad \begin{matrix} i \in \{01,20,48\} \\ j \in \{1,2,3\} \end{matrix}$$

donde

$$n_{TH_i Emp_j} = \begin{cases} 750 \dfrac{\sqrt{estab_{TH_i Emp_j}}}{\sum_{j=1} \sqrt{estab_{TH_i Emp_j}}} & \text{para empleo} < 10 \\ 2400 \dfrac{\sqrt{estab_{TH_i Emp_j}}}{\sum_{j \in 2,3} \sqrt{estab_{TH_i Emp_j}}} & \text{para empleo} > 10 \end{cases}$$

- Finalmente se añaden establecimientos hasta obtener un tamaño mínimo de 5 establecimientos en cada estrato.

Una vez calculados los tamaños teóricos necesarios por estrato, restamos las unidades que ya contiene el panel para obtener el número de unidades a extraer en cada estrato. Concretamente, en el año 2012 ha sido necesario extraer 771 establecimientos.

Sorteo

Se realiza un muestreo aleatorio simple en cada uno de los estratos, dando prioridad a los establecimientos que estén especificados en el marco como altas.

- Variables de equilibrio

Con el objetivo de obtener mejores estimaciones a nivel comarcal, la muestra correspondiente a los estratos empleo 2 y 3 (más de 10 empleados) ha sido equilibrada sobre el número de establecimientos en cada comarca (20 comarcas) y en las capitales.

- Sustitutos

Para completar la muestra se necesita una bolsa de sustitutos. Para ello, se extraerán 5 establecimientos en los estratos que no estén completos. En el año 2012 se han extraído 1.950 establecimientos reserva

Al igual que en la muestra principal, la muestra de sustitutos se equilibrará con el Método del Cubo sobre el número de establecimientos en cada comarca y las capitales.

Resultados

A continuación se muestran los resultados obtenidos con el Método del Cubo al equilibrar la el número de establecimientos por comarca y las capitales.

Distribución del número de establecimientos por comarca y capitales (más de 10 empleados)

	Poblacional	Muestral (elevado)
Valles Alaveses	50 (0.40 %)	64 (0.51 %)
Llanada Alavesa (sin capital)	102 (0.81 %)	69 (0.54 %)
Montaña Alavesa	14 (0.11 %)	19 (0.15 %)
Rioja Alavesa	105 (0.83 %)	93 (0.74 %)
Estribaciones del Gorbea	97 (0.77 %)	156 (1.23 %)
Cantábrica Alavesa	185 (1.47 %)	234 (1.86 %)
Arratia - Nervión	135 (1.07 %)	114 (0.91%)
Gran Bilbao (sin capital)	2.931 (23.26 %)	2.597 (20.61 %)
Durangaldea	648 (5.14 %)	556 (4.41 %)
Encartaciones	111 (0.88 %)	217 (1.72 %)
Gernika – Bermeo	162 (1.29 %)	271 (2.15 %)
Markina – Ondarroa	103 (0.82 %)	192 (1.52 %)
Plentzia – Mungia	200 (1.59 %)	333 (2.64 %)
Bajo Bidasoa	373 (2.96 %)	385 (3.06 %)
Bajo Deba	359 (2.85 %)	290 (2.30 %)
Alto Deba	366 (2.90%)	490 (3.88 %)
Donostialdea (sin capital)	910 (7.22 %)	841 (6.67 %)
Goierri	334 (2.65 %)	387 (3.07 %)

Tolosaldea	311 (2.47 %)	419 (3.32 %)
Urola Costa	390 (3.09 %)	263 (2.09 %)
Vitoria-Gasteiz	1.548 (12.28 %)	1.467 (11.64 %)
Bilbao	1.979 (15.70 %)	1.988 (15.78 %)
Donostia-San Sebastian	1.190 (9.44 %)	1.158 (9.19 %)
TOTAL	**12.603**	**12.603**

Los porcentajes están dados por columnas

- Notas:

1. Para el cálculo de los elevadores del número de establecimientos por comarca, se ha hecho una post-estratificación, agrupado los estratos de actividad en función de la agregación sectorial A38 (CNAE09), puesto que es la que se utiliza en difusión.

2. En las tres capitales, se han obtenido muy buenas estimaciones del número de establecimientos.

3. En lo que al resto de comarcas se refieren, pese a que la mayoría de ellas están bastante bien estimadas, podemos encontrar comarcas con un alto error relativo como Estribaciones del Gorbea, Encartaciones, Gernika-Bermeo, Markina-Ondarroa, Plentzia-Mungia, Tolosaldea o Urola-Costa.

4. En estas 7 comarcas el Método del Cubo no ha logrado un solución muestral que obtenga mejores resultados debido a las restricciones impuestas por el mismo diseño:

   - Pese a que el tamaño de la muestra era de unos 2.900 establecimientos, solo se han sorteado 410, puesto que el resto provenían tanto del panel como de estratos censales.

   - Además, de los 401 estratos definidos por el cruce de Territorio, actividad y empleo, solamente se seleccionan establecimientos en 173 estratos.

   - Finalmente, de los 173 estratos en donde realmente se realiza el sorteo, en 21 de ellos el establecimiento a seleccionar está determinado a priori (por tener que dar prioridad a las altas).

Muestra para la Encuesta de Pobreza y Desigualdades Sociales (EPDS)

La Encuesta de Pobreza y Desigualdades Sociales (EPDS), tiene una alta importancia para el Departamento de Justicia, Empleo y Seguridad Social, al vincularse a la evaluación y programación de sus prestaciones económicas. Por esa razón, resulta de especial importancia consolidar un diseño muestral que permita un acercamiento lo más correcto posible al colectivo de encuestación.

De forma general, el objetivo central de la EPDS es el conocimiento, estudio y evaluación de las distintas líneas de pobreza, y de su incidencia en Euskadi, así como de indicadores asociados de desigualdad social.

En el año 2012, se decide seleccionar la muestra para la EPDS utilizando el Método del Cubo. De esta manera, hemos logrado obtener una muestra equilibrada por sexo, edad y nacionalidad, además del tamaño familiar en cada uno de los Territorios Históricos.

Ficha Técnica

- Marco

  El marco de la muestra de la Encuesta de Pobreza y Desigualdades Sociales lo componen las viviendas familiares ocupadas de la Comunidad Autónoma de Euskadi y sus territorios históricos.

- Diseño muestral

  Se trata de una muestra bietápica con estratificación en la primera etapa y tamaño de la muestra fija en la segunda.

  Unidades muestrales

  Viviendas familiares ocupadas de la CAE.

  Tamaño de la muestra

  Se seleccionan alrededor de 4.000 unidades de encuestación, aportándose unas 8.000 unidades sustitutas (dos sustitutos por unidad muestral).

  Primera etapa: Muestra de secciones

  En la primera etapa se realiza un sorteo de las secciones censales de la CAE.

  - **Estratificación**

    Las unidades de la primera etapa se estratifican por el cruce de las siguientes variables:

    - Comarcas y cuadrillas

      01 = Añana; 02 = Ayala/Aiara; 03 = Campezo-Montaña Alavesa;
      04 = Laguardia-Rioja Alavesa; 05 = Salvatierra/Agurain;
      06 = Vitoria-Gasteiz; 07 = Zuia; 08 = Donostialdea;
      09 = Tolosaldea-Goierri; 10 = Alto-Deba; 11 = Bajo-Deba;
      12 = Margen Derecha; 13 = Bilbao; 14 = Margen Izquierda;
      15 = Bizkaia Costa; 16 = Duranguesado

- Tipologías

  Se realiza un análisis de las tipologías de las secciones censales de Eustat, específico para la EPDS. Para ello, se tienen en cuenta las variables básicas: edad, sexo, nacionalidad, relación con la actividad, nº de residentes en la vivienda y renta personal y familiar media.

  Una vez realizado un Análisis de Componentes Principales, las secciones son clasificadas en 7 tipologías.

- Predominio personas jóvenes:

  Con el objetivo de sobrerrepresentar la muestra en aquellas secciones caracterizadas por una fuerte presencia relativa de personas menores de 45 años, se realiza una clasificación de las secciones en dos grupos:

  1 = Secciones con predominio de jóvenes

  0 = Resto

  En la segunda etapa, se sortearán 24 viviendas en las secciones "jóvenes" y 16 viviendas en el resto.

o **Afijación**

El sorteo de las 4000 viviendas se ha realizado de acuerdo a las siguientes afijaciones:

1. Reparto proporcional a la raíz cuadrada del nº de viviendas por Territorio Histórico
2. Reparto proporcional a la raíz cúbica del nº de viviendas por comarcas/cuadrillas
3. Reparto proporcional al nº de viviendas por tipología y tipo de sección("joven"/"no-joven")

Se exigen un tamaño mínimo de 160 viviendas por comarca y 112 viviendas en las cuadrillas de Álava.

o **Sorteo**

El sorteo de las secciones ha sido probabilística y proporcional al tamaño (PPT), medido en número de viviendas ocupadas.

Segunda etapa: Muestra de viviendas

o **Afijación**

Para cada sección seleccionada en la primera etapa de la muestra, se seleccionan 16 o 24 viviendas en función del tipo de sección que se trate.

o **Sorteo**

Se realiza un sorteo aleatorio simple dentro de cada sección seleccionada en la primera etapa.

- Variables de equilibrio

La muestra ha sido equilibrada sobre las mismas variables tanto el la primera etapa como en la segunda. Con ello, aseguramos que la muestra final esté equilibrada sobre el marco de viviendas completo.

Las variables equilibradas son las siguientes:

- *Tamaño familiar*: Número de viviendas con 1 residente, 2 residentes, 3-4 residentes o más de 5 residentes por TH

- *Sexo*: Número de hombres y mujeres por TH.

- *Edad:* Número de individuos con menos de 34 años, entre 35-44 años, 45-54 años y más de 65 años por TH.

- *Nacionalidad*: Número de individuos nacionales y extranjeros por TH.

- *Número de individuos* en cada comarca/cuadrillas.

- Sustitutos

  Para completar la muestra se sortean un suplente y un reserva para cada una de las viviendas. Estos sustitutos han sido extraídos en cada una de las secciones censales seleccionadas en la primera etapa, equilibrando la muestra sobre las mismas variables que las viviendas titulares.

Resultados

A continuación se muestran los resultados obtenidos con el Método del Cubo para las variables de equilibrio.

Cada una de las tablas compara la distribución poblacional con la obtenida a partir de los elevadores de la muestra. Los porcentajes están dados por columnas

Distribución de las viviendas por Tamaño Familiar y Territorio

	Araba		Gipuzkoa		Bizkaia	
	Poblacional	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**
1 residente	35.528 (27,77%)	35.440 (27,70%)	68.232 (24,97%)	68.553 (25,09%)	109.535 (24,44%)	112.675 (25,14%)
2 residentes	37.537 (29,34%)	38.174 (29,84%)	78.075 (28,57%)	78.039 (28,56%)	130.825 (29,18%)	130.322 (29,07%)
3 - 4 residentes	47.391 (37,04%)	47.735 (37,31%)	108.714 (39,78%)	108.381 (39,66%)	180.827 (40,34%)	178.194 (39,75%)
Más de 5 residentes	7.485 (5,85%)	6.592 (5,15%)	18.248 (6,68%)	18.295 (6,69%)	27.079 (6,04%)	27.075 (6,04%)
TOTAL	**127.941**	**127.941**	**273.269**	**273.269**	**448.266**	**448.266**

Distribución por Sexo y Territorio Histórico

	Araba		Gipuzkoa		Bizkaia	
	Poblacional	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**
Hombres	157.836 (49,91%)	155.759 (49,63%)	344.561 (49,02%)	347.363 (49,48%)	553.674 (48,49%)	551.028 (48,53%)
Mujeres	158.392 (50,09%)	158.111 (50,37%)	358.350 (50,98%)	354.687 (50,52%)	588.197 (51,51%)	584.492 (51,47%)
TOTAL	**316.228**	**313.870**	**702.911**	**702.050**	**1.141.871**	**1.135.521**

Distribución por Edad y Territorio Histórico

	Araba		Gipuzkoa		Bizkaia	
	Poblacional	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**
Menos de 34 años	108.383 (34,27%)	109.676 (34,94%)	233.423 (33,21%)	234.644 (33,42%)	366.085 (32,06%)	363.674 (32,03%)
35 - 44 años	55.227 (17,46%)	49.691 (15,83%)	116.445 (16,57%)	116.922 (16,65%)	188.762 (16,53%)	194.045 (17,09%)
45 - 54 años	49.799 (15,75%)	49.939 (15,91%)	109.078 (15,52%)	107.384 (15,30%)	182.531 (15,99%)	179.632 (15,82%)
55 - 64 años	40.810 (12,91%)	43.836 (13,97%)	92.261 (13,13%)	91.599 (13,05%)	151.434 (13,26%)	146.342 (12,89%)
Más de 65 años	62.009 (19,61%)	60.729 (19,35%)	151.704 (21,58%)	151.501 (21,58%)	253.059 (22,16%)	251.828 (22,18%)
TOTAL	**316.228**	**313.870**	**702.911**	**702.050**	**1.141.871**	**1.135.521**

Distribución por Nacionalidad y Territorio Histórico

	Araba		Gipuzkoa		Bizkaia	
	Poblacional	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**
Nacional	286.633 (90,64%)	289.847 (92,35%)	658.599 (93,70%)	659.521 (93,94%)	1.067.272 (93,47%)	1.059.925 (93,34%)
Extranjero	29.595 (9,36%)	24.023 (7,65%)	44.312 (6,30%)	42.529 (6,06%)	74.599 (6,53%)	75.595 (6,66%)
TOTAL	**316.228**	**313.870**	**702.911**	**702.050**	**1.141.871**	**1.135.521**

Distribución del número de individuos por comarca/cuadrilla

	Poblacional	Muestral elevado
Añana	8.617 (0,40%)	8.350 (0,39%)
Ayala / Aiara	34.208 (1,58%)	33.894 (1,58%)
Campezo - Montaña Alavesa	3.156 (0,15%)	3.118 (0,14%)
Laguardia - Rioja Alavesa	11.414 (0,53%)	11.181 (0,52%)
Salvatierra/Agurain	12.255 (0,57%)	12.384 (0,58%)
Vitoria - Gasteiz	237.059 (10,97%)	235.576 (10,95%)
Zuia	9.519 (0,44%)	9.368 (0,44%)
Donostialdea	472.708 (21,87%)	472.950 (21,98%)
Tolosaldea - Goierri	114.584 (5,30%)	113.420 (5,27%)
Alto Deba	60.919 (2,82%)	60.945 (2,83%)
Bajo Deba	54.700 (2,53%)	54.734 (2,54%)
Margen Derecha	161.425 (7,47%)	157.625 (7,33%)
Bilbao	349.132 (16,16%)	348.884 (16,22%)
Margen Izquierda	386.068 (17,87%)	379.912 (17,66%)
Bizkaia Costa	126.504 (5,85%)	127.321 (5,92%)
Duranguesado	118.742 (5,49%)	121.778 (5,66%)
TOTAL	**2.161.010**	**2.151.441**

Muestra para el estudio de las mujeres en el ámbito rural vasco

El Departamento de Agricultura, Pesca y Alimentación quiere actualizar el estudio que se viene realizando desde 1998 sobre "La mujer en el ámbito rural vasco. Necesidades, demandas y carencias sociales."

En el año 2012, a diferencia de diseños anteriores, se va extraer una muestra de mujeres y otra de hombres de 15 y más años que residen en los municipios que el departamento ha señalado como rurales, por criterios de tamaño, densidad de población y proporción de PIB agrario. La muestra deberá estar compuesta por 250 hombres y 250 mujeres en cada uno de los Territorios Históricos de la CAE.

Además de esto, se decide seleccionar la muestra utilizando el Método del Cubo, obteniendo una muestra equilibrada de hombres y mujeres por edad, nacionalidad, nivel de estudios y tipo de vivienda (núcleo o diseminado) en cada uno de los TH.

Ficha Técnica

- Marco

  El marco de la muestra lo componen la población de 15 años y más, que residen en viviendas familiares de los 128 municipios señalados como rurales por el Departamento de Agricultura, Pesca y Alimentación.

- Diseño muestral

  Como el objetivo es obtener una muestra de mujeres y otra de hombres de igual tamaño en los municipios rurales, se ha optado por realizar una muestra bietápica con estratificación en la primera etapa. Las afijaciones de la primera y segunda etapa se calculan de modo que la muestra final de individuos es autoponderada por Territorio Histórico.

  De esta manera, una vez sorteados los municipios rurales, se sortearán el mismo número de hombres y mujeres dentro de cada municipio.

  Tamaño de la muestra

  Se seleccionan alrededor de 250 hombres y 250 mujeres en cada Territorio Histórico de la CAE. No se seleccionarán sustitutos, puesto que se ha optado por realizar una sobremuestra teniendo en cuenta la tasa de no respuesta estimada (46% en cada uno de los TH).

  Primera etapa: Muestra de municipios

  En la primera etapa se realiza un sorteo estratificado de los 128 municipios rurales de la CAE.

  - **Unidades muestrales**

    Municipios rurales de la CAE. Se trata de conglomerados de individuos de tamaños distintos.

  - **Estratificación**

    Las unidades de la primera etapa se estratifican por:

- Territorio Histórico

  01 = Araba; 20 = Gipuzkoa; 48 = Bizkaia

- Tamaño de los municipios

  La estratificación por tamaño de los municipios es óptima, es decir, minimiza la inercia intra-clase o varianza interna de cada estrato, tomando como referencia la inercia o varianza total.

  1 = [0-569]; 2 = [570-1154]; 3 = [1155-1884]; 4 = [1885-3400]

o **Afijación**

El objetivo final es sortear 250 hombres y 250 mujeres en cada uno de lo TH.

No se seleccionarán sustitutos, puesto que se ha optado por realizar un sobremuestra, teniendo en cuenta la tasa de no respuesta estimada (46% en cada uno de los TH).

Para calcular el número de municipios a sortear en cada estrato, se ha seguido el siguiente procedimiento:

1. Reparto proporcional al tamaño de los estratos (población) de 500 individuos por cada Territorio.
2. Se calcula el nº de municipios a sortear en cada TH, a partir de un múltiplo de la fracción de muestreo de la población.
3. Reparto proporcional al nº de municipios por estrato.
4. Se amplia la muestra de municipios para seleccionar aquellos que pertenezcan al estrato de tamaño igual a 4.

o **Sorteo**

Una vez obtenido el reparto teórico, el sorteo de los municipios rurales se realiza mediante muestreo aleatorio simple.

Segunda etapa: Muestra de hombres y mujeres

En la segunda etapa, debemos seleccionar los hombres y mujeres que van a ser encuestados.

o **Unidades muestrales**

Hombres y mujeres mayores de 15 años pertenecientes a los municipios rurales seleccionados en la primera etapa.

o **Afijación**

Para cada municipio rural seleccionado en la primera etapa de la muestra, se calcula en número de hombres y mujeres a sortear de manera proporcional al tamaño del municipio dentro del estrato. Es decir,

$$n_{MUN_i} = n_h \frac{Pob_{MUN_i}}{Pob_h}$$

donde MUN_i son aquellos municipios rurales seleccionados en la primera etapa y *h* el estrato correspondiente a dicho municipio.

o **Sorteo**

Se extraen dos muestras aleatorios simples e independientes dentro de las subpoblaciones de hombres y mujeres de cada municipio.

La muestra final es aproximadamente autoponderada por Territorios Históricos.

- Variables de equilibrio

  La muestra ha sido equilibrada sobre las mismas variables tanto el la primera etapa como en la segunda. Con ello, aseguramos que la muestra final esté equilibrada sobre el marco de individuos completo.

  Las variables equilibradas son las siguientes:

  - *Sexo*: Número de hombres y mujeres por TH.
  - *Edad*: Número de individuos entre 15-25 años, 26-39 años, 40-54 años, 55-64 años y más de 65 años por TH.
  - Nacionalidad: Número de individuos nacionales y extranjeros por TH
  - *Estudios:* Número de individuos con estudios primarios, medios o superiores por TH
  - *Tipo de vivienda*: Número de individuos residentes en viviendas de tipo núcleo o diseminado

Resultados

A continuación se muestran los resultados obtenidos con el Método del Cubo para las variables de equilibrio.

Cada una de las tablas compara la distribución poblacional con la obtenida a partir de los elevadores de la muestra. Los porcentajes están dados por columnas

Distribución por Edad y Territorio Histórico

SEXO = HOMBRES

	Araba		Gipuzkoa		Bizkaia	
	Poblacional	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**
15 - 25 años	1.705 *(9,70%)*	1.676 *(9,53%)*	1.231 *(10,41%)*	1.236 *(10,45%)*	1.769 *(8,90%)*	1.807 *(9,09%)*
26 - 39 años	3.706 *(21,08%)*	3.634 *(20,67%)*	2.958 *(25,01%)*	2.988 *(25,26%)*	4.354 *(21,91%)*	4.383 *(22,06%)*
40 - 54 años	5.746 *(32,68%)*	5.807 *(33,03%)*	3.396 *(28,71%)*	3.320 *(28,07%)*	6.169 *(31,05%)*	6.260 *(31,51%)*
55 - 64 años	2.698 *(15,35%)*	2.730 *(15,53%)*	1.802 *(15,23%)*	1.809 *(15,29%)*	3.191 *(16,06%)*	3.050 *(15,35%)*
Más de 65 años	3.727 *(21,20%)*	3.734 *(21,24%)*	2.442 *(20,64%)*	2.476 *(20,93%)*	4.386 *(22,07%)*	4.369 *(21,99%)*
TOTAL	**17.582**	**17.852**	**11.829**	**11.829**	**19.869**	**19.869**

SEXO = MUJERES

	Araba		Gipuzkoa		Bizkaia	
	Poblacional	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**
15 - 25 años	1.552 *(9,91%)*	1.624 *(10,37 %)*	1.164 *(10,73%)*	1.133 *(10,45%)*	1.716 *(8,99%)*	1.655 *(8,67%)*
26 - 39 años	3.351 *(21,39%)*	3.309 *(21,12%)*	2.709 *(24,98%)*	2.658 *(24,51 %)*	3.970 *(20,81%)*	4.058 *(21,27%)*
40 - 54 años	4.694 *(29,96%)*	4.749 *(30,31%)*	2.880 *(26,56%)*	2.870 *(26,47%)*	5.398 *(28,29%)*	5.403 *(28,32%)*
55 - 64 años	2.133 *(13,61%)*	2.067 *(13,19%)*	1.416 *(13,06%)*	1.481 *(13,66%)*	2.714 *(14,23%)*	2.708 *(14,19%)*
Más de 65 años	3.938 *(25,13%)*	3.918 *(25,01%)*	2.675 *(24,67%)*	2.703 *(24,93 %)*	5.281 *(27,68%)*	5.255 *(27,54%)*
TOTAL	**15.668**	**15.668**	**10.844**	**10.844**	**19.079**	**19.079**

Distribución por Nacionalidad y Territorio Histórico

SEXO = HOMBRES

	Araba		Gipuzkoa		Bizkaia	
	Poblacional	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**
Nacional	16.410 *(93,33%)*	16.403 *(93,29%)*	11.182 *(94.53%)*	11.218 *(94,83%)*	19.037 *(95,81%)*	19.000 *(95,63%)*
Extranjero	1.172 *(6,67%)*	1.179 *(6,71%)*	647 *(5,47%)*	611 *(5,17%)*	832 *(4,19%)*	869 *(4,37%)*
TOTAL	**17.582**	**17.852**	**11.829**	**11.829**	**19.869**	**19.869**

SEXO = MUJERES

	Araba		Gipuzkoa		Bizkaia	
	Poblacional	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**
Nacional	14.694 *(93,78%)*	14.673 *(93,65%)*	10.300 *(94,98%)*	10.278 *(94,78%)*	18.270 *(95,76%)*	18.251 *(95,66%)*
Extranjero	974 *(6,22%)*	995 *(6,35%)*	544 *(5,02%)*	566 *(5,22%)*	809 *(4,24%)*	828 *(4,34%)*
TOTAL	**15.668**	**15.668**	**10.844**	**10.844**	**19.079**	**19.079**

Distribución por Nivel de Estudios y Territorio Histórico

SEXO = HOMBRES

	Araba		Gipuzkoa		Bizkaia	
	Poblacional	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**
Estudios Primarios	7.304 *(41,54%)*	7.225 *(41,09%)*	5.287 *(44,70%)*	5.144 *(43,49 %)*	6.873 *(34,59%)*	6.813 *(34,29%)*
Estudios Medios	7.616 *(43,32%)*	7.630 *(43,40%)*	4.957 *(41,91%)*	5.123 *(43,41%)*	8.798 *(44,28%)*	8.915 *(44,87%)*
Estudios Superiores	2.662 *(15,14%)*	2.727 *(15,51%)*	1.585 *(13,40%)*	1.562 *(13,20%)*	4.198 *(21,13%)*	4.141 *(20,84%)*
TOTAL	**17.582**	**17.852**	**11.829**	**11.829**	**19.869**	**19.869**

SEXO = MUJERES

	Araba		Gipuzkoa		Bizkaia	
	Poblacional	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**
Estudios Primarios	6.774 *(43,23%)*	6.665 *(42,54%)*	4.928 *(45,44%)*	4.922 *(45,39%)*	7.587 *(39,77%)*	7.586 *(39,76%)*
Estudios Medios	5.459 *(34,84%)*	5.557 *(35,47 %)*	3.451 *(31,82%)*	3.441 *(31,73 %)*	6.148 *(32,22%)*	6.160 *(32,29%)*
Estudios Superiores	3.435 *(21,92%)*	3.446 *(21,99%)*	2.465 *(22,73%)*	2.482 *(22,89%)*	5.344 *(28,01%)*	5.333 *(27,95%)*
TOTAL	**15.668**	**15.668**	**10.844**	**10.844**	**19.079**	**19.079**

Distribución por Tipo de Vivienda y Territorio Histórico

SEXO = HOMBRES

	Araba		Gipuzkoa		Bizkaia	
	Poblacional	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**
Núcleo	16.555 *(94,16%)*	16.743 *(95,23%)*	7.530 *(63,66%)*	7.891 *(66,71%)*	11.750 *(59,14%)*	12.245 *(61,63 %)*
Diseminado	1.027 *(5,84%)*	839 *(4,77%)*	4.299 *(36,34%)*	3.938 *(33,29%)*	8.119 *(40,86%)*	7.624 *(38,37%)*
TOTAL	**17.582**	**17.852**	**11.829**	**11.829**	**19.869**	**19.869**

SEXO = MUJERES

	Araba		Gipuzkoa		Bizkaia	
	Poblacional	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**	**Poblacional**	**Muestral (elevado)**
Núcleo	14.781 *(94,34%)*	14.977 *(95,59%)*	7.223 *(66,61%)*	7.687 *(70,89%)*	11.555 *(60,56%)*	12.072 *(63,27%)*
Diseminado	887 *(5,66%)*	691 *(4,41%)*	3.621 *(33,39%)*	3.157 *(29,11%)*	7.524 *(39,44%)*	7.007 *(36,73%)*
TOTAL	**15.668**	**15.668**	**10.844**	**10.844**	**19.079**	**19.079**

Muestra para la Encuesta de Euskadi y Drogas

Euskadi y Drogas es una encuesta de periodicidad bienal, orientada a conocer los consumos de diversas sustancias por parte de la población vasca de 15 a 74 años de edad, así como su percepción respecto a diversas cuestiones relacionadas con las drogas y las drogodependencias.

En el año 2012, se decide seleccionar la muestra utilizando el Método del Cubo. De esta manera, se ha obtenido una muestra equilibrada del total de individuos por comarcas sanitarias, tamaños de municipios, sexo y nacionalidad.

Ficha Técnica

- Marco

  El marco de la muestra lo componen la población de 15 a 74 años de edad residentes en viviendas familiares de la Comunidad Autónoma de Euskadi y sus territorios históricos.

- Diseño muestral

  Se trata de una muestra estratificada de una sola etapa.

  Unidades muestrales

  Población entre 15 y 74 años (fecha de referencia: 15 de julio de 2012) residentes en viviendas familiares de la Comunidad Autónoma de Euskadi.

  Tamaño de la muestra

  Según las especificaciones de la operación, se seleccionarán n=2007 individuos titulares; y otros tantos suplentes y reservas.

  Estratificación

  Se realiza un muestreo estratificado por el cruce de las siguientes variables:

- Territorio Histórico

  01 = Araba; 20 = Gipuzkoa; 48 = Bizkaia

- Grupos de edad:

  6 grupos de edades decenales

  (15-24, 25-34, 35-44, 45-54, 55-64 y 65-74 años)

Afijación

Se ha establecido un criterio para cada uno de los niveles de estratificación:

1. Reparto proporcional a la raíz cuadrada del nº de individuos por Territorio
2. Para cada Territorio, afijación del tamaño doble para los grupos de edad más jóvenes (15-24 años, 25-34 años y 35-44 años).

Sorteo

Una vez obtenido el reparto teórico, se realiza un muestreo aleatorio simple en cada estrato.

- Variables de equilibrio

La muestra ha sido equilibrada sobre las siguientes variables:

- Número de individuos de 15 a 74 años en cada uno de las 11 comarcas sanitarias de la CAE: Alava, Gipuzkoa Oeste, Gipuzkoa Este, (Biz) Interior, (Biz) Ezkerraldea-Enkarterri, (Biz) Uribe y (Biz) Bilbao.
- Número de individuos de 15 a 74 años en los municipios, según su tamaño en población: Capitales, entre 50.000 y 100.000 habitantes, entre 25.000-50.000 habitantes, entre 10.000-25.000 habitantes y hasta 10.000 habitantes.
- Número de individuos por sexo.
- Número de individuos con nacionalidad española y extranjera.

- Sustitutos

Para completar la muestra, se necesitan dos bolsas de unidades sustitutas: una de suplentes y otra de reservas, ambas de 2007 unidades en cada caso.

Estas unidades sustitutas se extraerán respetando el mismo reparto muestral por estratos utilizado en la muestra original, equilibrando la muestra sobre las mismas variables que los titulares.

Resultados

A continuación se muestran los resultados obtenidos con el Método del Cubo para las variables de equilibrio.

Cada una de las tablas compara la distribución poblacional con la obtenida a partir de los elevadores de la muestra. Los porcentajes están dados por columnas:

Distribución del número de individuos por Comarca Sanitaria

	Poblacional	Muestral elevado
Alava	219.042 (13,28%)	218.966 (13,28%)
Gipuzkoa Oeste	218.155 (13,23%)	218.335 (13,24%)
Gipuzkoa Este	328.814 (19,94%)	329.009 (19,95%)
(Biz) Interior	227.787 (13,81%)	228.032 (13,83%)
(Biz) Ezkerraldea-Enkarterria	225.829 (13,70%)	224.429 (13,61%)
(Biz) Uribe	166.287 (10,08%)	166.029 (10,07%)
(Biz) Bilbao	263.028 (15,95%)	264.141 (16,02%)
TOTAL	**1.648.942**	**1.648.942**

Distribución del número de individuos por Tamaño de municipio

	Poblacional	Muestral elevado
Capitales	587.948 (35,66%)	589.033 (35,72%)
De 50.000 a 100.000	184.970 (11,22%)	184.638 (11,20%)
De 25.000 a 50.000	239.465 (14,52%)	239.354 (14,52%)
De 10.000 a 25.000	300.173 (18,20%)	300.088 (18,20%)
Hasta 10.000	336.386 (20,40%)	335.829 (20,37%)
TOTAL	**1.648.942**	**1.648.942**

Distribución del número de individuos por Sexo

	Poblacional	Muestral elevado
Hombres	823.310 (49,93%)	823.742 (49,96%)
Mujeres	825.632 (50,07%)	825.200 (50,04%)
TOTAL	**1.648.942**	**1.648.942**

Distribución del número de individuos por Nacionalidad

	Poblacional	Muestral elevado
Nacional	1.519.906 (92,17%)	1.518.872 (92,11%)
Extranjero	129.036 (7,83%)	130.070 (7,89%)
TOTAL	**1.648.942**	**1.648.942**

8. Conclusiones

Por último, vamos a mencionar ciertas conclusiones relativas al interés de realizar muestreos equilibrados, la elección de las variables de equilibrio y la relación del equilibrio con la estratificación y calibración.

Equilibrio y estratificación

Tanto para la estratificación como para el equilibrio, necesitamos conocer el valor de las variables auxiliares para todas las unidades de la población.

La mayor ventaja de la estratificación, es que nos permite dividir la población en subpoblaciones más homogéneas obteniendo estimadores más precisos, reduciendo la varianza de muestreo. La estratificación es tanto mejor cuantas más variables correlacionadas con la variable de interés intervengan.

Aún así, el utilizar demasiadas variables de estratificación, puede producir estratos demasiado pequeños, en donde el tamaño muestral no es suficiente; sin mencionar los problemas que pueda acarrear la no respuesta en dichos estratos, aunque esto se pueda arreglar mediante el colapso de estratos (post-estratificación).

Las variables de equilibrio, permiten que aquellas variables que no puedan entrar en la estratificación múltiple se añadan como variable de equilibrio, manteniendo todas las ventajas de la estratificación en lo que a la reducción de la varianza se refiere y añadiendo las ventajas propias del equilibrio.

Permiten también, trabajar en dominios definidos sobre el cruce varios estratos o áreas pequeñas.

Las variables de equilibrio pueden ser cuantitativas, mientras que las variables de estratificación siempre han de ser cualitativas o categóricas.

Elección de las variables de equilibrio

Las variables auxiliares escogidas para equilibrar la muestra, deben estar muy correlacionadas con las variables de interés y no demasiado correlacionadas entre ellas.

Al equilibrar la muestra sobre un gran número de variables auxiliares cualitativas, se obtienen totales estimados (o medias estimadas) con distribuciones prácticamente iguales a las de la población de origen.

El Método del Cubo, es muy interesante para la selección de las unidades primarias en una muestra multietápica. En el caso de seleccionar también una muestra equilibrada en la segunda etapa, las variables a equilibrar deben de haber sido equilibradas en la primera etapa previamente.

Equilibrio y calibración

A diferencia del equilibrio y la estratificación, para la calibración solo debemos conocer el valor de las variables auxiliares para los elementos de la muestra, así como los totales de estas variables en la población.

La mejor estrategia es usar equilibrio y calibración juntos (ver la simulación en *Deville and Tillé*, 2004), puesto que en general, se obtienen mejores resultados si calibramos una muestra sobre las mismas variables auxiliares utilizadas en el equilibrio.

Hay un caso en el que la calibración se puede utilizar sobre variables distintas a las de equilibrio: cuando se tratan de la misma variable medida en diferentes momentos.

Análisis de los resultados

A continuación, se van a mostrar los resultados obtenidos a la hora de calibrar dos muestras que previamente han sido equilibradas con el Método del Cubo (*Euskadi y Drogas 2012* y *Encuesta de Capital Social 2012)*.

En ambos casos, la calibración ha sido realizada a traves de la macro CALMAR (*calage sur marges*), "reponderando" los pesos muestrales de los individuos de la muestra para ajustarlos a los totales marginales de las variables auxiliares de calibración.

1. Calibración de la encuesta de Euskadi y Drogas 2012

Para la encuesta de Euskadi y Drogas 2012 (n=2007 individuos), se ha decidido calibrar la muestra sobre las siguientes variables:

- Cruce de las variables Territorio Histórico y Edad (variables de estratificación)
- Comarca sanitaria, tamaño de municipio y sexo (variables de equilibrio)

Partiendo de los pesos iniciales $w_{hi} = w_h \ \forall i$ (pesos iguales dentro de cada estrato), se han obtenido los pesos finales w_{hi}^* utilizando la macro CALMAR con el método ranking ratio para ajustar las estimaciones a los totales marginales de las variables de calibración.

Se define la variable $f = \frac{w_{hi}^*}{w_{hi}}$ como la razón entre los pesos finales y los pesos iniciales.

Analizando la distribución de esta variable, podemos determinar cuanto se han "deformado" los pesos iniciales para ajustarse a los totales marginales de las variables de calibración.

Este es un pequeño resumen de la distribución de la variable *f*:

Media	1
Mediana	0.9987
Moda	0.9978
Desviación estándar	0.0875
Coeficiente de variación	**8.75%**
Mínimo	0.8365
Máximo	1.2484

Como se puede observar, los pesos finales no están demasiado alejados de los pesos iniciales (incremento máximo del 24% y decremento máximo del 16%), manteniendo en buena medida los pesos de las unidades muestrales asociados a la estratificación.

2. Calibración de la Encuesta de Capital Social 2012

Para la Encuesta de Capital Social 2012 (n=4000 individuos), se ha decidido calibrar la muestra sobre el cruce de las siguientes variables:

- Territorio Histórico (Araba, Gipuzkoa y Bizkaia)
- Sexo (hombres y mujeres)
- Edad (15-24, 25-34, 35-44, 45-54, 55-64 y más de 65 años)

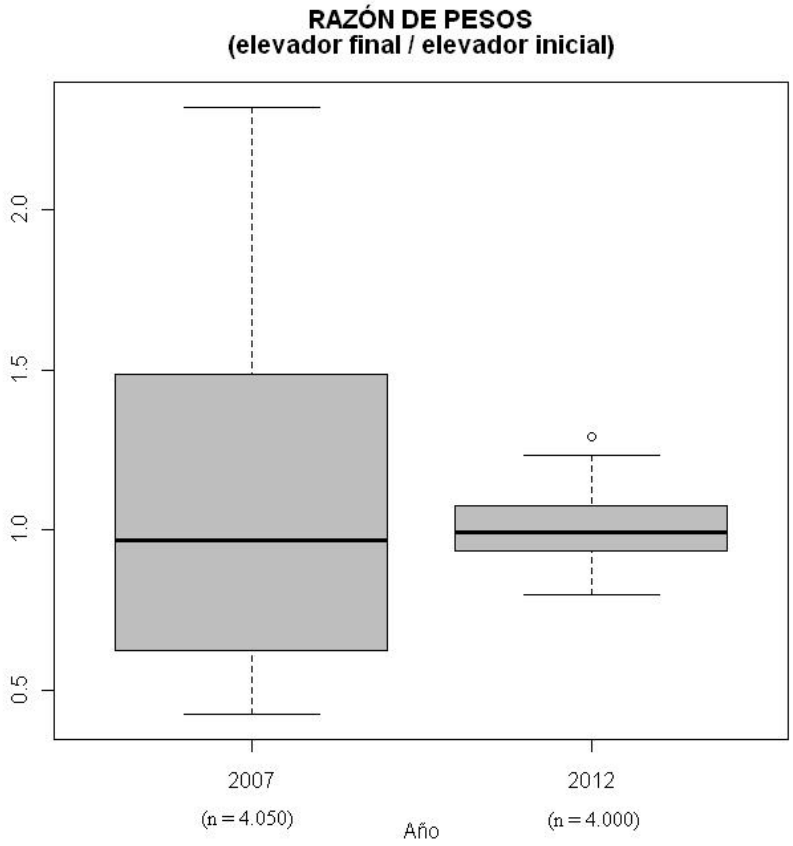
Por lo tanto, la muestra ha sido calibrada sobre 36 totales marginales.

Al igual que en el ejemplo anterior, se define la variable $f = \frac{w_{hi}^{*}}{w_{hi}}$ como la razón entre los pesos finales y los pesos iniciales. Los pesos finales w_{hi}^{*} han sido obtenidos utilizando la macro CALMAR con el método ranking ratio para ajustar las estimaciones a los totales marginales de las variables de calibración.

En esta ocasión, no solo se va a analizar la distribución de la variable *f*, sino que la vamos a comparar con los valores obtenidos para la Encuesta de Capital Social 2007.

Debemos recordar, que pese a que ambas encuestas tienen el mismo diseño muestral, la ECS 2012 ha sido seleccionada equilibrando la muestra con el Método del Cubo. Las variables de equilibrio utilizadas, han sido precisamente las mismas que las variables de calibración.

A continuación se muestran los resultados obtenidos para los años 2007 y 2012:



	2007	*2012*
Media	1.1139	1.0074
Mediana	0.9685	0.9944
Moda	2.0076	1.0287
Desviación estándar	0.5306	0.1125
Coeficiente de variación	**47.63%**	**11.17%**
Mínimo	0.4223	0.7965
Máximo	2.3236	1.2915

Al haber equilibrado la muestra de la ECS 2012 sobre las variables de calibración, hemos obtenido unos mejores resultados, obteniendo pesos finales mucho menos alejados que los obtenidos en la ECS 2007 (incremento máximo del 29% frente al 132% y un decremento máximo del 20% frente al 58%).

Interés del muestreo equilibrado

En el marco asistido por el modelo y basado sobre el modelo, un diseño de muestreo equilibrado con el estimador de Horvitz-Thompson es a menudo la estrategia óptima (ver Nedyalkova and Tillé, 2009). En realidad, cuando una muestra es totalmente equilibrada, las varianzas de los estimadores de H-T de las variables auxiliares son iguales a cero.

Las ventajas del muestreo equilibrado son las siguientes:

- Se trata de una optimización de diseños muestrales probabilísticos, sean unietapicos o multietapicos, donde las probabilidades de inclusión definidas por el diseño son la clave de partida para seleccionar muestras aleatorias.

- Aumenta la exactitud del estimador de H-T; es mas, la varianza del estimador sólo depende de la correlación entre las variables de interés y las variables de equilibrio (residuos de la regresión).

- Las muestras más desfavorables, extremas o lejanas a la media tienen una probabilidad casi nula de ser seleccionadas.

- El muestreo equilibrado, asegura que los tamaños de las muestras en áreas geográficas o dominios particulares no sean demasiado pequeñas.

9. Bibliografía

ADIN, A.; ARAMENDI, J.; GALBETE, E. AND IZTUETA, A. (2012)

El Método del Cubo: Un Método para seleccionar muestras equilibradas. Congreso Vasco de Sociología y Ciencia Política

ARDILLY, P. (1994)

Les Techniques de Sondage. Technip, Paris.

ARDILLY, P. AND TILLÉ, Y. (2006)

Sampling Methods: Exercises and Solutions. Springer, New York.

AZORÍN, F. AND SANCHEZ-CRESPO, J. L. (1986)

Métodos y Aplicaciones del Muestreo. Alianza Editorial, Madrid.

CHAUVET, G. AND TILLÉ, Y. (2005)

Fast SAS Macros for balancing Samples: user's guide. Software Manual, University of Neuchâtel.

CHAUVET, G. AND TILLÉ, Y. (2007)

Application of fast SAS macros for balanced samples to the selection of addresses. *Case Studies in Business, Industry and Government Statistics,* 1:173-182.

COCHRAN, W. (1977)

Sampling Techniques. Wiley, New York.

DEVILLE, J.-C. AND TILLÉ, Y. (2004)

Efficient balanced sampling: the cube method. *Biometrika*, 91:893-912.

DEVILLE, J.-C. AND TILLÉ, Y. (2005)

Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference,* 128:569-591.

KISH, L. (1965)

Survey Sampling. Wiley, New York.

NEDYALKOVA, D. AND TILLÉ, Y. (2009)

Optimal sampling and estimation strategies under linear model. *Biometrika*, 95:521-537.

SÄRNDAL, C.-E.; SWENSSON, B. AND WRETMAN, J. (1992)

Model Assisted Survey Sampling. Springer Verlag, New York.

TILLÉ, Y. (2000)

Ten years of balanced sampling with the cube method: an appraisal. *Demographic Statistical Methods Division Seminar of the U.S. Census Bureau.*

TILLÉ, Y. (2005)

Teoría de Muestreo. Gruope de Statistique, Université de Neuchâtel, Suisse.
http://www2.unine.ch/files/content/sites/statistics/files/shared/documents/curso_teoria_de_muestreo.pdf

TILLÉ, Y. AND MATEI, A. (2007)

The R Package Sampling. The Comprehensive R Archive Network, Manual of the Contributed Packages.
http://cran.r-project.org/web/packages/sampling/sampling.pdf

Tillé, Y. (2010)

Muestreo Equilibrado y Eficiente: el Método del Cubo. Instituto Vasco de Estadística, Vitoria-Gasteiz.
http://www.eustat.es/productosServicios/datos/Seminario_52.pdf