QÜESTIIÓ, vol. 23, 3, p. 543-558, 1999

DISEÑO DE MUESTRAS EN ENCUESTAS DE POBLACIÓN Y HOGARES

J. PORRAS PUGA
Instituto Nacional de Estadística*

En este articulo se presenta el esquema general del diseño muestral que el Instituto Nacional de Estadística (INE) utiliza en las encuestas dirigidas a la población. Este diseño se conoce como Encuesta General de Población y fue implantado en el año 1971, aunque desde entonces ha sufrido ligeros cambios en cuanto a la periodicidad, criterios de estratificación, tamaño, etc. Se utiliza tanto para las encuestas de tipo continuo (encuestas coyunturales) como de tipo esporádico (encuestas estructurales).

Sampling design for Household Surveys

Palabras clave: Marco, estratificación, probabilidad proporcional, estimador de razón

Clasificación AMS (MSC 2000): 62D05

* Instituto Nacional de Estadística (INE). Diseño de Muestras de Población y Hogares. Paseo de la Castellana, 183. 28071 Madrid. E-mail: juporras@ine.es

– Recibido en setiembre de 1999.
– Aceptado en noviembre de 1999.

1. INTRODUCCIÓN

La Encuesta de Población Activa (EPA) es una encuesta de tipo continuo y periodicidad trimestral cuyo objetivo es el conocimiento de las características de la población en relación con la actividad económica.

El INE viene realizando esta encuesta ininterrumpidamente desde 1964, habiendo sufrido desde entonces diversas modificaciones que han afectado tanto al cuestionario como al diseño de la muestra.

Desde 1971 el diseño de la EPA se enmarca en el de la Encuesta General de Población (EGP).

La EGP no es una encuesta en sí misma sino un diseño muestral válido para encuestas dirigidas a la población. El objetivo de la misma es mantener un diseño muestral actualizado que permita, en un momento dado, investigar las características de la población española en todos aquellos aspectos que interesen a la Administración del Estado.

El ámbito poblacional en la EGP es la población que reside en viviendas familiares principales. Estas son las utilizadas toda o la mayor parte del año como residencia habitual o permanente. Se excluye del ámbito poblacional la población residente en hogares colectivos.

El estudio básico que el INE realiza con este diseño, como se mencionó anteriormente, es la EPA. Otras encuestas realizadas en el marco de la EGP son, entre otras, las Encuestas de Presupuestos Familiares (1980-81 y 1990-91), la Encuesta de Fecundidad 1998 y la Encuesta sobre Discapacidades y Deficiencias (1986 y 1999).

En los apartados siguientes se presentan los aspectos más importantes del diseño de la encuesta.

2. MARCO

Se considera marco de una encuesta al conjunto de información que puede ser útil en cualquier momento del diseño de la misma.

En sentido restringido el marco está formado por la relación de unidades de muestreo. Las unidades de muestreo deben estar definidas de forma tal que su identificación sea inequívoca, no exista solapamiento entre ellas, a cada unidad se le pueda asignar una probabilidad de selección y que el conjunto de todas ellas coincida con la población que se pretende estudiar.

En sentido amplio, el marco está constituido además por toda la información complementaria, mapas, listas, comunicaciones, etc., que nos permita llegar a un mayor conocimiento de la población y que se pueda utilizar para la división de la población en estratos, selección de la muestra o formación de estimadores.

De acuerdo con el ámbito poblacional de la encuesta, para la selección de la muestra sería necesario disponer de un listado de todas las viviendas familiares existentes en el territorio nacional. Esto es difícil de conseguir ya que, por una parte, esta relación viene afectada por el paso del tiempo y, por otra, de errores tales como omisiones, duplicidades, etc.

Como consecuencia de lo anterior, se elige en lugar de viviendas áreas geográficas que son más estables en el tiempo y se muestrean dichas áreas.

Para definir el marco de la Encuesta es necesario partir de la división administrativa de España. Todo el Estado se encuentra dividido en 17 **Comunidades Autónomas** más Ceuta y Melilla y a su vez en 50 **provincias** de las cuales 47 son peninsulares y 3 insulares. Las provincias se encuentran divididas en **municipios** y éstos en distritos municipales. Hasta aquí tenemos la división administrativa oficial. Después el INE, juntamente con los Ayuntamientos, hace una nueva subdivisión de los distritos en **secciones censales**.

Las secciones se utilizan para todos los trabajos encomendados al INE en los que es necesario una división inframunicipal, entre otros para fines electorales como *secciones electorales*, lo cual exige de acuerdo con la Ley Electoral que cada sección incluya un máximo de 2.000 electores y un mínimo de 500. Por tanto, la sección censal puede considerarse como un área geográfica con límites perfectamente definidos, cuyo tamaño de población viene limitado por las condiciones antes expuestas.

El seccionado y su número varía considerablemente a lo largo del tiempo, por lo que con referencia 1 de enero de cada año y en cada Censo o Padrón se realiza una actualización del mismo. Por una parte, hay secciones que quedan despobladas y es necesario fusionarlas con otras y, por otra, también se produce el fenómeno contrario, es decir, las secciones crecen hasta superar los límites de población establecidos y es necesario dividirlas. Finalmente, para llegar a la vivienda familiar es posible confeccionar para cada sección censal una lista de viviendas familiares con sus direcciones postales, obtenidas del último Censo o Padrón.

Por tanto, el marco de la encuesta lo constituye:

– El marco de áreas, formado por las aproximadamente 32.000 secciones censales en que se encuentra dividido el territorio nacional.

– El marco de viviendas formado por la relación de viviendas familiares que se confecciona para cada sección censal seleccionada para la muestra.

3. UNIDADES DE MUESTREO Y CRITERIOS DE ESTRATIFICACIÓN

El tipo de muestreo utilizado es un **muestreo bietápico de conglomerados con submuestreo y estratificación de las unidades de primera etapa**. En cada provincia se diseña una muestra independiente.

Las unidades de primera etapa son las **secciones censales**. Con el objetivo de conseguir estimaciones fiables del cambio entre dos períodos de encuesta, la muestra de secciones permanece fija indefinidamente, con las excepciones siguientes:

a) Cuando los resultados obtenidos en los Censos arrojen variaciones sensibles en la estructura de la población que aconsejen una afijación distinta.

b) Se agoten los hogares consultables de la sección.

c) Cuando al actualizar las probabilidades de selección le corresponda salir de la muestra.

Estas unidades se estratifican atendiendo a un doble criterio:

Criterio geográfico (de estratificación)

Las secciones en cada provincia se agrupan según la importancia demográfica del municipio a que pertenecen.

Criterio socioeconómico (de subestratificación)

Dentro de cada estrato geográfico las secciones censales se agrupan en **subestratos**, atendiendo a la categoría socioeconómica de los hogares ubicados en 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 que tienen un número de habitantes tal que en la afijación proporcional dentro de la provincia le corresponden al menos 12 secciones en la muestra.

– Municipios que teniendo una situación demográfica destacada dentro de la provincia no hay otros similares con que agruparlos, aunque proporcionalmente le correspondan menos de 12 secciones en la muestra.

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, importantes en relación con la capital.

Estrato 3: Otros 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 se ha podido realizar una estratificación uniforme para todas ellas. Por ejemplo, en la provincia de Lugo solamente hay 10 municipios con menos de 2.000 habitantes, por lo que se han agrupado los estratos teóricos 8 y 9 en el estrato 8 que contiene a los municipios de menos de 5.000 habitantes. Por el contrario, la provincia de Burgos tiene más de 350 municipios de menos de 2.000 habitantes incluidos en el estrato 9 y, sin embargo, tiene agrupados los estratos teóricos 7 y 8 en el estrato 7 al no haber apenas municipios entre 2.000 y 5.000 habitantes. No obstante, siempre que ha sido posible, se ha procurado realizar una estratificación uniforme para todas las provincias pertenecientes a una misma Comunidad Autónoma.

Para la formación de los **subestratos** se tiene en cuenta la categoría socioeconómica de los hogares ubicados en la sección. Las secciones cambian de subestrato debido a la variación de la estructura de la población, por lo que la subestratificación se revisa en cada Censo, utilizando la información que éste proporciona sobre las características que intervienen en la definición de categoría socioeconómica.

Esta información permite clasificar la población económicamente activa de la sección en 18 categorías que a su vez se agrupan en cuatro grupos homogéneos. El primero agrupa a la población cuya actividad principal es la agricultura; el segundo al conjunto de trabajadores por cuenta propia. El tercer grupo representa al conjunto de directivos

y profesionales por cuenta ajena y al personal administrativo y el cuarto grupo al resto de los trabajadores.

Existen dieciséis subestratos, quince de los cuales se obtienen en función de los porcentajes de población de los grupos anteriores y el decimosexto (subestrato cero) está formado por aquellas secciones con un elevado porcentaje de población inactiva.

La definición de los quince primeros subestratos se establece según:

1) Haya un claro predominio de uno de los cuatro grupos sobre los otros tres; 2) predominen dos sobre los otros dos; 3) predominen tres grupos, y 4) no hay un claro predominio de ninguno de los cuatro grupos. En alguno de los estratos pueden no existir varios de los subestratos.

El criterio matemático para considerar que un grupo es de predominio se establece según que el grupo considerado sea superior a los dos tercios del grupo predominante. Así, por ejemplo, supongamos que los porcentajes de los grupos de población económicamente activa en una sección son: Grupo 1=20, Grupo 2=40, Grupo 3=30 y Grupo 4=10. El grupo más importante es el 2. Se verifica además que: porcentaje grupo 3 $>$ 2/3 porcentaje grupo 2, lo que no sucede con el resto de los grupos. Por tanto el subestrato al que pertenece la sección será el 23, es decir, predominan estos dos sobre el resto.

Las **unidades de segunda etapa** están constituidas por las viviendas familiares principales (ocupadas permanentemente) y los alojamientos fijos (chabolas, cuevas, etc.). No se consideran las viviendas secundarias (ocupadas sólo una parte del año) y las disponibles para alquiler o venta, ya que no forman parte del ámbito poblacional definido anteriormente.

Dentro de las unidades de segunda etapa no se realiza submuestreo alguno, recogiéndose información de todas las personas que tengan su residencia habitual en las mismas.

4. TAMAÑO Y AFIJACIÓN DE LA MUESTRA

Para la determinación del tamaño de muestra se partió de una función de coste de tipo lineal y de la expresión del coeficiente de variación para una proporción en el muestreo de conglomerados con submuestreo.

Se empleó la siguiente función de coste:

$$Q = n\,Q_S + n\,m\,Q_V \quad \text{con} \quad Q_S = Q_F + d\,Q_D$$

donde:

Q = Presupuesto total para el pago a los entrevistadores
Q_S = Coste por unidad primaria (sección)
Q_V = Coste por unidad última (vivienda)
n = Número de secciones
m = Número de viviendas por sección
Q_F = Coste fijo por sección
Q_D = Coste diario del trabajo de campo
d = Número de días necesarios para el trabajo de campo

Todas las variables eran conocidas excepto n y m.

El coeficiente de variación para una proporción viene dado por:

$$C^2(\hat{P}) = \frac{V(\hat{P})}{\hat{P}^2} = \frac{1-\hat{P}}{\hat{P}} \cdot \frac{1+\delta(m-1)}{n\,m} = \frac{1-\hat{P}}{\hat{P}}\, F(\delta,\, m,\, n)$$

siendo:

$$F(\delta,\, m,\, n) = \frac{1+\delta(m-1)}{n\,m}$$

y δ el coeficiente de correlación intraclásica, que para el caso de la población activa se ha calculado y vale 0,05.

El mínimo de la expresión $C^2(\hat{P})$ respecto de las variables m y n se obtiene calculando el mínimo de la expresión $F(\delta,\, m,\, n)$ que es independiente de $\hat{P}$.

Para distintos valores de m compatibles con el trabajo de campo,

$$m = 4, 6, 8, 10, 11, 14, 17, 18, 19, \ldots, 91, 100$$

y los correspondientes valores de n dados por:

$$n = \frac{Q}{Q_S + m\,Q_V}$$

se obtienen distintos valores para $F(\delta,\, m,\, n)$.

El valor mínimo de $F(\delta,\, m,\, n)$ respecto de m y n correspondió a m = 20 y n = 3.000. En base a este resultado la muestra se fijó en un total de 3.060 secciones.

Posteriormente, con objeto de lograr una mayor representatividad en algunas Comunidades Autónomas y al mismo tiempo dar cumplimiento a las exigencias de la Unión Europea en cuanto al tamaño de la muestra en las Encuestas de Empleo, se ha ampliado la muestra en diversas ocasiones hasta alcanzar el tamaño actual de 3.484 secciones.

Para la afijación entre las provincias se tuvieron en cuenta los siguientes aspectos:

a) Disponer en cada provincia de un tamaño mínimo de muestra que permita dar estimaciones de la misma.

b) Los resultados nacionales deben tener la mayor fiabilidad posible.

Para compatibilizar estas condiciones se ha aceptado una **afijación de compromiso entre la uniforme y la proporcional**, a base de agrupar provincias de importancia demográfica similar y asignarles de 36 a 144 secciones (actualmente estos límites son de 39 a 156 secciones).

Dentro de cada provincia la afijación entre estratos es proporcional al tamaño de cada uno de ellos, si bien se han potenciado los estratos donde se encuentran los municipios de mayor tamaño, ya que se espera que la mayor parte de las características que se estudian estén correlacionadas con los niveles económico-social y cultural de los habitantes y es precisamente en estos estratos donde, en general, la dispersión debe ser mayor y donde el costo por entrevista es menor.

Dentro de los estratos, la afijación entre subestratos es estrictamente proporcional al tamaño (medido en número de viviendas familiares).

El tamaño de muestra final de unidades de segunda etapa depende de los objetivos de la encuesta, variando desde 16.000 viviendas en la Encuesta de Fecundidad a 75.000 investigadas en la Encuesta de Discapacidades.

5. SELECCIÓN DE LA MUESTRA

La selección de la muestra se realiza de tal forma que dentro de cada estrato cualquier vivienda familiar tenga la misma probabilidad de ser seleccionada, es decir, se tengan **muestras autoponderadas dentro de cada estrato**.

Para ello, las unidades de primera etapa (secciones censales) se seleccionan con probabilidad proporcional al número de viviendas familiares principales, según los datos del último Censo o Padrón. Dentro de cada sección seleccionada en primera etapa, se selecciona un número fijo de viviendas familiares, m, con igual probabilidad mediante la aplicación de un muestreo sistemático con arranque aleatorio.

Por tanto, la probabilidad de selección de la vivienda i, perteneciente a la sección j del estrato h, donde se han afijado K_h secciones sería

Siendo

$$P(v_{ijh}) = P(S_{jh}) \cdot P(v_{ijh}/S_{jh}) = K_h \cdot \frac{V_{jh}}{V_h} \cdot \frac{m}{V_{jh}} = K_h \cdot \frac{m}{V_h}$$

$P(S_{jh})$ = Probabilidad de selección de la sección j del estrato h
$P(v_{ijh}/S_{jh})$ = Probabilidad de selección de la vivienda i condicionada a la selección de la sección j.
v_{jh} = Total de viviendas de la sección j.
V_h = Total de viviendas del estrato h.
m = Número fijo de viviendas investigado en cada sección.

Como se ve, esta probabilidad no depende de i ni de j, es decir, la probabilidad de selección de una vivienda no depende de la sección a la que pertenece, sino solamente del estrato.

6. ESTIMADORES

Se utilizan **estimadores de razón** separados tomando como variable auxiliar las Proyecciones Demográficas de población elaboradas por el INE.

La expresión del estimador de una determinada característica X es la siguiente:

$$\hat{X} = \sum_h \sum_{i=1}^{n_h} \frac{P_h}{p_h} x_{hi}$$

extendiéndose el sumatorio en h a los estratos de una provincia, una comunidad autónoma o al total nacional, y donde:

P_h = Proyección de la población que reside en viviendas familiares, en el estrato h.
p_h = Número de personas que habitan en las viviendas de la muestra, en el estrato h, en el momento de la entrevista.
n_h = Número de viviendas en las secciones de la muestra en el estrato h.
X_{hi} = Valor de la característica investigada en la vivienda i-ésima, del estrato h.

A la expresión de este estimador se llega de la siguiente manera:

Al ser la muestra autoponderada en cada estrato, un estimador insesgado de la característica X se puede obtener mediante un estimador de expansión simple, que tiene la expresión:

$$\hat{X} = \sum_h \frac{V_h}{v_h} x_h$$

donde:

V_h = Total de viviendas en el estrato h.
v_h = Viviendas en la muestra en el estrato h.
x_h = Valor muestral de la característica investigada en el estrato h.

Este estimador tiene el inconveniente de que el valor de V_h sólo es conocido en el momento del Censo, por lo que sería necesario estimarlo para los períodos intercensales. Para evitar esto se recurre a un estimador de razón, utilizando como variable auxiliar las proyecciones demográficas de población estimadas por el INE para cada período de encuesta.

El estimador separado de razón tiene la expresión:

$$\hat{X}_r = \sum_h \frac{\hat{X}_h}{\hat{P}_h} P_h$$

siendo:

$\hat{P}_h$ = Población estimada en el estrato h a partir de la muestra

$$\hat{P}_h = \frac{V_h}{v_h} p_h$$

p_h = Población en las viviendas de la muestra en el estrato h.

Sustituyendo $\hat{X}_h$ y $\hat{P}_h$ en la expresión de $\hat{X}_r$ se obtiene:

$$\hat{X}_r = \sum_h \frac{\frac{V_h}{v_h} x_h}{\frac{V_h}{v_h} p_h} P_h = \sum_h \frac{P_h}{p_h} x_h$$

que corresponde a la expresión inicial del estimador.

7. ACTUALIZACIONES EN EL MARCO DE LA ENCUESTA

Las continuas variaciones de población, bien en sus características, bien en su distribución espacial, exigen realizar actualizaciones en el marco que necesariamente repercuten en la estructura muestral.

En el marco de la EGP se consideran tres tipos de actualizaciones:

Actualización en el marco de viviendas con carácter restringido y exclusivo para las secciones de la muestra. Esta actualización, tiene por objeto incorporar las viviendas principales, *altas* de la sección, en el listado de viviendas de la misma.

Actualización en el marco de secciones, consecuencia de las modificaciones producidas por diversas incidencias como particiones, fusiones o variaciones de límites en las secciones seleccionadas. En cada uno de estos casos es necesario determinar la

probabilidad de selección de las nuevas secciones, así como el número de entrevistas a realizar en las mismas.

Actualización con carácter general relativa a todas las secciones y viviendas de la población, la cual se realiza cada cinco años coincidiendo con el *Censo de Población* o el *Padrón Municipal de Habitantes*.

7.1. Actualización en el marco de viviendas

Cada trimestre se actualiza el marco de viviendas en una sexta parte de las secciones seleccionadas para la muestra, con objeto de poder incorporar al marco aquellas viviendas, tanto de nueva construcción como las que se han transformado en viviendas familiares, las cuales no existían como tales cuando se realizó el Censo o Padrón. Estas viviendas se incorporan a la muestra con una probabilidad igual a la original de las viviendas de la sección.

Cada seis trimestres se produce, por tanto, una actualización completa del marco de viviendas en las secciones seleccionadas para la muestra.

Como consecuencia de esta actualización y con objeto de mantener la muestra autoponderada, el número de viviendas a seleccionar en cada sección varía según la expresión:

$$m' = m \cdot \frac{V'_s}{V_s}$$

De esta manera la muestra es autoponderada.

$$P(v_{ijh}) = P(S_{jh}) \cdot P(v_{ijh}/S_{jh}) = K_h \cdot \frac{V_s}{V_h} \cdot \frac{m\,\dfrac{V'_s}{V_s}}{V'_s} = K_h \cdot \frac{m}{V_h}$$

7.2. Actualización en el marco de secciones

Se consideran los siguientes casos:

1º) Partición de secciones

Es el caso de una sección S en la que el crecimiento del número de viviendas principales exige que se escinda en diversas partes S_1, $S_2, \ldots, S_K$, bien para formar nuevas secciones o para incorporarse a otras ya existentes.

Se plantea el problema de determinar las probabilidades de selección de las nuevas secciones para conocer cual es la que va a permanecer en la muestra, así como el número de viviendas a entrevistar en la misma para que la muestra sea autoponderada.

Se distinguen dos casos:

a) La sección S se fragmenta para formar dos o más secciones completas. En este caso se opera como sigue

Llamamos:

V_s = Número de viviendas de la sección S según el último Censo.
V'_s = Número de viviendas de la sección S después de actualizada.
V_{sj} = Número de viviendas de la parte j de la sección S según datos del último Censo.
V'_{sj} = Número de viviendas de la parte j de la sección S después de actualizada.

Se selecciona una de las nuevas secciones S_j con probabilidad proporcional a su tamaño actualizado V'_{sj}/V_{sj}.

El número de viviendas que deben ser objeto de entrevista es:

$$m' = m \cdot \frac{V'_s}{V_s}$$

las cuales son seleccionadas sistemáticamente.

De esta manera la muestra es autoponderada.

b) La sección S se fragmenta para anexionarse a una o más secciones existentes.

En este caso:

Se selecciona uno de los fragmentos con probabilidad proporcional a su tamaño según el último Censo, V_{sj}/V_s, y la nueva sección S'_j a donde se haya incorporado dicha parte quedará automáticamente seleccionada.

El número de viviendas que han de ser entrevistadas viene dado por:

$$m' = m \cdot \frac{V'_{sj}}{V_{sj}}$$

siendo

$V'_{s'j}$ = Número de viviendas principales en la actualidad en la nueva sección S'_j.
$V_{s'j}$ = Número de viviendas principales que existían en el último Censo o Padrón dentro de los límites de la nueva sección S'_j.

2º) Fusión de secciones

Debido a que algunas secciones por los movimientos migratorios y naturales de la población van quedando vacías se procede a su fusión con otra u otras, de forma que en caso de ser seleccionada tengan unidades que investigar.

Si la sección S_j seleccionada para la muestra se fusiona con otra para formar la nueva sección S, ésta queda incorporada automáticamente a la muestra y el número de viviendas a entrevistar es:

$$m' = m \cdot \frac{V'_S}{V_S}$$

siendo

V'_S = Número de viviendas principales en la actualidad en la nueva sección S.
V_S = Número de viviendas principales, según último Censo o Padrón, dentro de los límites de la nueva sección S.

3º) Variación de límites

Éste es el caso de una sección que se forma con fragmentos de dos o más secciones por reajuste en sus límites.

Para el cálculo de la probabilidad de selección, este caso puede considerarse como un proceso en dos etapas: la primera de partición de cada sección y la segunda de fusión adecuada de las secciones resultantes de la partición.

7.3. Actualización de carácter general

Al obtenerse los resultados de un nuevo Censo o Padrón se procede a actualizar las probabilidades de selección de las secciones y a ajustar el número de entrevistas por sección.

Este procedimiento se realiza de tal forma que las probabilidades de selección de las secciones sean proporcionales al número de viviendas que en ese momento tenga cada una. En principio esto podría lograrse partiendo de cero y seleccionando una muestra nueva, pero ello provocaría una ruptura total con la muestra antigua, lo cual es arriesgado en el caso de encuestas continuas como es la EPA. Por ello se arbitra un procedimiento que sin distorsionar las probabilidades de selección que realmente corresponden a cada sección mantenga la muestra con las mínimas variaciones.

El procedimiento que se sigue es el siguiente:

Sea S una sección perteneciente al estrato h, seleccionada en un Censo o Padrón, C, con probabilidad

$$P_S = \frac{V_S^C}{V_h^C} = \frac{\text{Viviendas en } S \text{ según Censo } C}{\text{Viviendas en el estrato } h \text{ según Censo } C}$$

y supongamos que en el siguiente Censo o Padrón, C', le corresponde una probabilidad de selección dada por

$$P_{S'} = \frac{V_S^{C'}}{V_h^{C'}} = \frac{\text{Viviendas en } S \text{ según Censo } C'}{\text{Viviendas en el estrato } h \text{ según Censo } C'}$$

Se compara P_S con P'_S pudiendo ocurrir uno de los dos siguientes casos:

1) Si $P'_S > P_S$ la sección S permanece en la muestra con probabilidad P'_S, ya que si fue seleccionada con una probabilidad P_S inferior a la que actualmente le corresponde, con mayor motivo hubiera salido seleccionada aplicándole su probabilidad actual P'_S.

2) Si $P'_S < P_S$ la sección permanece en la muestra con probabilidad P'_S/P_S y sale de la muestra con probabilidad $1 - P'_S/P_S$.

Este criterio motivará la salida de la muestra de un cierto número de secciones. Estas serán sustituidas por otras secciones del mismo estrato pero seleccionadas de **entre las que no perteneciendo a la muestra hayan aumentado de probabilidad**.

Con este criterio se mantiene el esquema de que la probabilidad que tiene una sección de pertenecer a la muestra es la que realmente le corresponde, es decir, proporcional al número de viviendas actuales.

8. COMENTARIO FINAL

Con este tipo de diseño muestral y las correspondientes actualizaciones, el INE mantiene un marco actualizado sobre el que realiza todas las investigaciones de tipo social y económico dirigidas a la población.

El desarrollo de nuevas técnicas de estimación, así como la disponibilidad del futuro Padrón Continuo permitirá introducir mejoras en este diseño.

ENGLISH SUMMARY

SAMPLING DESIGN FOR HOUSEHOLD SURVEYS

J. PORRAS PUGA
Instituto Nacional de Estadística*

This paper present the general sampling design used by the Instituto Nacional de Estadística (INE) for household surveys. This design is named Encuesta General de Población *and it was implemented in 1971. Since then has had several methodological changes, related to periodicity, criteria of strata, etc. It is used in continuous and non continuous surveys.*

Keywords: Frame, stratification, proportional probability, ratio estimator

AMS Classification (MSC 2000): 62D05

* Instituto Nacional de Estadística (INE). Diseño de Muestras de Población y Hogares. Paseo de la Castellana, 183. 28071 Madrid. E-mail: juporras@ine.es

– Received September 1999.
– Accepted November 1999.

1. INTRODUCTION

The Spanish Labour Force Survey is a continuous survey that has been conducted by the INE since 1964. Since 1971 this survey uses the general design of the **Encuesta General de Población** (EGP). The EGP is an updated sampling design used by the INE in Household Surveys, that are useful for the Government Policy. It includes the population living in private dwellings. The institutions (hotels, hospital,...) are excluded from the sample.

2. DESIGN OF THE E G P

A two-stage sampling is used with stratification of the first stage units. The first stage units are the enumeration areas. These are stratified, within each province, using the population size of the municipality. Within each strata, they are substratified according to the socio-economic characteristic of the population in the enumeration areas.

The second stage units are the private dwellings.

The sample size was calculated using the criterion of minimum variance for the estimator of a proportion with fixed budget and a lineal cost function.

The number of primary units was established in 3.060.

Nowadays the sample size is 3.484 enumeration areas.

The units are selected in such a way to obtain self -weighted samples within each stratum. The first stage units are selected with proportional probability to the size and second stage units are selected with equal probability.

The design use **Ratio Estimator**, and the auxiliary variable is the Population Projection.

The frame is updated in two ways:

- Updating of the dwellings in the sampling primary units. Each of these units is updated every six quarter.
- General updating of all the primary units with the information obtained from the Census or from the Population Register.