INSTITUTO NACIONAL DE ESTADÍSTICA

Encuesta de Población Activa

Diseño de la Encuesta y Evaluación de la calidad de los datos

Informe Técnico

Madrid, 2024
Unidad de Diseño de Muestras

Índice

I Introducción

La Encuesta de Población Activa (EPA), es una encuesta de tipo continuo dirigida a investigar características socioeconómicas de la población, que viene siendo realizada por el INE desde 1964.

Desde su implantación ha sufrido diferentes modificaciones, siempre dirigidas a mejorar la información que proporciona y a adaptarse a la reglamentación europea.

El presente informe tiene por objeto recoger los aspectos metodológicos del diseño actual, así como la evaluación de la calidad de los datos de la misma.

El INE agradece de antemano cuantas sugerencias se presenten para posibles mejoras futuras de la encuesta.

II Diseño de la Encuesta

1 Objetivos

La EPA tiene como objetivo principal el conocimiento de la actividad económica del país, en lo relativo al componente humano. Su diseño está orientado a proporcionar información de las principales categorías poblacionales en relación con el mercado de trabajo así como obtener clasificaciones de estas categorías según distintas variables.

Las diferentes fuentes estadísticas (Censo, Encuestas de Salarios, Paro registrado, etc.) que proporcionan información sobre estos temas, no son adecuadas para satisfacer los objetivos de la encuesta por diferentes motivos.

En el caso del Censo: 1) Su larga periodicidad impide conocer la situación en períodos intercensales.

2) Los datos del Censo son insuficientes para dar una visión detallada de la situación laboral.

Las encuestas Industrial y de Salarios sólo aportan información sobre la población asalariada.

El Paro Registrado publicado por el Servicio de Empleo Público Estatal (SEPE) y la Afiliación a la Seguridad Social presentan dificultades en la obtención de series homogéneas por ser variable la normativa legal que los rige, y además no recogen información sobre muchas de las variables investigadas en la encuesta.

Asimismo, el Reglamento (UE) 2019/1700 del Parlamento Europeo y del Consejo de 10 de octubre de 2019, por el que se establece un marco común europeo para las encuestas dirigidas a las personas y hogares, determina la necesidad de realizar una encuesta para la investigación de la población activa.

Se justifica así la necesidad de una encuesta continua diseñada y concebida expresamente para conocer el grado de actividad económica de la población, junto a otras características estrechamente relacionadas con dicha actividad.

La encuesta está diseñada para dar resultados detallados a nivel nacional. Para las comunidades autónomas y las provincias se ofrece información sobre las principales características al nivel de desagregación que permiten los coeficientes de variación de los estimadores.

Como definición de población económicamente activa se ha tomado la aceptada por la Oficina Internacional de Trabajo (OIT), según la cual se establece ésta como el *conjunto de personas, que en un período de referencia dado, suministran mano de obra para la producción de bienes y servicios económicos o que están disponibles y hacen gestiones para incorporarse a dicha producción*.

De acuerdo con lo anterior la encuesta considera como población económicamente activa la constituida por las personas de 16 y más años que en la semana de referencia satisfacen las condiciones necesarias para su inclusión entre las personas ocupadas o paradas de acuerdo con las definiciones dadas para la encuesta.

Con el objeto de dar cumplimiento a las exigencias de la Unión Europea que establecen un marco común para las encuestas a hogares ya mencionado así como al Reglamento de Ejecución (UE) 2019/2240 de 16 de diciembre de 2019, que fija otras características

de la encuesta europea en el ámbito de la población activa, ha sido necesario introducir una serie de adaptaciones con efecto a partir del primer trimestre de 2021.

2 Ámbito de la Encuesta

El ámbito abarcado por la encuesta se desglosa en los tres apartados siguientes:

2.1 ÁMBITO POBLACIONAL

La Encuesta va dirigida a la población que reside en viviendas familiares principales, es decir, las utilizadas todo el año o la mayor parte de él como residencia habitual o permanente.

Se excluyen de la investigación los llamados *hogares colectivos*, ejemplo de los cuales son los hospitales, hoteles, cuarteles, conventos, etc.

Sí se incluyen las familias que formando un grupo independiente residan en estos establecimientos, como puede ocurrir con los directores de los centros, conserjes y porteros. En definitiva, teóricamente, sólo queda excluida de la muestra aquella población que carezca de residencia familiar, que constituye solamente el 0,9 por ciento de la población total según datos del Censo 2011.

2.2 ÁMBITO GEOGRÁFICO

La encuesta se realiza en todo el territorio nacional.

2.3 ÁMBITO TEMPORAL

La EPA es una encuesta continua con periodicidad trimestral, extendiéndose las entrevistas a lo largo de las trece semanas del trimestre.

Hay que distinguir:

Período de referencia de los resultados de la encuesta: el trimestre.

Período de referencia de la información recogida: se ha adoptado, como norma general, la semana anterior (de lunes a domingo) a la de la fecha en que se realiza la entrevista. Dicha semana se denomina *semana de referencia* y todos los datos deben referirse a ella, salvo las excepciones que figuran en el documento *Encuesta de Población Activa. Descripción de la encuesta, definiciones e instrucciones para la cumplimentación del cuestionario*.

3 Marco de la Encuesta

Para definir el marco de la Encuesta es necesario partir de la división administrativa de España, que aparece de la forma siguiente:

Toda la Nación se encuentra dividida en 17 comunidades autónomas y dos ciudades autónomas, que constituyen los NUTS 2(Nomenclature of Territorial Units for Statistics) aprobados por el Parlamento europeo. Las comunidades autónomas se dividen a su vez en 50 provincias (NUTS 3) de las cuales 47 son peninsulares y 3 insulares. Las provincias se encuentran divididas en municipios y éstos en distritos municipales.

A partir de lo anterior el INE juntamente con los Ayuntamientos hace una nueva subdivisión de los distritos municipales en secciones censales.

Las secciones se utilizan para todos los trabajos encomendados al INE en los que es necesaria 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 varían a lo largo del tiempo, por lo que es necesario realizar una actualización continua 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.

En el apartado 5 de este documento se analiza en detalle cómo estas actualizaciones afectan a la encuesta y su tratamiento.

El marco utilizado para la selección de la muestra tanto de unidades de primera etapa como de segunda, es el Marco de Direcciones Georreferenciadas (MDG), que es un sistema de Información estructurado y jerarquizado compuesto por todas las entidades que forman parte del modelo territorial: comunidad autónoma, provincia, municipio, distrito, sección, calle y resto de elementos que componen la dirección postal. En él todas las direcciones están estructuradas, tienen identificadores únicos y tienen un alto porcentaje de georreferenciación. Además estas direcciones tienen asociados los habitantes padronales y sus entidades territoriales están ajustadas al callejero de Censo Electoral.

Asimismo, el MDG se encuentra sincronizado con la Infraestructura de Datos Espaciales del INE (IDEINE), lo que le aporta una capa de visualización en dicho entorno.

Se actualiza anualmente, a partir de la información sobre altas, bajas y modificaciones tanto de la población como del territorio, que transmiten mensualmente los ayuntamientos para actualizar el Padrón. Y el mantenimiento de los límites de las secciones censales, consistente en la partición, desaparición o creación de nuevas secciones, se hace con la información que reportan las Delegaciones Provinciales del INE.

4 Diseño de la muestra

4.1 TIPO DE MUESTREO. UNIDADES MUESTRALES

Se utiliza un muestreo bietápico con estratificación de las unidades de primera etapa en todo el territorio nacional a excepción de las ciudades autónomas de Ceuta y Melilla, donde el muestreo es en una etapa.

Las unidades de primera etapa (UPM) están constituidas por las **secciones censales**. La muestra de secciones permanece fija en el tiempo, excepto cuando se han agotado, por haberse visitado todas las viviendas encuestables, o cuando en el proceso de actualización del seccionado algunas secciones les corresponden salir de la muestra debido al reajuste de las probabilidades de selección asociadas a las variaciones de población (ver apartado 5).

En todos los casos las secciones que salen de la muestra son sustituidas por otras secciones seleccionadas aleatoriamente.

Las unidades de segunda etapa están constituidas por las viviendas familiares principales (ocupadas permanentemente) y los alojamientos fijos (chabolas, cuevas, etc.). No se consideran encuestables las viviendas secundarias o de temporada (ocupadas sólo una parte del año), ni 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.

En el caso de Ceuta y Melilla se lleva a cabo en cada una de ellas un muestreo aleatorio simple, donde las unidades seleccionadas son las viviendas familiares principales, recogiéndose información de todas las personas que tengan su residencia habitual en las mismas.

4.2 ESTRATIFICACIÓN DE LAS UNIDADES DE MUESTREO

En cada comunidad autónoma las unidades de primera etapa se estratifican siguiendo un criterio geográfico, que asigna el estrato según el tamaño, medido en términos de población, del municipio al que pertenece la sección.

Las unidades de primera etapa se estratifican atendiendo a un doble criterio:

A. Criterio geográfico (de estratificación)

Las secciones se agrupan en estratos dentro de cada provincia, de acuerdo al tamaño del municipio al que pertenecen, medido en términos de población.

B. Criterio socioeconómico (de subestratificación)

Las secciones censales se agrupan en subestratos dentro de cada uno de los estratos, según las características socioeconómicas de las mismas.

4.2.1 Estratos

Los estratos teóricos considerados responden a los siguientes tamaños:

- Estrato 0: Municipios de 500.000 habitantes o más
- Estrato 1: Municipios capital de provincia con menos de 500.000 habitantes.
- Estrato 2: Municipios con más de 100.000 habitantes, excepto los anteriores.
- Estrato 3: Municipios de 50.000 a 100.000 habitantes, excepto los anteriores.
- Estrato 4: Municipios de 20.000 a 50.000 habitantes, excepto los anteriores.
- Estrato 5: Municipios de 10.000 a 20.000 habitantes.
- Estrato 6: Municipios con menos de 10.000 habitantes.

En algunas provincias ha sido necesario unir estratos contiguos, bien porque no hay municipios en alguno de ellos, bien porque la población es demasiado reducida y por tanto no le correspondería muestra en el reparto proporcional de la misma entre estratos.

4.2.2 Subestratos

En el proceso de formación de los subestratos, dentro de cada estrato, se han considerado dos grupos de secciones:

a- **Secciones del estrato 6.** Se considera que este grupo de secciones pertenecientes a municipios pequeños presenta una variabilidad relativamente pequeña respecto de las variables objetivo y en todo caso bien explicada por el territorio al que pertenecen. Por ello se les asigna como subestrato la comarca (LAU1-Local Administrative Units) del municipio al que pertenecen. De esta forma se consigue que, además de distribuir la muestra en grupos homogéneos, la representación muestral del territorio permita obtener en un futuro estimaciones más desagregadas mediante técnicas de estimación en pequeñas áreas.

b- **Resto de secciones**. Las secciones se agrupan dentro de sus estratos mediante la aplicación de técnicas de análisis de conglomerados. En este caso, al tratarse de municipios más grandes y tener por ello prácticamente garantizada la representación muestral de la comarca (LAU1) a la que pertenecen, se ha considerado prioritario utilizar la información auxiliar disponible para formar grupos homogéneos de secciones y mejorar con ello la precisión de las estimaciones.

La información auxiliar utilizada para realizar el análisis en este segundo grupo de secciones procede de los datos agregados a este nivel territorial del Fichero Precensal de 2018 y de la Agencia Estatal de Administración Tributaria (AEAT). Se han elegido aquellas características que se considera que están más correlacionadas con las variables objeto de estudio en la Encuesta de Población Activa.

Las variables auxiliares utilizadas, al nivel de sección, han sido:

- Porcentaje de parados
- Porcentaje de inactivos
- Porcentaje de ocupados

- Porcentaje de extranjeros
- Porcentaje de personas entre 0 y 19 años
- Porcentaje de personas entre 15 y 24 años
- Porcentaje de personas de 65 o más años
- Porcentaje de personas con nivel de estudios realizados 1, 2 ó 3 según la clasificación del censo 2001, esto es, analfabetos, sin estudios o nivel de estudios de primer grado
- Porcentaje de personas con nivel de estudios realizados 4, 5, 6 ó 7, es decir, ESO, EGB, Bachillerato, FP
- Porcentaje de personas con nivel de estudios realizados 8, 9 ó 10, es decir, diplomatura, licenciatura o doctorado

Finalmente, la variable fiscal utilizada ha sido:

- Renta total por vivienda con perceptores

Previamente al análisis de conglomerados se han estandarizado las variables dentro de cada estrato con media 0 y desviación típica 1, con la excepción de las variables porcentaje de parados, porcentaje de jóvenes y la variable fiscal, que se han estandarizado con desviación típica 2. Con ello se pretende que estas últimas variables tengan una ponderación superior al resto y por tanto una mayor influencia en el proceso de formación de los subestratos.

El algoritmo utilizado para obtener los conglomerados (subestratos) ha sido el de Ward (1963). Este es un algoritmo multivariante de análisis de conglomerados jerarquizado, basado en la minimización de las distancias entre conglomerados. En cada etapa, se agrupan dos conglomerados, de forma que la suma de cuadrados de las distancias entre conglomerados se minimiza sobre todas las particiones posibles obtenidas agrupando dos conglomerados de la etapa anterior. De esta forma se pasa de una primera etapa, con tantos conglomerados como secciones en el estrato, hasta la última etapa con todas las secciones en un único conglomerado.

Este procedimiento se ha realizado aplicando el procedimiento CLUSTER, del módulo SAS/STAT de SAS.

Posteriormente se utilizó el procedimiento TREE, también de SAS, que permite visualizar un gráfico con el proceso de formación de los conglomerados. Este gráfico de árbol facilita la decisión sobre el número final de conglomerados a considerar en cada estrato.

La cantidad de subestratos por estrato se asigna en función de la variabilidad interna de los conglomerados, y también considerando el número de secciones de cada uno, de forma que al final no resulten subestratos demasiado pequeños y por ello con difícil representación muestral.

4.3 TAMAÑO DE LA MUESTRA

Con el objetivo de cumplir, con la adecuada holgura, los criterios de precisión impuestos por el nuevo Reglamento Europeo, se han calculado los tamaños muestrales mediante un procedimiento de minimización del coste de la encuesta, sujeto a unas restricciones de precisión en términos de acotación de la varianza del estimador. Para definir el problema de optimización necesitamos las variables de decisión, las restricciones y la función objetivo del mismo.

El problema de optimización que se ha planteado tiene como variables de decisión:

- n_d^* con $d=1, \ldots, 17$ es el número de secciones censales en la muestra de la Comunidad autónoma d, excluyendo Ceuta y Melilla.
- m^* es el número de viviendas seleccionadas de cada sección censal, que se fija como una constante independiente de la Comunidad Autónoma.

La decisión de seleccionar el mismo número de viviendas dentro de cada sección censal se ha tomado para facilitar la organización y gestión de las labores de recogida. Como en Ceuta y Melilla el muestreo es aleatorio simple, no hay que considerar los tamaños muestrales asociados a la primera etapa sino el número de viviendas que se recoge en cada ciudad autónoma.

Los requisitos de precisión impuestos por la reglamentación europea dan lugar a las siguientes restricciones:

- Criterio 1: Requisito de precisión para el paro a nivel nacional:

$$\sqrt{\hat{V}(\hat{p}_u)} \leq L_u$$

Donde $\hat{V}(\hat{p}_u)$ es la estimación de la varianza para la proporción estimada del paro en la población de 16-74 años a nivel nacional y L_u es un 75% de la cota superior impuesta por reglamento.

- Criterio 2: Requisito de precisión para el empleo a nivel nacional:

$$\sqrt{\hat{V}(\hat{p}_e)} \leq L_e$$

Donde $\hat{V}(\hat{p}_e)$ es la estimación de la varianza para la proporción estimada del empleo en la población de 16-74 años a nivel nacional y L_e es un 75% de la cota superior impuesta por reglamento.

- Criterio 3: Requisito de precisión para el paro en cada Comunidad Autónoma:

$$\sqrt{\hat{V}(\hat{p}_{u,d})} \leq L_{u,d} \quad d = 1, \dots, 19$$

donde $\hat{V}(\hat{p}_{u,d})$ es la estimación de la varianza de la proporción estimada del paro en la población de 16-74 años en la comunidad autónoma *d* y $L_{u,d}$ es un 75% de la cota superior en la comunidad d impuesta por reglamento.

La función objetivo a optimizar es la función del coste de la encuesta. Se ha partido de un coste $Q_d(n_d^*, m^*)$ para la comunidad autónoma d que recoge los costes derivados de realizar las entrevistas EPA mediante los dos métodos de recogida que se usan principalmente: CAPI (computer-assisted personal interviewing) y CATI (computer-assisted telephone interviewing). Así pues la expresión para la función de costes sería:

$$Q_d(n_d^*, m^*) = C_1 n_d^*(\alpha + \beta m^*) + C_2 m^* n_d^* f_2$$

Donde:

- C_1 = Coste diario medio por investigar una unidad primaria (sección) por CAPI
- La recta $\alpha + \beta m^*$ = número de días necesarios para investigar m^*viviendas en una sección. Los parámetros α y β son definidos en base a la experiencia adquirida en la recogida en períodos previos de la encuesta
- C_2 = Coste medio por investigar una vivienda por entrevista CATI.
- f_2 = Proporción de viviendas que son investigadas por CATI.

Sumando en todas las comunidades obtenemos la expresión final de la función objetivo:

$$\sum_{d=1}^{19} Q_d(n_d^*, m^*)$$

Las restricciones vienen dadas por los límites de precisión establecidos más arriba.

Este problema se ha resuelto usando el procedimiento PROC OPTMODEL de SAS para programación no lineal.

Los tamaños resultantes de la resolución del problema de optimización han sido revisados para garantizar unos mínimos de muestra por provincias y acotar los cambios con respecto a los antiguos tamaños muestrales.

El número de **secciones en la muestra, es de 5.298**, investigándose **13 viviendas por sección**. De ellas, 288 corresponden al incremento muestral recogido por el Instituto

Galego de Estadística en virtud del convenio firmado con el INE. En el caso de las ciudades autónomas de Ceuta y Melilla, el muestreo será en una única etapa y se seleccionan 260 viviendas principales en ambos casos. En el siguiente cuadro aparece la distribución del número de secciones entre las distintas Comunidades Autónomas.

Comunidad Autónoma	Secciones
01 Andalucía	751
02 Aragón	230
03 Asturias, Principado de	188
04 Balears, Illes	191
05 Canarias	239
06 Cantabria	159
07 Castilla y León	531
08 Castilla-La Mancha	336
09 Cataluña	490
10 Comunitat Valenciana	366
11 Extremadura	210
12 Galicia	576
13 Madrid, Comunidad de	325
14 Murcia, Región de	168
15 Navarra, Com.Foral de	174
16 País Vasco	239
17 Rioja, La	125
18 Ceuta (*)	260
19 Melilla (*)	260

(*) Número de Viviendas investigadas

4.4 AFIJACIÓN

Este apartado recoge los criterios seguidos para la distribución de las secciones de la muestra entre las provincias, dentro de la provincia entre estratos y dentro de éstos entre subestratos. Se ha adoptado una afijación de compromiso entre uniforme y proporcional para la distribución de la muestra de la Comunidad Autónoma entre las provincias que la integran.

Dentro de cada provincia la afijación entre estratos es proporcional al tamaño de cada uno de ellos medido en población. Dentro de los estratos, la afijación entre subestratos es estrictamente proporcional al tamaño (medido en población).

En el cuadro que figura a continuación se encuentra la distribución de la muestra de secciones por provincias y estratos.

	ESTRATO							
	0	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
01 Araba/Álava		37				4	8	49
02 Albacete		27			13		24	64
03 Alicante/Alacant		24	15	29	37	11	18	134
04 Almería		20		18		17	18	73
05 Ávila		16					29	45
06 Badajoz		27		11	17	10	59	124
07 Balears, Illes		66		17	53	29	26	191
08 Barcelona	89		59	43	43	25	31	290
09 Burgos		33			13		22	68
10 Cáceres		19			15		52	86
11 Cádiz		10	28	33	17	8	9	105
12 Castellón/Castelló		20		5	19	5	17	66
13 Ciudad Real		12			22	17	26	77
14 Córdoba		31			16	9	21	77
15 Coruña, A		48	18	14	44	26	60	210
16 Cuenca		12					33	45
17 Girona		9			23	13	26	71
18 Granada		23			17	18	28	86
19 Guadalajara		16				12	22	50
20 Gipuzkoa		20		7	15	21	16	79
21 Huelva		17			17	9	18	61
22 Huesca		12				13	23	48
23 Jaén		12		5	13	13	24	67
24 León		23			23		37	83
25 Lleida		17				8	31	56
26 Rioja, La		59			9	14	43	125
27 Lugo		28	0	0	0	16	48	92
28 Madrid	170		71	41	15	10	18	325
29 Málaga	44		11	38	19		21	133
30 Murcia		51	24	18	47	28		168
31 Navarra		55			20	26	73	174
32 Orense		30	0	0	0	14	46	90
33 Asturias		40	50	23	19	28	28	188
34 Palencia		22					23	45
35 Palmas, Las		42	11	20	32	10	7	122
36 Pontevedra		16	58	0	40	40	30	184
37 Salamanca		29				7	28	64
38 Santa Cruz de Tenerife		24	17	22	30	8	16	117
39 Cantabria		48		14	23	17	57	159
40 Segovia		15					30	45
41 Sevilla	55		10	9	29	23	23	149
42 Soria		19					26	45
43 Tarragona		12	9		21	6	25	73
44 Teruel		12					33	45
45 Toledo		12		19		15	54	100
46 Valencia/València	52			19	46	19	30	166
47 Valladolid		54			10		27	91
48 Bizkaia		35	10	7	26	13	20	111
49 Zamora		16				4	25	45
50 Zaragoza	97					11	29	137
51 Ceuta (*)		(260)						
52 Melilla (*)		(260)						
Total (**)	**507**	**1170**	**391**	**412**	**803**	**577**	**1438**	**5298**

(*) Número de Viviendas investigadas

(**) Sin contabilizar Ceuta y Melilla

4.5 SELECCIÓN DE LA MUESTRA

La selección de la muestra se ha realizado 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**. Este tipo de muestras proporciona pesos de diseño iguales por estrato en los estimadores. Para ello, las unidades de primera etapa (secciones censales) se seleccionan con probabilidad proporcional al número de viviendas familiares principales, según los datos de población. Dentro de cada sección seleccionada en primera etapa, se selecciona un número fijo de viviendas familiares con igual probabilidad mediante la aplicación de **un muestreo sistemático con arranque aleatorio**. Como se dijo anteriormente (ver apartado 4.3), en la encuesta se seleccionan 13 viviendas por sección.

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:

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

siendo:

m=13

$P(S_{jh})$ = Probabilidad de selección de la sección j del estrato h

$P\left(V_{ijh}/S_{jh}\right)$ = Probabilidad de selección de la vivienda i condicionada a la selección de la sección j.

V_{jh} = Total de viviendas de la sección j del estrato h.

V_h = Total de viviendas del estrato h.

Como se observa, esta probabilidad no depende de i ni de j, es decir, ni de la vivienda ni de la sección, y por lo tanto la muestra es autoponderada.

En Ceuta y Melilla se seleccionan 260 viviendas en cada ciudad autónoma usando muestreo aleatorio simple. Por tanto, cada vivienda V_i tiene la misma probabilidad de ser seleccionada:

$$P(\mathrm{V_i}) = \frac{260}{V}$$

Donde V es el total de viviendas en Ceuta o Melilla, según el caso.

4.6 DISTRIBUCIÓN DE LA MUESTRA EN EL TIEMPO

Cada período de la encuesta es de un trimestre, siendo cada una de las secciones de la muestra visitada en una de las 13 semanas del mismo.

La distribución de la muestra es uniforme en el tiempo, lo que equivale a que en cada provincia el número de secciones por semana es constante.

Además se ha procurado que la distribución de secciones muestrales por provincia, estrato y semana sea homogéneo, al igual que por provincia, turno de rotación(ver apartado 4.7) y semana.

4.7 TURNOS DE ROTACIÓN

Como se ha dicho en el párrafo anterior, cada período de la encuesta es de un trimestre, repitiéndose ésta sucesivamente.

Las secciones censales permanecen fijas en la muestra indefinidamente (salvo lo comentado en el apartado 4.1), sin embargo las viviendas familiares son renovadas parcialmente cada trimestre de encuesta, a fin de evitar el cansancio de las familias. Esta renovación se efectúa en una sexta parte de las secciones.

A estos efectos, la muestra total de secciones se halla dividida en seis submuestras que denominamos *Turnos de rotación*. Cada sección viene identificada por un código de cinco dígitos. El último dígito nos expresa el turno de rotación a que pertenece, estando numerado del 1 al 6.

Cada trimestre se renuevan las viviendas que pertenecen a las secciones de un determinado turno de rotación. Por tanto cada vivienda permanece en a muestra durante seis trimestres consecutivos, al cabo de los cuales sale de la misma para ser reemplazada por otra de la misma sección.

Estas viviendas se incorporan a la muestra con una probabilidad igual a la original de las viviendas de la sección.

Por tanto, las viviendas de las secciones de cada turno de rotación colaboran en la encuesta el mismo número de trimestres. Este número de colaboraciones está asociado al turno y varía desde uno hasta un máximo de seis trimestres de permanencia en la encuesta.

La distribución del número de secciones por estrato y semana es similar en cada turno de rotación.

4.8 ESTIMADORES

Hasta el año 2001, se han utilizado **estimadores de razón** tomando como variable auxiliar las cifras de población residente en viviendas familiares principales, deducidas de las Estimaciones de la Población Actual elaboradas por el INE.

Sendo la expresión del estimador de una determinada característica Y, en un trimestre de encuesta la siguiente:

$$\hat{Y} = \sum_{h} \frac{P_h}{p_h} \sum_{i=1}^{n_h} y_{hi} \qquad (1)$$

extendiéndose el sumatorio h a los estratos de una provincia, una comunidad autónoma o al total nacional,según el nivel geográfico de la estimanción.

En esta fórmula:

P_h *:* es la población residente en viviendas familiares principales, en el estrato h, referida a la mitad del trimestre.

p_h : es el número de personas que habitan en las viviendas de la muestra, en el estrato h, en el momento de la entrevista.

n_h : es el número de viviendas en las secciones de la muestra en el estrato h.

y_{hi} : es el valor de la característica investigada en la vivienda i-ésima, del estrato h.

A partir del primer trimestre de 2002, se aplican **Técnicas de reponderación** a los estimadores con objeto de ajustar las estimaciones de la encuesta a la información procedente de fuentes externas.

La técnica de reponderación consiste en lo siguiente:

Se considera una población $U = \{u_1 ..... u_N\}$ de la cual se extrae una muestra

$s = \{u_1 .... u_k ..... u_n\}$

La expresión (1) puede escribirse de la siguiente forma:

$$\hat{Y} = \sum_{k \in s} d_k \, y_k$$

donde:

y_k : Valor de la característica investigada en la unidad muestral k.

d_k : Factor de elevación de la unidad k obtenido mediante la expresión $\frac{P_h}{p_h}$, siendo h el estrato al que pertenece la unidad.

$\sum_{k \in s}$: Sumatorio extendido a todas las unidades de la muestra s.

Se dispone de J variables auxiliares cuyos valores son conocidos para la muestra y cuyos totales son conocidos para la población

$$X_j = \sum_{k \in U} x_{jk}$$

Se trata de encontrar un nuevo estimador

$$\hat{Y}_w = \sum_{k \in s} w_k \, y_k$$

donde los nuevos pesos w_k cumplan las siguientes condiciones:

$\forall \; j = 1.....J$

- Sean próximos a los pesos iniciales d_k
- Verifiquen la ecuación de equilibrado

  $\sum_{k \in s} w_k \, x_{jk} = X_j$

El planteamiento del problema es encontrar unos valores w_k que hagan mínima la expresión:

$$\sum_{k \in s} d_k \; G\left(\frac{w_k}{d_k}\right) \qquad \text{con la condición} \qquad \sum_{k \in s} w_k \; X_k = X$$

siendo:

G = Función de distancia.

X = Vector de dimensión (J,1) con los totales de las variables auxiliares.

X_k = Vector de dimensión (J,1) con los valores de las variables auxiliares en la unidad muestral k.

La solución del problema depende de la función de distancia G que se utilice.

Si se considera la función de distancia lineal de argumento $z = \frac{w_k}{d_k}$:

$$G(z) = \frac{1}{2}(z-1)^2, \quad z \in R$$

el problema se resuelve mediante la utilización de los multiplicadores de Lagrange que conducen a la obtención de un conjunto de factores w_k que verifican las condiciones de equilibrado y proporcionan las mismas estimaciones que el estimador de regresión generalizado.

En el caso particular de la EPA se ha optado por utilizar la función de distancia lineal pero truncada (para evitar las soluciones negativas del sistema de ecuaciones), con

objeto de aprovechar las propiedades del estimador de regresión, de pequeña varianza y mínimo sesgo.

Para la solución práctica de este problema se ha utilizado el software CALMAR (CALage sur MARges) programado en SAS por el INSEE (Institut National de la Statistique et des Études Économiques) de Francia.

Desde el año 2021, en el proceso de calibrado de la EPA para estimaciones de viviendas y de población se utilizan variables auxiliares definidas en dos niveles de desagregación geográfica:

Provincia

- Población por grupos de edad (0-14,15-29,30-49, 50 y más) y sexo.

Comunidad autónoma

- Población por grupos quinquenales de edad y sexo.
- Población de 15 o más años, por nacionalidad española o extranjera.
- Viviendas por tamaño (1, 2, 3 y 4 o más) (*)

De esta forma con los estimadores actuales utilizados en la EPA se estiman correctamente los hogares por tamaño, la población por grupo de edad y sexo y el total de españoles y extranjeros mayores de 15 años por comunidad autónoma, así como los totales provinciales por grupos más agregados de edad y sexo.

En las ciudades autónomas de Ceuta y Melilla se agrupan las variables auxiliares para permitir el proceso de calibrado, dado el tamaño de la muestra:

- Población por grupos edad (0-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65 y más) y sexo considerando Ceuta y Melilla de forma conjunta.
- Viviendas por tamaño considerando Ceuta y Melilla de forma conjunta: (1, 2, 3 y 4 o más).
- Población de 15 y más años de Ceuta y Melilla, para cada una de ellas.
- Población de 15 y más años, por nacionalidad española o extranjera considerando Ceuta y Melilla de forma conjunta.

5 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.

(*) ***Nota:*** *Hasta 2020 el término hogar equivale a vivienda, pues se consideran miembros del hogar a todos los residentes en la vivienda. A partir de 2021, con la entrada en vigor de la nueva normativa europea asociada al Reglamento Marco de Estadísticas Sociales 2019/1700, para formar parte del hogar hay que compartir un presupuesto común y es posible que en una misma vivienda se encuentre más de un hogar.*

En el marco de la EPA se consideran cuatro tipos de actualizaciones:

Actualizaciones en el marco de secciones muestrales, consecuencia de las modificaciones (ver apartado 5.1) 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.

Actualizaciones en el marco de viviendas, con carácter anual se dispone del Marco de Direcciones Georreferenciadas actualizado, que tiene por objeto incorporar todos los cambios producidos en las viviendas del territorio nacional (cambios en la dirección postal, altas).

Actualización de las probabilidades de selección del seccionado. Con ella se pretende que, realizando la menor cantidad de cambios posible, la muestra de secciones sea equivalente a una muestra seleccionada en el año de la actualización. Se realiza cada tres años.

Actualización con carácter general relativa a todas las secciones y viviendas de la población, en la cual se revisa la definición de estratos y subestratos y se actualiza la probabilidad de selección de la sección. Se lleva a cabo con la información procedente de los Censos de Población (Ver apartado 5.2.2).

5.1 MODIFICACIONES EN LAS SECCIONES DE LA MUESTRA

Se consideran los siguientes casos:

5.1.1 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 ... 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 siga siendo 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:

1) 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.

2) Se selecciona una de las nuevas secciones S_j con probabilidad proporcional a su tamaño actualizado V'_{Sj} / V'_S

3) El número de viviendas que deben ser objeto de entrevista es

$$m_j = m.\frac{V'_S}{V_S}$$

Con m = 13

De esta manera la muestra continúa siendo autoponderada.

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

En este caso:

1) 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.

2) El número de viviendas que han de ser entrevistadas viene dado por

$$m_j = m.\frac{V'_{S'_j}}{V_{S'_j}}$$

Siendo

m = 13

$V'_{S'j}$ = 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 dentro de los límites de la nueva sección S'_j.

5.1.2 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 seleccionadas tengan unidades que investigar.

El caso de fusión de secciones no es sino un caso particular de la partición estudiada en el apartado 5.1.1.B.

Por tanto si la sección S_j seleccionada 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:

$$m_j = m.\frac{V'_{S'}}{V_{S'}}$$

siendo

m = 13

V'_S = Número de viviendas principales en la actualidad en la nueva sección S

V_S = Número de viviendas principales, según último Censo, dentro de los límites de la nueva sección S.

5.1.3 Variación de límites

Este 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.

En todos los casos antes expuestos, las nuevas secciones se incorporan a la muestra cuando por *Turno de rotación* corresponde renovar las familias en las secciones afectadas por dichas incidencias.

5.2 RENOVACIÓN DE LA MUESTRA. ACTUALIZACIÓN DE PROBABILIDADES DE SELECCIÓN

Cada tres años se lleva a cabo una actualización de las probabilidades de selección de las secciones en base a los datos más actuales de población disponibles.

Los cambios producidos en la muestra de secciones como consecuencia de la actualización se incorporan a la misma por turno de rotación es decir durante un periodo de seis trimestres, igual que en el caso de la renovación de viviendas. Por esta razón y con objeto de proporcionar una cierta estabilidad en las series temporales de la encuesta, las actualizaciones de las probabilidades del seccionado se realizan cada dos o tres años.

La forma más directa de actualizar las probabilidades de selección del seccionado es la selección de una nueva muestra a partir del marco disponible más actualizado. Pero un cambio tan radical en una encuesta continua, como es la EPA, genera los siguientes problemas:

- Pérdida de precisión en las estimaciones de variaciones trimestrales interanuales, al disminuir considerablemente la muestra común entre ambos periodos.
- Posible presencia de discontinuidades en la serie temporal de la encuesta, debidas a la causa citada en el apartado anterior, por el efecto del número de entrevista en la información y una carga de trabajo variable en el entrevistador.

Por ello se decidió arbitrar un procedimiento que, sin distorsionar las probabilidades de selección que realmente corresponden a cada sección, mantenga la muestra de secciones con las mínimas variaciones.

Se consideran dos tipos de actualizaciones de las probabilidades de selección en función de la información disponible para las mismas.

5.2.1 Actualizaciones realizadas cada tres años

En este caso no se modifica la definición de los estratos y se mantiene el que ya tiene asignado cada municipio, aunque su población haya cambiado y superado el límite del estrato inferior o el del superior. El procedimiento utilizado para la actualización es el propuesto por L. Kish y A. Scott(JASA 1971).

Sea S una sección perteneciente al estrato h, cuya probabilidad de selección en la anterior actualización (t-1) viene dada por:

$$Ps = \frac{V_s}{V_h} = \frac{\text{Viviendas en la sección S en } (t-1)}{\text{Viviendas en el estrato h en } (t-1)}$$

y supongamos que en el momento de la actualización(t), le corresponde una probabilidad de selección dada por

$$P's = \frac{V'_s}{V'_h} = \frac{\text{Viviendas en la sección S en } (t)}{\text{Viviendas en el estrato h en } (t)}$$

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.

5.2.2 Actualizaciones realizadas decenalmente

Decenalmente se procede a revisar las definiciones de estratos y subestratos, y a asignar a cada municipio el que le corresponda con arreglo a sus nuevas cifras de población.

Debido a lo anterior se producen múltiples cambios de estratos y el procedimiento de Kish-Scott resulta demasiado complejo y sin garantía de que sea óptimo, en el sentido de que no se demuestra que realice el menor número de cambios.

Por ello, en este caso se utiliza el método propuesto por J. M. Brick, R. Morganstein y CH. L. Wolter(Westat 1987), basado en el método de Kish y Scott del aparatado anterior.

Si las siguientes expresiones son las probabilidades de pertenecer a la muestra de la sección 'S' en la última actualización y en la nueva, respectivamente:

$$\pi_{hs} = n_h * \frac{V_s}{V_h} \qquad \pi'_{h^*s} = n'_{h^*} * \frac{V'_s}{V'_{h^*}}$$

donde n_h y n'_{h^*} son el número de secciones afijadas por estrato en 't-1' y en 't', en los estratos h y h* respectivamente.

Entonces:

- Si π'_{h^*S} es mayor que π_{hS} y la sección está en la muestra, entonces continúa en ella.
- Si π'_{h^*S} es mayor que π_{hS} y la sección **no** está en la muestra, entrará en la misma con probabilidad:

$$\frac{(\pi'_{h^*_s} - \pi_{hs})}{1 - \pi_{hs}}$$

- Si π'_{h^*S} es menor que π_{hS} y la sección estaba en la muestra, continúa en ella con probabilidad:

$$\frac{\pi'_{h^*_s}}{\pi_{hs}}$$

- Si π'_{h^*S} es menor que π_{hS} y la sección no estaba en la muestra, no tiene posibilidad de entrar en la misma.

Actuando de esta forma se demuestra que la probabilidad de una sección s de pertenecer a la muestra es π'_{h^*S} , es decir, la probabilidad actualizada en t en el nuevo estrato.

La principal cualidad de este algoritmo es la sencillez de su aplicación. Por el contrario, presenta el inconveniente de que no proporciona una muestra de tamaño fijo por estrato y por ello es necesario realizar un último ajuste, eliminando las secciones sobrantes con probabilidad igual y seleccionando las que falten con probabilidad proporcional al tamaño.

III Evaluación de la calidad de los datos

1 Introducción

Cuando a partir de los datos de una encuesta por muestreo se pretende estimar un parámetro poblacional, bajo la hipótesis de que se está utilizando un estimador apropiado, una estimación de aquel será de alta calidad si los datos en los que se basa la misma son de alta calidad. Y a la inversa, si los datos de la encuesta son de baja calidad, las estimaciones también serán de baja calidad.

Además, el tamaño muestral en el que se basan las estimaciones constituye también un importante determinante de la calidad. Incluso aunque los datos sean de gran calidad, una estimación basada en un número muy pequeño de observaciones resultará poco fiable. Por consiguiente, la calidad de un estimador de un parámetro poblacional es una función del ***error total de encuesta***, que engloba, por un lado, un error que deriva únicamente del hecho de seleccionar una muestra en lugar de llevar a cabo un censo completo, denominado ***error de muestreo***, y por otro, un error relacionado con los procedimientos de recogida y procesamiento de los datos, conocido como ***error ajeno al muestreo***.

La optimización del diseño de una encuesta implica encontrar un equilibrio entre los errores de muestreo y los ajenos al muestreo.

2 Errores de muestreo

Trimestralmente se calculan los errores de muestreo de las estimaciones de algunas de las principales características investigadas.

Para la obtención de los errores de muestreo se utiliza el método *Jackknife*.

Este procedimiento se basa en la formación de submuestras, en las que cada una de ellas se obtiene eliminando una unidad primaria de la muestra total. En las ciudades autónomas de Ceuta y Melilla, dado que el muestreo utilizado no es bietápico, en cada estrato se forman unas unidades primarias ficticias al dividir la muestra de viviendas en grupos aleatorios.

A partir de cada submuestra, o muestra jackknife, se obtiene la estimación trimestral de la característica cuyo error de muestreo queremos obtener. Esta estimación se calcula de igual manera que la estimación trimestral sobre la muestra completa, es decir, incluidos los ajustes de falta de respuesta y de calibración.

Una vez calculadas todas las estimaciones con cada una de las muestras jackknife, así como la estimación con la muestra completa, el estimador de la varianza viene dado por la expresión:

$$\hat{V}(\hat{Y}) = \sum_{h} \frac{n_h - 1}{n_h} \sum_{j \in h} \left(\hat{Y}_{(hj)} - \hat{Y}\right)^2$$

donde:

$\hat{Y}_{(hj)}$ es la estimación basada en la muestra jackknife obtenida al quitar de la muestra completa la unidad primaria j del estrato h.

$\hat{Y}$ es la estimación basada en la muestra completa.

n_h es el número de unidades primarias en el estrato h.

En las tablas se publica el error de muestreo relativo en porcentaje (coeficiente de variación), que viene dado por la siguiente expresión:

$$C\hat{V}(\hat{Y}) = \frac{\sqrt{\hat{V}(\hat{Y})}}{\hat{Y}} \cdot 100$$

3 Errores ajenos al muestreo

Los errores ajenos al muestreo son errores que se presentan en cualquiera de las etapas del desarrollo de la encuesta. Provienen, entre otras, de las siguientes fuentes:

1) **Errores de especificación y medida**: estos errores se producen cuando no coincide lo que se pretende medir o averiguar mediante la encuesta y lo que realmente se obtiene en el proceso de la entrevista. Pueden tener múltiples causas: conceptos o definiciones que no están bien especificados o son confusos para los informantes, preguntas que no están redactadas correctamente, formulación inadecuada por parte del entrevistador, respuesta inadecuada por parte del entrevistado,…

2) **Errores de marco**: tienen lugar cuando hay elementos de la población que están omitidos o duplicados en el marco de muestreo o cuando hay elementos incluidos en el mismo que no deberían estarlo (elementos erróneamente incluidos).

3) **Errores por falta de respuesta**: aquí se pueden distinguir tres tipos:

   - **Falta de respuesta de la unidad**: ocurre cuando un elemento de la muestra no colabora en la encuesta por diferentes motivos (negativa a colaborar, ausencia, incapacidad para contestar, etc.).
   - **Falta de respuesta a una o varias preguntas**: ocurre cuando el cuestionario se ha cumplimentado solo parcialmente, debido a que ha habido preguntas que han quedado sin contestar.
   - **Respuesta incompleta**: se produce cuando en preguntas abiertas el informante proporciona alguna información, pero la respuesta es demasiado escueta como para permitir una codificación adecuada.

Los métodos para la evaluación de estos errores son generalmente costosos y difíciles de llevar a la práctica. En la EPA, actualmente, el estudio de los errores ajenos al muestreo se centra en el análisis de los errores debidos a defectos del marco y a la falta de respuesta de las unidades informantes

Los errores de marco y los debidos a la falta de respuesta originan situaciones, denominadas incidencias, que hacen que algunas unidades no lleguen a colaborar en la encuesta.

Por un lado, se lleva a cabo una cuantificación de las incidencias según diversas variables, como son el método de recogida de datos utilizado, el número de entrevista (primera o sucesiva), el estrato a que pertenece la unidad (capitales de provincia y resto

de municipios). Igualmente se obtiene la distribución de incidencias por comunidades autónomas.

Por otro lado, se realiza un estudio específico de aquellas unidades seleccionadas que son encuestables pero que se negaron a facilitar los datos solicitados.

Para estas unidades que se niegan a colaborar en la encuesta se cumplimenta un ***cuestionario de negativas***, en el que se intenta recoger una serie de características mínimas básicas.