INSTITUTO NACIONAL DE ESTADISTICA

Encuesta de Población Activa

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

Informe Técnico

Madrid, Octubre 2009
Edición revisada: abril 2012
Área de Diseño de Muestras y Evaluación de Resultados

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. El diseño de la encuesta se enmarca en el de la Encuesta General de Población (EGP).

Desde su implantación ha sufrido diferentes modificaciones, siempre dirigidas a una mejora en la realización de la encuesta.

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.

3) Los datos son obtenidos por autoenumeración por lo que existen, en muchos casos dificultades de interpretación de los conceptos utilizados, por parte del informante.

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

El Paro Registrado 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.

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.

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 toda o la mayor parte del año 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 0,6 por ciento de la población total según datos del Censo 2001.

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 varía considerablemente a lo largo del tiempo, por lo que con referencia 1 de enero de cada año, coincidiendo con la revisión del Censo Electoral 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.

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.

Las unidades de primera etapa están constituidas por las **secciones censales**. La muestra de secciones permanece fija indefinidamente con las excepciones siguientes:

a) Salen de la muestra aquellas secciones en las que ya se han visitado todas las viviendas encuestables.

b) Cuando en el proceso de actualización del seccionado(ver apartado 5) a algunas secciones les corresponda salir de la muestra, bien por los cálculos probabilísticos, bien por cambios en la afijación por estratos.

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 (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.

4.2 ESTRATIFICACIÓN DE LAS UNIDADES DE MUESTREO

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 con la importancia demográfica del municipio al que pertenecen.

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

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 Alicante desaparece el estrato 9 por no tener población suficiente, y por ello los municipios de menos de 2.000 habitantes pasan al estrato 8. 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.

Cada diez años, con la información procedente de los Censos de Población, se actualiza la definición de los estratos en cada provincia.

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 de los estratos 7, 8 y 9. 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.

b- Resto de secciones. Estas 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 del Censo 2001 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:

Porcentajes 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

Se han tomado además las 18 variables de porcentajes de la población ocupada de la sección según su condición socioeconómica, cuya clasificación es:

01 Empresarios agrarios con asalariados

02 Empresarios agrarios sin asalariados

03 Miembros de cooperativas agrarias

04 Directores y jefes de empresas o explotaciones agrarias

05 Resto de trabajadores agrarios

06 Profesionales, técnicos y asimilados por cuenta propia

07 Empresarios no agrarios con asalariados

08 Empresarios no agrarios sin asalariados

09 Miembros de cooperativas no agrarias

10 Directores de empresas no agrarias y altos funcionarios

11 Profesionales técnicos y asimilados por cuenta ajena

12 Jefes departamento administración, comerciales o servicios empresas no agrarias

13 Resto del personal administrativo y comercial

14 Resto del personal de los servicios

15 Contramaestres y capataces no agrarios

16 Operarios cualificados y especializados no agrarios

17 Operarios sin especializar no agrarios

18 Profesionales de las Fuerzas Armadas

Finalmente, las variables fiscales utilizadas han sido:

Renta total por vivienda con perceptores

Renta Capital mobiliario e inmobiliario sobre renta total.

Renta agraria sobre renta total

(No se ha dispuesto de estas últimas variables en el País Vasco)

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 las tres variables fiscales, 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.

Además se han eliminado del análisis de conglomerados aquellas variables que representan menos del 1 por ciento del total del estrato en cada provincia, para tratar de evitar subestratos demasiado pequeños.

Se ha utilizado un algoritmo acumulativo que obtiene conglomerados jerárquicos. Se parte de tantos conglomerados como secciones y, en cada etapa, se unen los conglomerados más parecidos entre sí, es decir, con distancia mínima. Al final todas las secciones forman un único conglomerado. Por último se determina el punto intermedio del proceso de agrupación, en función del número y el tamaño de conglomerados que se considere más adecuado.

La distancia entre secciones es la euclídea y se define sobre las variables estandarizadas tal y como se ha explicado anteriormente. Entre conglomerados la distancia viene dada por el método de Ward, que tiende a producir conglomerados con un número parecido de secciones.

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

4.3 TAMAÑO DE LA MUESTRA

En el momento de implantación de la encuesta, el tamaño se estableció mediante la aplicación de un procedimiento de mínima varianza para coste fijo. Se partió de un presupuesto (Q) y a partir de él se procedió a determinar el número de secciones(n) y el número de viviendas(m) que minimizan la varianza de las estimaciones. Para ello se utilizó una función de coste de tipo lineal y la expresión del coeficiente de variación para una proporción en el muestreo de conglomerados con submuestreo.

Se empleó la siguiente función de coste:

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

donde:

Q = Presupuesto total

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 P viene dado por

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

siendo:

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

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 $CV^2(\hat{P})$ respecto de las variables m y n se obtiene calculando el mínimo de la expresión F (δ, 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, ...91, 100

y los correspondientes valores de n dados por

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

se obtienen distintos valores para F (δ, m, n).

El valor mínimo de F (δ, 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 3.000 secciones, investigándose una media de 20 viviendas por sección.

Posteriormente la muestra ha tenido diferentes ampliaciones con el objeto de dar cumplimiento a las exigencias de la Unión Europea y mejorar la representación de áreas más desagregadas. A partir del primer trimestre de 2005 se establece un tamaño muestral de 3.588 secciones y 18 viviendas por sección, excepto en las provincias de Madrid, Barcelona, Sevilla, Valencia y Zaragoza, en las cuales el número de entrevistas por sección es de 22. En el tercer trimestre de 2009 se firma un convenio con la Comunidad Autónoma de Galicia por el que se incrementa la muestra en esta comunidad hasta un total de 468 secciones y se asignan estratos separados a los municipios de Santiago de Compostela y Ferrol. **El tamaño de muestra final para el total nacional se establece en 3822 secciones**.

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.

En la afijación provincial se tuvieron en cuenta los siguientes aspectos:

a) Los resultados nacionales deben tener la mayor fiabilidad posible. A este respecto hay que recordar que, en general, cuanto más lejos se esté de la afijación óptima por provincias y estratos, mayor es la pérdida de precisión en la estimación nacional para un tamaño fijo de la muestra.

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

c) En cada provincia el número de secciones debe ser múltiplo de trece. Con ello se facilita la distribución de la muestra entre las semanas del trimestre. Para compatibilizar las tres condiciones antes expuestas se ha adoptado una afijación de compromiso entre la uniforme y la proporcional, a base de agrupar provincias de importancia demográfica similar y asignarles de 3 a 12 entrevistadores, es

decir de 39 a 156 secciones muestrales (con las excepciones de Ceuta y Melilla, que debido a su reducido tamaño poblacional tienen un solo agente y por tanto 13 secciones en la muestra cada una de ellas).

Dentro de cada provincia la afijación entre estratos es proporcional al tamaño de cada uno de ellos, si bien se ha 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-sociales y culturales 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).

En el cuadro 1 figura la distribución de las secciones de la muestra por provincias y estratos.

Cuadro 1

Distribución de las secciones de la muestra por provincias y estratos

Provincias	Estratos 1	2	3	4	5	6	7	8	9	Total
02 Albacete	18				8		3	5	5	39
03 Alicante/Alacant	18	10		15	12	12	8	3		78
04 Almería	16			9		5	3	6		39
01 Araba/Álava	30					3	6			39
33 Asturias	30	33		9	23	19	7		9	130
05 Ávila	13						3	8	15	39
06 Badajoz	20			6	10	6	15	11	10	78
07 Balears, Illes	44				27	12	12	9		104
08 Barcelona	55		33	19	21	12	10	3	3	156
48 Bizkaia	29	7		9	15	8	4	6		78
09 Burgos	20				7		3		9	39
10 Cáceres	19				7	5	10	12	25	78
11 Cádiz	13	13	6	26	7	6	7			78
39 Cantabria	35	10			8	11	9	9	9	91
12 Castellón/Castelló	26				26	6	4	7	9	78
13 Ciudad Real	13	9			14	15	12	8	7	78
14 Córdoba	34				13	9	11	11		78
15 Coruña, A	42	14	12		24	26	26	12		156
16 Cuenca	10						8	6	15	39
20 Gipuzkoa	26			6	13	20	7	6		78
17 Girona	15				19	12	10	13	9	78
18 Granada	28			5	5	12	12	10	6	78
19 Guadalajara	20				4			6	9	39
21 Huelva	15					13	3	8		39
22 Huesca	13					10	6		10	39
23 Jaén	17	7			15	12	12	15		78
24 León	24	10				10	15		19	78
25 Lleida	15					5	3	6	10	39
27 Lugo	26					14	14	24		78
28 Madrid	92		30	15	9	4	6			156
29 Málaga	36			10	18	5	9			78
30 Murcia	36	18		6	26	12	6			104
31 Navarra	36				9	9	8	15	14	91
32 Ourense	30					12	10	14	12	78
34 Palencia	20						5	5	9	39
35 Palmas, Las	44			9	28	14	9			104
36 Pontevedra	18	52			24	36	16	10		156
26 Rioja, La	33					9	7	7	9	65
37 Salamanca	20					4	3		12	39
38 Santa Cruz de Tenerife	25	15			24	10	10	7		91
40 Segovia	16						5	3	15	39
41 Sevilla	52			11	20	18	10	6		117
42 Soria	18						8		13	39
43 Tarragona	19	12			12	9	9	9	8	78
44 Teruel	10					4	9		16	39
45 Toledo	13	13				7	11	21	13	78
46 Valencia/València	45			10	24	19	7	7	5	117
47 Valladolid	36					4	6		6	52
49 Zamora	16					4			19	39
50 Zaragoza	59					4		6	9	78
51 Ceuta	13									13
52 Melilla	13									13
Total	1.384	223	81	165	472	447	397	314	339	3.822

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 del último Censo o del Padrón Continuo. 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. Para esta encuesta se ha determinado seleccionar 18 viviendas por sección (ver apartado 4.3)

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:

$$P(V_{ijh}) = P(S_{jh}) \times P(V_{ijh}/S_{jh}) = K_h \times \frac{V_{jh}}{V_h} \times \frac{18}{V_{jh}} = K_h \times \frac{18}{V_h}$$

siendo:

$P(S_{jh})$ = 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.

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.

4.6 DISTRIBUCIÓN DE LA MUESTRA EN EL TIEMPO

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. 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 totalidad de la muestra está dividida en tres submuestras independientes representativas, cada una de ellas, de toda la población.

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 las excepciones que figuran 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 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 la muestra durante seis trimestres consecutivos, al cabo de los cuales sale de la misma para ser reemplazada por otra de la misma sección.

Para poder realizar la renovación de viviendas de una forma adecuada cada trimestre se actualiza el marco de viviendas de las secciones del turno de rotación cuyas viviendas son entrevistadas por sexta y última vez. De esta forma se pueden incorporar a la muestra en el siguiente período aquellas viviendas, tanto de nueva construcción como las que se han transformado en viviendas familiares, las cuales, cuando se realizó el último Censo o Padrón no existían o se encontraban desocupadas o destinadas a otras finalidades diferentes a la de vivienda principal.

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

La distribución del número de secciones por estrato y semana es similar en cada turno de rotación. De esta forma se trata de evitar posibles sesgos de medida en las estimaciones, debidos al diferente comportamiento de las familias colaboradoras en función del tiempo que lleven en la encuesta.
Cada turno de rotación puede ser considerado, por tanto, como una submuestra representativa. Este hecho facilita la obtención de estimaciones de variables estructurales mediante la unión de dichas submuestras.

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, siendo 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, y donde:

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 $U = \{u_1 \ldots\ldots u_N\}$ de la cual se extrae una muestra

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

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 verifi-

can 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.

Como variables auxiliares se utilizan:

- Población de 16 y más años por grupos de edad y sexo a nivel de comunidad autónoma(22 grupos).
- Población de 16 y más años por comunidad autónoma y nacionalidad española o extranjera.
- Población de 16 y más años por provincia
- Población menor de 16 años por grupos de edad y sexo(6 grupos) a nivel de comunidad autónoma
- Población menor de 16 años por provincia

De esta forma con los estimadores actuales utilizados en la EPA se estima correctamente la población por grupo de edad y sexo y el total de españoles y extranjeros mayores de 16 años por comunidad autónoma.

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

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.

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 restringido y exclusivo para las secciones de la muestra. Esta actualización, como ya se dijo en el apartado 4.7, tiene por objeto incorporar las viviendas principales *altas* de la sección en la relación de viviendas de la misma.

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

$$m_j = 18 \frac{V'_s}{V_s}$$

las cuales son seleccionadas sistemáticamente.

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

$$m_j = 18 \frac{V'_{S'_j}}{V_{S'_j}}$$

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 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:

$$m = 18 \frac{V'_S}{V_S}$$

siendo

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

V_s = Número de viviendas principales, según último Censo, 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 COMO CONSECUENCIA DE LA ACTUALIZACIÓN DE LAS PROBABILIDADES DE SELECCIÓN

Cuando se dispone de información, procedente de los ficheros electorales, Censos de Población o del Padrón Continuo, se procede a actualizar las probabilidades de las secciones y a ajustar a 18 el número de entrevistas por sección.

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 tres tipos de problemas:

- Pérdida de información imprescindible para la selección y visita de las viviendas que resulten elegidas en la segunda etapa. Esta información, que es necesario rehacer, tiene aspectos tangibles como los directorios de viviendas o la planimetría de la zona, y otros intangibles pero no menos importantes, como el conocimiento por parte de la población de la sección de la figura del entrevistador, hecho éste que facilita el acceso a las familias y disminuye notablemente la falta de respuesta.
- 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 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 con información procedente del Padrón Continuo.

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 con información procedente del Censo de Población.

Al disponer de una información más completa 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:

$$\left(\pi'_{h^*s} - \pi_{hs}\right) / \left(1 - \pi_{hs}\right)$$

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

$$\pi'_{h^*s} \ / \ \pi_{hs}$$

- 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 que resulta de aplicación bastante sencilla en una situación compleja. 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

Los errores que afectan a toda encuesta pueden agruparse en dos grandes grupos:

Errores de muestreo, que se originan por la obtención de resultados sobre las características de una población, a partir de la información recogida en una muestra de la misma.

Errores ajenos al muestreo, que son comunes a toda investigación estadística, tanto si la información es recogida por muestreo como si se realiza un Censo. Estos errores se presentan en cualquier fase del proceso estadístico:

- Antes de la toma de datos: por deficiencias del marco e insuficiencias en las definiciones y cuestionarios.

- Durante la toma de datos: por defectos en la labor de los entrevistadores e incorrecta declaración por parte de los informantes.

- Tras la recogida de los datos: errores en la depuración, codificación, grabación, tabulación , etc. de los resultados.

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 de las *semimuestras reiteradas*.

Este procedimiento consiste en obtener sucesivas semimuestras de la muestra inicial. A partir de cada semimuestra se calcula la estimación de la característica de la que queremos obtener el error de muestreo. Una vez calculadas todas las estimaciones con cada una de las semimuestras, así como la estimación con la muestra completa, el estimador de la varianza viene dado por:

$$\hat{V}(\hat{Y}) = \frac{1}{r}\sum_{i=1}^{r}(\hat{Y}_i - \hat{Y})^2$$

donde:

r : es el número de semimuestras obtenidas, esto es el número de reiteraciones

$\hat{Y}_i$: es la estimación obtenida con la i-ésima reiteración

Para cada reiteración se repite el proceso de estimación general, es decir, se aplica la técnica de reponderación utilizando el software CALMAR.

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

En el caso de la EPA el número de reiteraciones que se utiliza es de 40. Para formarlas se procedió de la siguiente forma:

a) Se agruparon todas las secciones de cada estrato por pares, procurando que las dos secciones de cada par pertenecieran al mismo turno de rotación de la EPA.

b) Se asignó aleatoriamente la primera sección de cada par a 20 reiteraciones y la otra sección a las otras 20.

De esta forma cada reiteración queda constituida por un número de secciones equivalente al 50 por ciento de la muestra (semimuestra) y cada sección aparece en la mitad de las reiteraciones.

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

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

3 Errores ajenos al muestreo

El estudio de los errores ajenos al muestreo presenta numerosas dificultades debido a la gran variedad de causas que los originan, así como a las hipótesis en que se basan los modelos teóricos que, en general, no se cumplen en la realidad, lo que lleva a obtener resultados aproximados.

En la EPA el análisis de los errores ajenos al muestreo se basa en el modelo matemático elaborado por la Oficina de Censos de los Estados Unidos, debido a Hansen, Hurwitz y Bershad, y que, operativamente, consiste en repetir las entrevistas de la encuesta en una submuestra de la muestra de viviendas originalmente seleccionada. Posteriormente se cotejan los datos obtenidos en ambas ocasiones, con objeto de investigar las inconsistencias y cuantificar los errores mediante la aplicación de diversos índices de calidad.

Aparte de la *entrevista repetida* 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 recogen una serie de características básicas, como son el sexo, la edad y la relación con la persona principal de la persona que rehúsa ser entrevistada, así como la edad, el sexo, la nacionalidad, los estudios terminados, la relación con la actividad, la rama de actividad y la ocupación de la persona principal.

3.1 ENCUESTA DE EVALUACIÓN

Esta encuesta persigue un doble objetivo:

- Controlar el trabajo de recogida de la información en todas las comunidades autónomas.
- Evaluar la calidad de los resultados.

La comparación de los resultados obtenidos en la encuesta de evaluación (entrevista repetida, ER) con los obtenidos en la entrevista original (EO) permite evaluar dos grandes tipos de errores ajenos al muestreo:

a) Errores de cobertura, producidos por la omisión o por la inclusión errónea de unidades (viviendas y personas) en la encuesta original.

b) Errores de contenido, que afectan a las características investigadas en las personas encuestables.

El trabajo de campo se lleva a cabo por agentes especializados, los cuales realizan la entrevista repetida a lo sumo tres semanas después de la original, refiriéndose los datos de ambas entrevistas al mismo período de tiempo.

El hecho de que más del 70 por ciento de las negativas por primera vez se producen en la primera entrevista a las familias, unido a la existencia de dificultades técnicas para la realización de la encuesta de evaluación (ER) con CATI, han determinado que se investiguen en ER únicamente **secciones que en EO se encuentran en primera entrevista**. El método de recogida utilizado en estas secciones, tanto en EO como en ER, es CAPI.

Como consecuencia de lo anterior se dispone de menos muestra en la encuesta de evaluación, respecto a años anteriores, por lo que las cuatro muestras trimestrales se van a agrupar para ofrecer los resultados en cómputo anual, a fin de que éstos sean más representativos.

Para la selección trimestral de la muestra de la encuesta de evaluación se han creado cuatro zonas, agrupando en cada una varias comunidades autónomas, de forma que cada una de éstas esté incluida en una y sólo una de las zonas.

Cada semana se investigan las secciones (de primera entrevista) de la muestra en una de las zonas, siendo aleatoria la asignación de las semanas a las zonas, de modo que cada una de éstas se investigue al menos en tres de las semanas del trimestre.

De este modo se investigan aproximadamente entre 130 y 150 secciones cada trimestre.

En las secciones seleccionadas se repite la entrevista en la mitad de las viviendas, utilizándose en ER un cuestionario ligeramente reducido respecto al de EO, es decir, con algunas preguntas menos.

Con este procedimiento se investiga un número de viviendas de entre 1.300 y 1.500, lo que representa aproximadamente un 2 por ciento de la muestra de la EPA.

Además de la encuesta de evaluación, y con objeto de detectar errores cometidos en el proceso de actualización de las secciones de la muestra, cada trimestre se selecciona una muestra de cincuenta secciones (una de cada provincia, salvo Ceuta y Melilla) para evaluar la calidad de las actualizaciones.

3.2 ERRORES DE COBERTURA

Con la comparación de los resultados obtenidos en ambas entrevistas se obtienen indicadores sobre la cobertura de viviendas y la de personas, así como indicadores sobre los errores de contenido.

Cobertura de viviendas: se obtienen las viviendas que son encuestables en ambas entrevistas, las encuestables en ER y no en EO y viceversa.

Cobertura de personas: para estudiar los errores en la cobertura de personas, éstas se clasifican en:

- Personas cotejables, son aquellas que ambos agentes han considerado encuestables.

- Personas omitidas, son aquellas cuyos datos ha recogido el agente de ER por considerarlas encuestables, pero de las que no existe información en la EO.

- Personas erróneamente incluidas, que figuran en la EO pero no en la ER, por considerar el agente de entrevista repetida que no eran encuestables.

3.3 ERRORES DE CONTENIDO

Los datos sobre errores de contenido se basan en la información suministrada en las dos entrevistas por las personas clasificadas como cotejables.

Para facilitar el análisis de los datos se confeccionan dos tipos de tablas: tablas de concordancia y tablas de indicadores de calidad.

Así para una característica C con las modalidades M_1, ..., M_k, la tabla de concordancia presenta el siguiente formato:

E.R. \ E.O.	Total personas	M_1	M_2	...	M_j	...	M_k
Total personas	**n**	$\mathbf{n_{.1}}$	$\mathbf{n_{.2}}$	...	$\mathbf{n_{.j}}$	...	$\mathbf{n_{.k}}$
M_1	$\mathbf{n_{1.}}$	n_{11}	n_{12}	...	n_{1j}	...	n_{1k}
M_2	$\mathbf{n_{2.}}$	n_{21}	n_{22}	...	n_{2j}	...	n_{2k}
..	...	...	...	...	...	...	...
..	...	...	...	...	...	...	...
..	...	...	...	...	...	...	...
M_i	$\mathbf{n_{i.}}$	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ik}
..	...	...	...	...	...	...	...
..	...	...	...	...	...	...	...
..	...	...	...	...	...	...	...
M_k	$\mathbf{n_{k.}}$	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kk}

n_{ij} representa el número de personas clasificadas en la modalidad M_i según la ER y en la M_j según la EO.

La diagonal principal (n_{ii}) representa el número de personas que han sido idénticamente clasificadas en ambas entrevistas.

Para cada modalidad M_i de la característica C se puede obtener la siguiente tabla reducida:

E.R. \ E.O.	Con la Modalidad M_i	Sin la Modalidad M_i	Total
Con la Modalidad M_i	a	b	a + b
Sin la Modalidad M_i	c	d	c + d
Total	a + c	b + d	n

Comparando con la tabla anterior tenemos las siguientes equivalencias:

$a = n_{ii}$ número de personas clasificadas en la modalidad M_i en ambas entrevistas.

$b = n_{i.} - n_{ii}$ número de personas clasificadas en la modalidad M_i en ER y en otra diferente en EO.

$c = n_{.i} - n_{ii}$ número de personas clasificadas en la modalidad M_i en EO y en otra distinta en ER.

$d = n - n_{i.} - n_{.i} + n_{ii}$ número de personas que se han clasificado en modalidad diferente a la M_i en ambas entrevistas.

$n = a + b + c + d$ total de personas que se han clasificado en ambas entrevistas respecto a la característica C estudiada.

En base a estas tablas reducidas se definen los siguientes indicadores de calidad para la modalidad M_i:

a) Porcentaje de idénticamente clasificados

$$P.I.C.(M_i) = \frac{a}{a+b} \times 100 = \frac{n_{ii}}{n_{i.}} \times 100$$

Varía entre cero y cien. Es un indicador de la estabilidad de respuesta. Su valor óptimo (100) expresa que todas las personas pertenecientes según la ER a la modalidad M_i se clasificaron de igual forma en la EO.

b) Índice de cambio neto

$$I.C.N.(M_i) = \frac{c-b}{a+b} \times 100 = \frac{n_{.i} - n_{i.}}{n_{i.}} \times 100$$

Puede ser positivo ($c > b$ o $n_{.i} > n_{i.}$) o negativo ($b > c$ o $n_{i.} > n_{.i}$). Es un indicador del sesgo de respuesta, expresado como porcentaje del número de personas clasificadas en M_i según la ER.

c) Tasa de diferencia neta

$$T.D.N.(M_i) = \frac{c-b}{n} \times 100 = \frac{n_{.i} - n_{i.}}{n} \times 100$$

Similar al anterior, pero en este caso es un porcentaje respecto al total de personas que se han clasificado en ambas entrevistas respecto a la característica de referencia.

d) Índice de cambio bruto

$$I.C.B.(M_i) = \frac{c+b}{a+b} \times 100 = \frac{n_{.i} + n_{i.} - 2n_{ii}}{n_{i.}} \times 100$$

Puede ser nulo o positivo. Es un indicador de la varianza de respuesta.

e) Tasa de diferencia bruta

$$T.D.B.(M_i) = \frac{c+b}{n} \times 100 = \frac{n_{.i} + n_{i.} - 2n_{ii}}{n} \times 100$$

Similar al anterior, pero referido al total de personas clasificadas en ambas entrevistas respecto a la característica en estudio.

Para comparar la calidad general de las distintas características evaluadas se utiliza el **índice de consistencia global**, que para cada característica C se obtiene a

partir de la tabla en la que aparecen todas las modalidades de la misma. Se define como

$$I.C.G.(C) = \frac{\sum_{i=1}^{k} n_{ii}}{n} x100$$

Un valor de I.C.G. = 100 indica inexistencia de errores en la clasificación.