

# Diseño de la Muestra

---

## A Introducción

Siguiendo las orientaciones dadas por la Oficina Estadística de la Unión Europea (EUROSTAT) se ha seleccionado una muestra probabilística representativa de la población de los hogares privados, obtenida como una submuestra de la relación de viviendas utilizadas en una de las últimas encuestas de hogares realizada por el INE. Ello nos ha permitido disponer de un directorio actualizado así como de algunas características de dichas viviendas, que han sido utilizadas en el proceso de selección de la muestra.

Este marco estaba constituido por 8.000 secciones censales, de las cuales se han seleccionado probabilísticamente 999.

A su vez el marco muestral originalmente utilizado fue la relación de secciones censales del Censo de Población y Viviendas 1991.

Por razones de tipo práctico se ha eliminado de la muestra un pequeño grupo de provincias, quedando éstas representadas por otras provincias de características similares y que pertenecen a su misma Comunidad Autónoma.

Dadas las especiales dificultades que presenta una investigación de tipo panel y con objeto de alcanzar el tamaño mínimo de muestra exigido por EUROSTAT, se han sustituido todas las incidencias habidas durante el trabajo de campo, negativas, ausencias y viviendas no encuestables.

Para ello además de la muestra de viviendas titulares se seleccionó una muestra de viviendas reservas que se ha utilizado en la sustitución de aquellas viviendas que han presentado algún tipo de incidencia.

La sustitución ha sido controlada, en el sentido de que se ha sustituido cada vivienda por otra con las mismas características. La variable de control utilizada fue el tamaño del hogar

Las características más importantes del diseño de la muestra se detallan a continuación.

---

## B Tipo de muestreo y estratificación

Se ha utilizado un muestreo bietápico estratificado entre las unidades de primera etapa.

Dado que entre otros objetivos del panel figura el de poder facilitar ciertas clasificaciones a nivel de NUT2 (comunidades autónomas), se ha seleccionado una muestra independiente en cada una de ellas.

Las unidades de primera etapa han sido las secciones censales, que se consideran conglomerados de viviendas.

Aunque para esta encuesta no se ha calculado, el efecto de diseño obtenido en otras encuestas que emplean diseños semejantes, para características relacionadas con la actividad que se presentan en un 20 por ciento de la población, está en torno a 2.

Las unidades de segunda etapa son las viviendas familiares principales. En ellas no se realiza submuestreo y se investiga a todos los hogares y personas que son

miembros del hogar de acuerdo con las definiciones dadas para la encuesta.

La variable de estratificación utilizada ha sido el tamaño del municipio al que pertenece la sección.

Para llegar a la formación de los estratos se consideran los siguientes tipos de municipios:

**1. Municipios autorrepresentados:** Son aquellos, que dada su categoría dentro de la provincia, deben tener siempre secciones en la muestra.

Son municipios autorrepresentados:

- La capital de la provincia
- Municipios de importancia similar o superior a la capital de la provincia
- Municipios que teniendo una situación demográfica destacada dentro de la provincia no existen otros similares con que agruparlos.

**2.- Municipios correpresentados:** Son aquellos que dentro de la misma provincia forman parte de un grupo de municipios demográficamente similares y que son representados en común.

De acuerdo con esta clasificación, en líneas generales, los estratos teóricos considerados responden a los siguientes conceptos:

Estrato 1: Municipio capital de provincia

Estrato 2: Municipios autorrepresentados, de importancia similar o superior a la capital de provincia (caso por ejemplo de Gijón o Vigo)

Estrato 3: Municipios autorrepresentados, importantes en relación con la capital o municipios mayores de 100.000 habitantes

Estrato 4: Municipios entre 50.000 y 100.000 habitantes

Estrato 5: Municipios entre 20.000 y 50.000 habitantes

Estrato 6: Municipios entre 10.000 y 20.000 habitantes

Estrato 7: Municipios entre 5.000 y 10.000 habitantes

Estrato 8: Municipios entre 2.000 y 5.000 habitantes

Estrato 9: Municipios menores de 2.000 habitantes

Hay que tener en cuenta que dada la diferente distribución de tamaños de los municipios entre las distintas provincias no es eficiente realizar una estratificación uniforme para todas ellas. Sin embargo, se ha procurado realizar una estratificación uniforme para todas las provincias pertenecientes a una misma Comunidad Autónoma. Así por ejemplo, en Andalucía y Canarias apenas existen municipios menores de 2.000 habitantes, por ello en estas Comunidades el estrato menor corresponde a **Municipios menores de 5.000 habitantes**, englobando en un sólo estrato los estratos teóricos 8 y 9. Por el contrario, en Castilla y Aragón adquieren gran importancia los municipios inferiores a 2.000 habitantes y apenas existen municipios intermedios entre 5.000 y 10.000 habitantes. En estas Comunidades se han agrupado los estratos teóricos 7 y 8, creando un estrato de municipios entre 2.000 y 10.000 habitantes y manteniendo plenamente el estrato 9.

---

## C Tamaño de la muestra y distribución

El tamaño de la muestra, a nivel global, para todos los países de la Unión Europea es de 76.500 hogares, lo que supone un total de aproximadamente 155.000 personas entrevistadas.

La distribución de la muestra entre los diferentes países se ha realizado teniendo en cuenta además de la necesidad de información el tamaño y las características del país.

A España le ha sido asignada una muestra de 8.000 hogares.

Se ha fijado un total de 8 viviendas por sección, para una muestra final de 999 secciones distribuidas por todo el territorio nacional.

Para lograr los objetivos de la encuesta, se exige una muestra mínima de 400 hogares para cada NUT1 (Grandes regiones) y de 200 hogares para cada NUT2 (Comunidad Autónoma), por lo que la afijación de las 999 secciones entre Comunidades Autónomas se ha realizado asignando una parte de las secciones uniformemente y el resto de modo proporcional.

La distribución de la muestra de secciones aparece en el cuadro I.

### **Cuadro I. Distribución de la muestra**

<b>NUT1</b>	<b>NUT2</b>	<b>SECCIONES</b>
1. Noroeste	Galicia	63
	Asturias	40
	Cantabria	32
2. Nordeste	País Vasco	54
	Navarra	32
	Rioja	29
	Aragón	42
3. Madrid	Madrid	108
4. Centro	Castilla-León	61
	Castilla-La Mancha	48
	Extremadura	40
5. Este	Cataluña	113
	Comunidad Valenciana	80
	Baleares	35
6. Sur	Andalucía	124
	Murcia	40
	Ceuta-Melilla	-
7. Canarias	Canarias	58

La afijación en cada comunidad se ha hecho primero entre las provincias y luego entre los estratos, siguiendo el criterio de afijación proporcional a la población, aunque se han potenciado los estratos que contienen a los municipios de mayor tamaño por ser éstos donde la población tiene características con mayor variabilidad y también donde se presenta un mayor número de incidencias.

#### **D Selección de la muestra**

La selección de las unidades primarias, dentro de la muestra general de la cual se ha obtenido, se ha realizado en cada estrato con probabilidad proporcional al tamaño.

La selección de la muestra de viviendas en cada sección, se ha realizado mediante

un muestreo sistemático con arranque aleatorio.

---

## E Estimadores

Los estimadores utilizados en esta encuesta son estimadores de expansión corregidos con información facilitada por fuentes externas, para ajustar las distribuciones obtenidas mediante la muestra a las distribuciones poblacionales.

Para el cálculo de las estimaciones se sigue un proceso que permite obtener el peso o ponderación final de cada hogar en cuatro pasos, que se aplican cada uno después del anterior y que permite conocer la aportación de cada uno de ellos al peso final. (Los dos últimos pasos han sido realizados por EUROSTAT)

Estos factores se calculan normalizados de tal forma que permitan la comparación entre sí de todos los hogares de la muestra.

Cada persona dentro del hogar tiene el mismo factor de elevación.

Los pasos a seguir son los siguientes:

### Paso 1. Factor de diseño

Este factor es la inversa de la probabilidad de selección del hogar.

De acuerdo con el procedimiento de selección de la muestra, la probabilidad de una vivienda  $i$  en un estrato  $h$  se puede calcular, aproximadamente, mediante la expresión:

$$P(V_{ih}) = \frac{g_{nh}}{V_h}$$

siendo:

$n_h$  = número de secciones afijadas en la encuesta en el estrato  $h$

$V_h$  = Total de viviendas según Censo en el estrato  $h$

El factor de diseño normalizado tiene la expresión:

$$W_{ih}^{(1)} = \frac{\frac{V_h}{g_{nh}} \cdot m}{\sum_h \frac{V_h \cdot m_h}{g_{nh}}}$$

siendo:

$m_h$  = muestra efectiva de hogares en el estrato  $h$

$$m = \sum_h m_h$$

### Paso 2. Factor de ajuste de la falta de respuesta

Este factor trata de recoger el efecto de las distintas tasas de respuesta que se pueden alcanzar en diferentes partes de la muestra.

La expresión normalizada de dicho factor es:

$$W_{ih}^{(2)} = \frac{m'_h \cdot m}{m_h \cdot m'}$$

siendo:

$m'_h$  muestra teórica de viviendas en el estrato h

$m_h$  = muestra efectiva de viviendas completamente entrevistadas en el estrato

$m = \sum m_h$  = total muestra efectiva

$m' = \sum_h m'_h$  = total muestra teórica 4

### Paso 3. Factor para corregir la distribución de los hogares

Se introduce para ajustar la distribución de hogares estimada a partir de la muestra, después de aplicar las ponderaciones obtenidas en los pasos 1 y 2, a la distribución poblacional de hogares conocida a través de fuentes externas.

Esta información deberá ser fiable y actualizada y puede proceder de un Censo, una encuesta o cualquier otra fuente.

El ajuste se puede realizar para distintas características.

Las características de control utilizadas han sido:

- Distribución de hogares según número de activos
- Distribución de hogares según número de personas que los habitan

Ambas se han obtenido a partir de las estimaciones de la Encuesta de Población Activa, en el cuarto trimestre de 1994.

El planteamiento general del problema es el siguiente:

Sea:

$\hat{P}_k$  = Proporción estimada de hogares que poseen la modalidad k de una determinada característica de clasificación ( $k = 1 \dots K$ ), siendo estas modalidades mutuamente excluyentes.

$$\hat{P}_k = \frac{\sum_i d_{io} y_{ik}}{\hat{Y}}$$

$\hat{Y}$  = Total de hogares en la población, estimado a partir de la muestra.

$$\hat{Y} = \sum_i \sum_{k=1}^K d_{io} y_{ik}$$

$d_{io}$  = Peso asignado al hogar i después de aplicar los pasos 1 y 2.

$y_{ik}$  = Variable que toma los valores 1 ó 0 según que el hogar  $i$  posea o no la modalidad  $k$  de la característica.

$\sum_i$  se extiende a todos los hogares de la muestra.

$P_k$  = Proporción poblacional en la modalidad  $k$  de la característica utilizada para el ajuste.

El objetivo es encontrar una nueva ponderación  $d_i$  ( $d_i=f_i \cdot d_{i0}$ ) de tal forma que se verifique:

$$\hat{P}_k = \frac{\sum_i d_i y_{ik}}{\sum_i \sum_{k=1}^K d_i y_{ik}} = P_k$$

con la condición de que la distancia entre  $d_i$  y  $d_{i0}$  sea mínima.

Existen diversas soluciones al problema antes planteado que dependen de la función de distancia elegida.

En nuestro caso se ha aplicado la metodología desarrollada por el Instituto de Estadística de Francia, adoptada por EUROSTAT para esta encuesta.

El procedimiento nos permite obtener el factor  $f_i$ , que una vez normalizado y siguiendo la notación anterior denominamos  $W_i^{(3)}$ .

#### Paso 4. **Factor para corregir la distribución de personas**

Este factor se introduce, para ajustar la distribución de personas en la muestra a la distribución de personas en la población para determinadas características demográficas significativas.

Las características utilizadas para el ajuste son la edad y el sexo, siendo la distribución poblacional empleada la Proyección Demográfica de Población referida a mitad del cuarto trimestre de 1994.

El planteamiento es análogo al descrito en el paso 3.

Sea:

$\hat{P}_c$  = Proporción de personas con la categoría  $c$  estimada a partir de la muestra ( $c=1 \dots C$ )

$$\hat{P}_c = \sum_i \frac{d_{il} x_{ic}}{\hat{X}}$$

siendo:

$x_{ic}$  = Número de personas que pertenecen a la categoría  $c$  de la variable de ajuste (grupos de edad y sexo) en el hogar  $i$ .

$\hat{X}$  = Total de personas estimado en la muestra

$$\hat{X} = \sum_i \sum_{c=1}^C d_{i1} x_{ic}$$

$d_{i1}$  = Peso asignado al hogar después de aplicar los tres pasos anteriores

$P_c$  = Proporción de la categoría  $c$  obtenida de las Proyecciones Demográficas de Población.

El objetivo, análogamente al paso 3, es ajustar el peso del hogar  $d_{i1}$  por un factor  $g_i$  tal que al aplicar este peso corregido  $d_{i1} \cdot g_i$  la distribución estimada de la muestra coincida con la distribución dada por la Proyección de Población obtenida de fuentes externas.

El ajuste se hace de tal forma que el factor  $g_i$  se aproxime a 1 tanto como sea posible, aplicando el método de mínimos cuadrados generalizados que asigna el mismo factor a todos los miembros del hogar.

A este factor normalizado le denominamos  $W_i^{(4)}$  siguiendo la nomenclatura utilizada en fases anteriores

### **Ponderación final**

Como factor de ajuste global, a cada hogar se le asigna el obtenido como producto de los factores calculados en cada uno de los pasos anteriores.

$$W_{ih}^{(final)} = W_{ih}^{(1)} \times W_{ih}^{(2)} \times W_i^{(3)} \times W_i^{(4)}$$

Este peso final se normaliza, es decir la media para todas las personas es igual a 1.

### **Factor de expansión**

Las ponderaciones anteriores son válidas para la estimación de medias y proporciones.

Para estimar el total de cualquier característica investigada, el cociente entre la población total y la correspondiente población muestral,  $P/p$ , se multiplica por las ponderaciones muestrales normalizadas. Con lo que el factor para totales quedaría:  $P/p \cdot W_{ih}$ .

---

## **F Estimadores para ciclos posteriores al primero**

El cálculo de las ponderaciones para el análisis transversal y longitudinal en las encuestas tipo panel, a lo largo del tiempo, envuelve una serie de técnicas complejas.

Las ponderaciones iniciales asignadas a las personas muestrales en el ciclo 1, han de ser ajustadas para reflejar los cambios en la población de estudio así como la evolución de la muestra en el tiempo.

Estas ponderaciones ajustadas son las que vamos a denominar **ponderaciones básicas**. En principio, los datos de la muestra original con las ponderaciones básicas pueden ser utilizadas para el análisis transversal y longitudinal. Sin embargo, y puesto que el panel contiene también información sobre personas no muestrales, que han entrado en la encuesta por residir en un hogar con una o más

personas muestrales, se utilizará una aproximación para incluir en el análisis la información facilitada por estas personas no muestrales.

Las ponderaciones necesarias para análisis específicos se obtienen por transformaciones sencillas de los pesos básicos.

---

## F.1 CÁLCULO DE LAS PONDERACIONES BÁSICAS

### a) **Peso inicial**

En el ciclo 1, a cada hogar  $i$  y a cada miembro  $k$  del mismo, se le asigna el mismo peso que en el apartado anterior denominamos  $W_{ih}^{(final)}$ .

Llamamos a este factor **peso inicial** que a partir de ahora notamos por

$$u_i^{(1)} = u_{ik}^{(1)}$$

definido para todos los hogares de la muestra y para todas las personas muestrales.

### b) **Peso básico**

En cada ciclo  $t$ , este peso se define para todas las personas muestrales. Se consideran *personas muestrales* aquellas que formaron parte de la muestra en el ciclo 1, y que aún siguen perteneciendo a la misma, más los niños nacidos de alguna mujer muestral.

El peso básico de una persona en el ciclo  $t$  se obtiene por ajuste de su peso básico asignado en el ciclo  $t-1$ , es decir

$$u_{ik}^{(t)} = u_{ik}^{(t-1)} \cdot f_{ik}^{(t-1 \rightarrow t)}$$

Para calcular el factor de ajuste  $f_{ik}^{(t-1 \rightarrow t)}$  hay que tener en cuenta los siguientes factores que recogen los cambios producidos en la población y la muestra entre los ciclos  $t-1$  y  $t$ :

#### 1º) **Ajuste de la falta de respuesta**

Los pesos del ciclo  $t-1$  se multiplican por un factor inversamente proporcional a la probabilidad de la persona de colaborar en el ciclo  $t$ , habiendo colaborado en el ciclo  $t-1$ .

Esta probabilidad se determina en función de ciertas características de hogares o de personas.

#### 2º) **Ajuste con fuentes externas**

En cada ciclo, los pesos deben ser ajustados para que los datos estimados a partir de la muestra sean consistentes con la distribución de la población según diferentes características demográficas.

3º) **Los niños nacidos entre  $t-1$  y  $t$**  de una mujer muestral recibe el peso básico de su madre.

Los niños nacidos de una madre que no es una persona muestral no se consideran personas muestrales.

Estos pesos básicos se definen a nivel de persona pero no de hogar.

Todas las personas no muestrales tienen un peso básico igual a cero.

---

## F.2 ESTIMADORES PARA EL ANÁLISIS TRANSVERSAL

A cada hogar  $i$  se le asigna un peso obtenido como la media de los pesos básicos de sus miembros adultos, es decir:

$$w_i^{(t)} = \frac{1}{s+n} \sum_{k=1}^s u_{ik}^{(t)}$$

siendo:

$u_{ik}^{(t)}$  = peso básico asignado al adulto muestral  $k$  en el hogar  $i$ , en el ciclo  $t$

$s$  = el número de adultos muestrales

$n$  = el número de adultos no muestrales.

De la misma forma que los hogares entrevistados en el ciclo 1 ponderados por sus pesos iniciales  $u_i^{(1)}$  proporcionan una muestra representativa de hogares en el periodo en que se inició el panel, los hogares entrevistados en cualquier ciclo posterior  $t$ , ponderados por los pesos  $w_i^{(t)}$  proporcionan una muestra representativa en el tiempo  $t$ .

Para el análisis transversal de personas, a todos los miembros muestrales y no muestrales, adultos y niños se le asigna el peso del hogar calculado anteriormente, es decir

$$W_{i,j}^{(t)} = W_i^{(t)}$$

Todas las personas entrevistadas en el ciclo  $t$  ponderadas con su **peso compartido**  $w_{ij}^{(t)}$  son, para el objetivo de representar a la población transversal en el periodo  $t$ , equivalentes a las personas muestrales entrevistadas en  $t$  ponderadas con sus pesos básicos  $u_{ik}^{(t)}$ .

Utilizando lo primero en lugar de lo último nos permite una extensión de la base muestral, por la inclusión en el análisis de las personas no muestrales que residen con personas muestrales.

---

## F.3 ANÁLISIS LONGITUDINAL DE PERSONAS

El análisis longitudinal ha de ser específico para un intervalo de tiempo, y limitado a los individuos que pertenecen a la muestra durante dicho intervalo.

La transformación de los pesos básicos es similar a la indicada en el análisis transversal, es decir considerando el intervalo de tiempo transcurrido entre  $t$  y  $T$

$$w_{ij}^{(t \rightarrow T)} = \frac{1}{s^{(T)} + n^{(t \rightarrow T)}} \sum_{k=1}^{s^{(T)}} u_{ik}^{(T)}$$

siendo:

$u_{ik}^{(T)}$  = Peso básico asignado al adulto muestral  $k$  del hogar  $i$  en el ciclo  $T$

$s^{(T)}$  = Total de adultos muestrales en  $T$

$n^{(t \rightarrow T)}$  Número de adultos no muestrales presentes en  $T$  y que entraron a formar parte de la muestra en  $t$  o antes.

Todas las personas del hogar tienen el mismo peso.