PRIMEROS CONCEPTOS EN LA INVESTIGACIÓN POR MUESTREO

CONCEPTO DE MUESTREO

En toda investigación estadística existe un conjunto de elementos sobre los que se toma información. Este conjunto de elementos es lo que se denota con el nombre de población o universo estadístico. Cuando el estadístico o el investigador toma información de todos y cada uno de los elementos de la población estadística se dice que está realizando un ***censo***. Sin embargo, esto no es muchas veces posible, ya sea por el coste que resulta de la toma de información, o bien porque la toma de información lleve consigo la destrucción de los elementos en cuestión, o que la población tenga infinitos elementos, o por otras causas.

Este problema lleva al investigador a tomar la información sólo de una parte de los elementos de la población estadística, proceso que recibe el nombre de ***muestreo***. El conjunto de elementos de los que se toma información en el proceso de muestreo se llama ***muestra*** y el número de elementos que la componen ***tamaño muestral***. Existen varios tipos de muestreo, dependiendo de que la población estadística sea finita o infinita, materia sobre la que existe amplia literatura estadística, pero nosotros consideraremos solamente el ***muestreo en poblaciones finitas***. El investigador utiliza la muestra para la toma de información, pero lo importante es que dicha muestra sea representativa.

Por lo tanto, entenderemos por muestra un subconjunto lo más representativo posible de una población. Naturalmente, el estudio del colectivo requerirá un cuidadoso proceso de muestreo, a los efectos de elaborar ese subconjunto en las condiciones más adecuadas de representatividad, puesto que la inferencia se caracterizará por aplicar al colectivo las conclusiones obtenidas a partir de la muestra. En este sentido, el método de selección de la muestra reviste una singular importancia, dado que dependiendo de cómo se haya constituido ésta se seguirán unos u otros resultados. Precisamente los métodos de selección de la muestra serán el núcleo principal a tratar en este libro.

Con la finalidad de medir el grado de representatividad de la muestra lo mejor posible es necesario utilizar muestreo probabilístico. Diremos que el muestreo es probabilístico cuando pueda establecerse la probabilidad de obtener cada una de las muestras que sea posible seleccionar, esto es, cuando la selección de muestras constituya un fenómeno aleatorio probabilizable. Dicha selección se verificará en condiciones de azar, siendo susceptible de medida la incertidumbre derivada de la misma. Esto permitirá medir los errores cometidos en el proceso de muestreo.

Se denomina ***inferencia estadística*** *o* ***estadística inductiva*** a la metodología consistente en inferir resultados, predicciones y generalizaciones sobre la población estadística, basándose en la información contenida en las muestras representativas previamente elegidas por métodos de muestreo formales. La inferencia estadística está basada en la teoría de la probabilidad, pero tiene un carácter diferente. En inferencia estadística se consideran fenómenos en los que se manifiesta la regularidad estadística y se construyen modelos probabilísticos para describirlos.

Podemos definir los ***métodos de muestreo*** como el conjunto de técnicas estadísticas que estudian la forma de seleccionar una ***muestra lo suficientemente representativa*** de una población cuya información permita inferir las propiedades o características de toda la población cometiendo un ***error medible y acotable***.

A partir de una muestra, seleccionada mediante un determinado método de muestreo, se estiman las características poblacionales (media, total, proporción, etc.) con un error cuantificable y controlable. Las estimaciones se realizan a través de funciones matemáticas de la muestra denominadas ***estimadores***, que se convierten en variables aleatorias al considerar la variabilidad de las muestras. Los errores se cuantifican mediante varianzas, desviaciones típicas o errores cuadráticos medios de los estimadores, que miden la precisión de los mismos.

La teoría del muestreo proporciona una técnica estadística de carácter muy práctico que sencillamente busca obtener datos de una población (hogares, empresas, árboles, etc.) en su totalidad, utilizando tan sólo una parte reducida de la misma, denominada muestra, aunque como es lógico pagando algún coste (calculable) en cuanto a la precisión de las medidas poblacionales inferidas.

De forma metafórica podríamos decir que una muestra, que se supone representativa de una población, es similar a lo que representa una maqueta respecto del edificio del que ofrece una imagen. La muestra, al igual que la maqueta, será mejor o peor, según el grado de representatividad que ofrezca. La teoría del muestreo traslada la información aportada por la muestra a toda la población, dando lugar a lo que se conoce en muestreo como ***elevación del dato muestral a la población*** que se estudia. En la metáfora de la maqueta el factor de elevación sería la escala de la misma, que permite pasar un dato de la maqueta a su correspondiente dato para el edificio real que representa.

POBLACIÓN, MARCO Y MUESTRA

Una tarea importante para el investigador es definir cuidadosa y completamente la población antes de recolectar la muestra. Inicialmente una población es una colección de elementos acerca de los cuales deseamos hacer alguna inferencia. Esta población inicial que se desea investigar se denomina ***población objetivo***.

Pero el muestreo de toda la población objetivo no es siempre posible. Existirán problemas que van a impedir obtener información de algunos de sus elementos. Entre estos problemas cabría destacar las negativas a colaborar, las ausencias, la inaccesibilidad a algunos elementos o los errores en los intrumentos de medida de la característica que se estudia en los elementos de la población. Por lo tanto la población objetivo se ve restringida a la hora de obtener la información de sus elementos, dando lugar al concepto de ***población investigada***, que es la población que realmente es objeto de estudio.

Por otra parte, una unidad de muestreo puede ser un simple elemento de la población, en cuyo caso estamos ante una ***unidad elemental de muestreo***. Pero también pueden considerarse unidades de muestreo que sean grupos no solapados de elementos de la población que cubren la población completa, en cuyo caso estaríamos ante una ***unidad de muestreo compuesta*** de varias unidades elementales, también denominada a veces ***unidad primaria***. De esta forma se puede establecer una ***jerarquía de unidades de muestreo*** en el sentido de que el primer nivel lo formarían las unidades elementales, el segundo nivel lo formarían grupos de unidades elementales, el tercer nivel lo formarían grupos de unidades de segundo nivel, y así sucesivamente. En el muestreo aleatorio simple suelen utilizarse unidades elementales, en el muestreo estratificado y por conglomerados monoetápico se utilizan unidades de muestreo compuestas de segundo nivel (estratos y conglomerados respectivamente). En el muestreo polietápico se generaliza a unidades de niveles superiores según el número de etapas de dicho tipo de muestreo. En todo caso, las ***unidades de muestreo*** han de ser grupos no solapados (de intersección vacía) de elementos de la población que cubran la población objetivo. En el caso de que las unidades de muestreo sean elementales, una unidad de muestreo y un elemento de la población son idénticos.

Pero para poder seleccionar el conjunto de unidades de muestreo que componen la muestra, será necesario disponer de un listado material de unidades de muestreo. Esta relación de unidades de muestreo, de la que se selecciona la muestra, se denomina ***marco***. Lo ideal sería disponer de un marco tal que la lista de unidades muestrales que lo componen coincida con la población objetivo. Pero en la práctica el marco contiene impurezas debidas a desactualizaciones, errores, omisiones y otras causas que hacen que el marco no coincida con la población objetivo, lo que no impide que el marco sea la contrapartida en el mundo real de la población objetivo. De todas formas, la separación entre el marco y la población objetivo ha de ser lo suficientemente pequeña como para permitir que se hagan inferencias acerca de la población basándose en una muestra obtenida del marco.

La imperfección del marco suele tener como origen la existencia de ***duplicaciones*** de algunas unidades, ***omisiones*** de otras y la presencia de unidades extrañas y vacías. Mantener un listado de unidades de muestreo actualizado es imposible en la práctica. Existirán unidades que deberían estar en el marco y que sin embargo por problemas de actualización u otros problemas similares se omiten en el mismo, y al revés, existirán unidades que aparecen en el marco y que ya no debieran estar en el mismo. Se suele denominar ***unidad vacía*** a una unidad de muestreo erróneamente incluida en el marco y que no pertenece a la población objetivo (aunque esté relacionada de alguna forma con la población objetivo). Se suele denominar ***unidad extraña*** a una unidad que aparece en el marco pero que no es realmente del marco y que de ninguna manera debiera constar en el mismo (no hay ninguna relación posible entre la unidad y la población objetivo). Por ejemplo, ante una encuesta para analizar características de la población española en la que se toma como marco un listado de viviendas, serán unidades vacías las viviendas deshabitadas. Como ejemplo adicional supongamos que para estimar la producción de leche en un país se toma como marco una lista de explotaciones agrícolas, en cuyo caso, las explotaciones que no se dediquen total o parcialmente a la producción de leche (que por cierto, serán muchas) son unidades extrañas, ya que de ninguna manera deberían aparecer en el marco.

Si eliminamos del marco las unidades erróneamente incluidas en él (unidades extrañas, unidades vacías y duplicaciones) y a su vez le añadimos las omisiones, obtenemos la población objetivo. Este proceso se conoce como ***depuración de marcos imperfectos***.

Por otra parte, si eliminamos del marco las unidades de las que no se puede obtener información para unos recursos dados (unidades inaccesibles, unidades que no colaboran ni responden, unidades ausentes, unidades medidas erróneamente, etc.), obtenemos la población investigada.

El concepto de ***marco en sentido restringido*** incluye únicamente el listado de unidades del que se va a extraer la muestra, pero también puede considerarse el concepto de ***marco en sentido amplio***, incluyendo adicionalmente al listado de unidades, la ***información complementaria***, es decir aquella información que puede y debe utilizarse para mejorar el diseño en los procesos de estratificación, selección, estimación, etc. Como ejemplos de esta información complementaria podríamos citar el conocimiento de una variable auxiliar correlacionada con la variable en estudio cuyos valores permiten realizar convenientemente estratos, o el conocimiento completo de una variable auxiliar correlacionada con la variable en estudio que nos va a servir de apoyo para utilizar un procedimiento de estimación indirecta (razón, regresión, diferencia), o el conocimiento de las estimaciones de determinadas características provenientes de una encuesta similar anterior o de una encuesta piloto.

A continuación se presenta un esquema de los conceptos vistos hasta ahora.

CONCEPTOS

- ***POBLACIÓN OBJETIVO → Población que se desea investigar***
  - *Problemas*
    - ***Inaccesibilidad a algunas unidades***
    - ***Negativas a colaborar***
    - ***Ausencias***
    - ***Errores en los instrumentos de medida***
- ***POBLACIÓN INVESTIGADA → Población que realmente es objeto de estudio (teniendo en cuenta los problemas citados)***
- ***POBLACIÓN MARCO → Lista de unidades de muestreo de entre las que se selecciona la muestra. Es la contrapartida en el mundo real de la población objetivo***
- ***UNIDAD EXTRAÑA → Unidad que no es realmente del marco y que no tiene relación de ningún tipo con la población objetivo***
- ***UNIDAD VACÍA → Unidad erróneamente incluida en el marco y que no pertenece a la población objetivo aunque esté directamente relacionada con ella***
- ***DUPLICACIONES → Unidades repetidas en el marco***
- ***OMISIONES → Unidades omitidas en el marco que son del marco realmente***
- ***MARCOS IMPERFECTOS → Marcos con unidades vacías, unidades extrañas, duplicaciones y omisiones***
- ***MARCO EN SENTIDO AMPLIO → Incluye información complementaria (variables auxiliares, encuestas piloto)***
- ***MUESTRA → Conjunto de unidades de muestreo seleccionadas de un marco o de varios marcos***
- ***UNIDAD ELEMENTAL DE MUESTREO → Elemento más simple de la población***
- ***UNIDAD DE MUESTREO COMPUESTA O PRIMARIA → Se compone de varias unidades elementales***

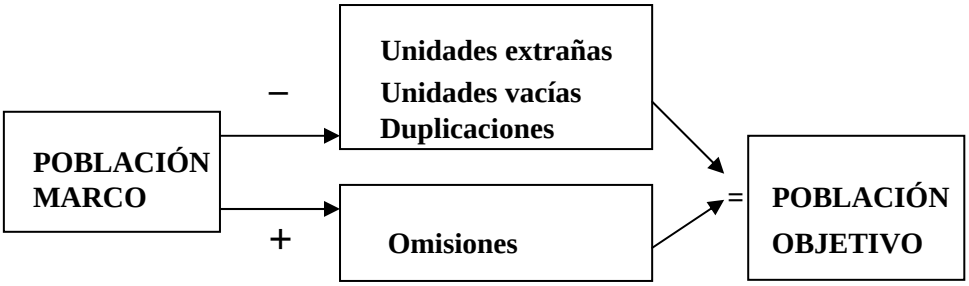
Un marco puede ser un listado de unidades elementales o de unidades compuestas, dependiendo del tipo de unidades de muestreo que se vayan a seleccionar en el proceso de muestreo. Cuando el marco es de unidades compuestas, puede ser posible disponer, adicionalmente al listado de unidades compuestas, de listados parciales de unidades simples dentro de cada unidad compuesta (por ejemplo, para realizar submuestreo). En este caso se dice que disponemos de ***marcos múltiples***. La existencia de marcos múltiples puede hacer el muestreo mucho más eficiente. Por ejemplo, los residentes de una ciudad pueden ser muestreados de una lista de manzanas de la ciudad relacionada con una lista de residentes dentro de las manzanas. El segundo marco puede no estar disponible hasta que las manzanas sean seleccionadas y estudiadas con cierto detalle.

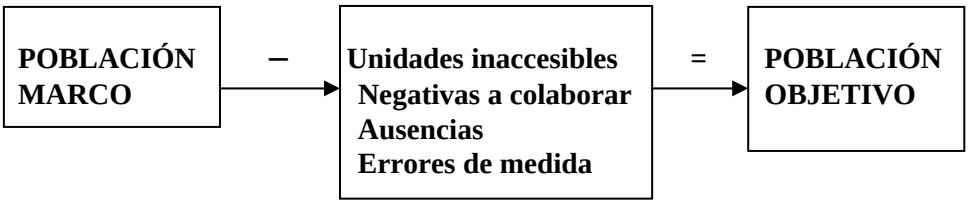
En general, una ***muestra*** es una colección de unidades de muestreo seleccionadas de un marco o de varios marcos.

Vamos a considerar un sencillo ejemplo que ilustre los conceptos definidos anteriormente. Supongamos que se trata de medir, mediante una encuesta, la posible influencia en el resultado de unas elecciones de una emisión de bonos justo antes de las mismas. En este caso la población objetivo estaría constituida por los votantes reales de la comunidad con derecho a voto en las inminentes elecciones. Evidentemente no será posible obtener respuesta de algunos de estos votantes, bien sea por problemas de inaccesibilidad a su domicilio derivados de errores en las direcciones, bien sea por su negativa a colaborar o contestar, bien sea porque no se encuentren en su domicilio en el momento de la encuesta, bien sea porque el cuestionario que se les pasa es erróneo, o por cualquier otro motivo. Si de la población de votantes reales restamos los votantes afectados por los problemas que acabamos de citar, obtendríamos la población investigada.

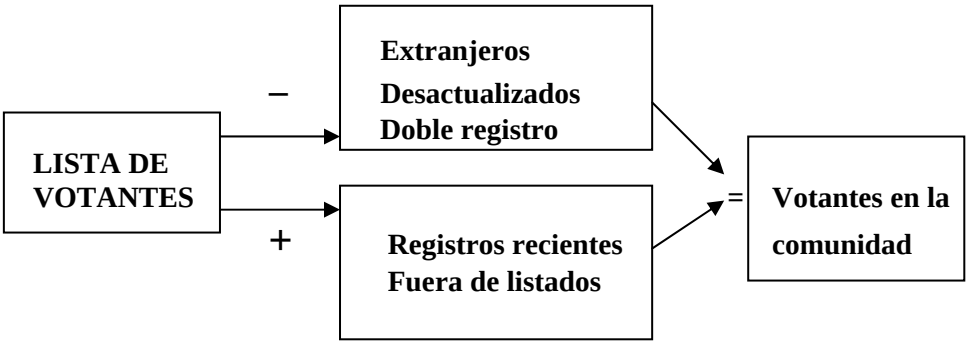
Para seleccionar la muestra de votantes que han de contestar a nuestra encuesta necesitamos un listado apropiado. En nuestro caso el listado ideal, es decir el marco, sería la relación oficial lo más actualizada posible de personas con derecho a voto registradas en la comunidad. Pero este listado ideal presentará diversos problemas. Habrá votantes en la lista que no podrán ejercer su derecho al voto el día de las elecciones porque se hayan cambiado recientemente de distrito electoral y se hayan inscrito en otra comunidad (unidades vacías). Puede haber votantes incluidos por error en la lista que sean extranjeros o que no tengan la edad para votar y que en ningún caso deberían estar en la lista (unidades extrañas). Puede haber votantes, incluidos por error en la lista más de una vez (duplicaciones). Puede haber votantes que no aparezcan en la lista y que hayan adquirido recientemente el derecho a voto, bien por haber entrado en las últimas fechas en edad de votar o bien por haberse domiciliado en la comunidad también en las últimas fechas (omisiones). Por lo tanto, para que el marco cubra lo mejor posible la población objetivo (la coincidencia es en la práctica es imposible), será necesario eliminar del marco las unidades extrañas, vacías y duplicaciones, y añadir las omisiones. Ya estaremos entonces en condiciones de seleccionar la muestra de entre la relación de votantes de este marco depurado.

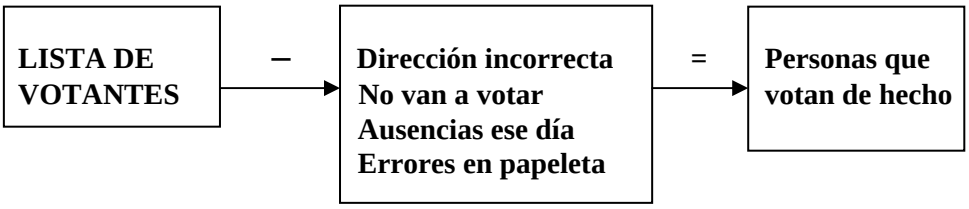
A continuación se presenta un esquema de los conceptos más importantes relativos a población, marco y muestra.



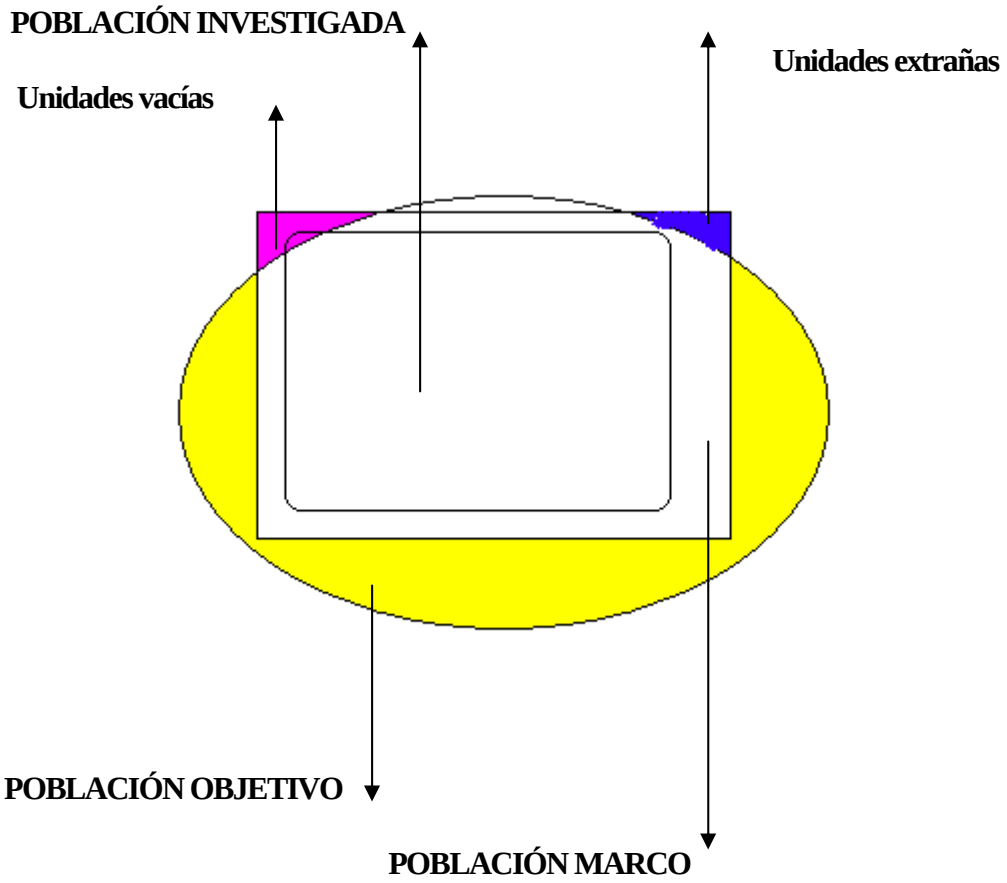


A continuación se presentan los conceptos del esquema anterior aplicados al ejemplo de los votantes expuesto anteriormente





La figura siguiente muestra un esquema en el que se identifican mediante diagramas de Venn los distintos conceptos estudiados en este capítulo.



LAS DISTINTAS FASES DE LA INVESTIGACIÓN POR MUESTREO

En su sentido amplio la finalidad de una encuesta por muestreo es obtener información para satisfacer una necesidad definida. La necesidad de recopilar datos muestrales de forma ordenada surge en todo campo de la actividad humana, por lo que es muy importante que el estadístico tenga una buena idea del trabajo que debe hacer en una encuesta por muestreo y de las limitaciones que confronta. A la hora de llevar a cabo una encuesta por muestreo es necesario tener en cuenta determinadas fases para su correcta planificación y ejecución.

Objetivos

La primera tarea de toda encuesta por muestreo es fijar en términos concretos los objetivos de la misma. Por lo general ocurre que el promotor de la encuesta no está seguro de lo que quiere ni de la forma en que va a utilizar los resultados. Es tarea del estadístico discutir con los promotores para hacerlos pensar en términos concretos. No aclarar la finalidad de la encuesta disminuirá su valor en última instancia, encontrándose al final de la misma con que los resultados no eran los que realmente se querían.

Por lo tanto, es necesario establecer los objetivos de la encuesta de una forma clara y concisa, y remitirse a estos objetivos conforme se vaya progresando en el diseño e instrumentación de la encuesta. Es vital mantener unos objetivos lo suficientemente simples para que sean entendidos por quienes trabajan en la encuesta y logrados con éxito cuando finalice la misma.

A la hora de fijar los objetivos es necesario tener presentes determinados factores como son:

- ¿Qué información se necesita para cumplirlos?
- ¿Cuál es el motivo de la encuesta?
- ¿Existe información disponible de antemano de encuestas piloto u otras encuestas similares que pueda ser aprovechada?
- ¿Existe información complementaria que se pueda utilizar para mejorar los procesos de estratificación, selección o estimación?
- ¿De qué medios materiales y personales se dispone?
- Límites presupuestarios y temporales
- Legislación y restricciones administrativas
- Oportunidad de fechas

Delimitación de la población objetivo y la población investigada

Un vez que se tiene claro el objetivo de la encuesta, es necesario definir cuidadosamente la población que va a ser muestreada, teniendo siempre presente que se va a obtener una muestra de esa población que ha de ser definida de tal manera que la selección de la muestra sea realmente factible. Por lo tanto tiene que estar clara la cobertura de la encuesta, eliminando de la población objetivo la parte de población ideal no accesible para obtener la población investigada.

En muchos casos, las dificultades prácticas para manejar ciertos segmentos de la población podrían apuntar a la eliminación de los mismos del campo de la encuesta. Por ejemplo, en una encuesta sobre la población, podría resultar muy difícil cubrir a la población trashumante.

En una investigación sobre la agricultura en la que se tiene la intención de considerar toda pequeña propiedad de tierra para determinar qué se cultiva, las consideraciones prácticas podrían obligar a la omisión de lugares como los pequeños huertos familiares. En una encuesta industrial se tendrían que omitir todas las fábricas que emplean menos de dos personas si se considera que sería muy difícil incluirlas en la muestra. De este modo, la población que se procurará cubrir (población objetivo) será por lo general diferente de la que es en realidad objeto de muestreo (población investigada). Los resultados que se obtengan a partir de la población investigada se aplicarán a toda la población objetivo, presentando adicionalmente información sobre los sectores omitidos (análisis de la falta de respuesta y de los errores de respuesta). Esto se hace mediante procedimientos no muy exactos pero que pueden arrojar alguna luz sobre el tema de la encuesta.

Establecimiento del marco

Con el fin de cubrir con la encuesta la población objetivo, debe haber alguna lista, mapa o algún otro material aceptable (marco) que sirva como guía al universo que se cubrirá. El marco debe examinarse para asegurarse que está razonablemente libre de defectos. Si no está al día, debe considerarse la posibilidad de actualizarlo. Existen técnicas específicas de depuración de marcos imperfectos, que serán abordadas en los últimos capítulos de este libro, cuya finalidad es eliminar del marco las unidades extrañas y vacías, así como cualquier otro tipo de errores u omisiones. Como la depuración total de marco es imposible, será necesario presentar datos sobre los errores de cobertura (errores cometidos en el muestreo por el desajuste entre población marco y población objetivo).

La investigación por muestreo se favorece por la existencia de una cierta ***infraestructura estadística previa***, que llamaremos ***infraestructura estadística deseable***, pero que además demanda como condición necesaria una ***infraestructura estadística mínima***, siendo ésta imprescindible para llevar a cabo el diseño muestral que posibilite la investigación por muestreo.

Dentro de esta infraestructura estadística mínima, indispensable para la investigación por muestreo, se enmarca la existencia de directorios o marcos convenientemente correctos y actualizados. Como infraestructura estadística deseable, añadida a la mínima y que favorece la investigación por muestreo aunque no la limita a ultranza, reseñamos la conveniente existencia de las infraestructuras complementarias siguientes:

- Infraestructura de definiciones, de habitual uso, relativa a variables (de gastos, ingresos, consumos...), y a unidades elementales o derivadas (hogar, empresa, local, unidad de producción...).

- Infraestructura de clasificaciones estadísticas. Una clasificación estadística constituye un instrumento básico que posibilita la coherencia entre la recogida, la tabulación y el análisis de los datos, siendo un elemento armonizador.

- Infraestructura de planimetría. Cuando la población sujeta a estudio o el diseño muestral se apoyan en una dimensión espacial, es necesario establecer una cartografía adecuada y codificación territorial (delimitación de espacios geográficos municipales, o de secciones...). Ello facilita la investigación muestral en aspectos de correcta localización geográfica y jerarquización de las unidades de muestreo.

- Infraestructura estadística de datos complementarios relativos a las unidades de la población de los directorios o marcos (por ejemplo, conocer el número de trabajadores de las empresas).

Resumiendo, en cuanto a lo que a infraestructura estadística se refiere, la situación se sintetiza en la necesidad de disponer de una información mínima y homogénea de las unidades de la población muestreada para posibilitar la investigación por muestreo y obtener posteriormente una información añadida a la preexistente. Por ejemplo, en el símil de maqueta de un edificio con muestra de una población, la información de infraestructura nos va a permitir construir la muestra (construir la maqueta) y disponer de los estimadores (de la escala) que permite elevar los datos de la muestra (maqueta) a la población (edificio real).

Es interesante resaltar que la informática ha jugado un papel considerable al potenciar, en cantidad y calidad, todos los elementos que participan en la investigación por muestreo. Concretamente, y en relación con el marco, contribuye a agilizar los procesos de creación de infraestructura estadística. Como ejemplo de marco informatizado para la realización de encuestas dirigidas a empresas e instituciones podríamos citar la reciente creación en el Instituto Nacional de Estadística español (INE) del Directorio Central de Empresas, en siglas DIRCE. El DIRCE trata de reunir en un directorio único todas las empresas españolas, siendo su objetivo básico hacer posible la realización de encuestas por muestreo dirigidas precisamente a las empresas, pues, evidentemente, una empresa no podrá ser seleccionada en una encuesta si no figura reseñada en el directorio, debidamente actualizado, del que se selecciona.

El DIRCE, actualmente con referencia a 1 de enero de 1995, relaciona por primera vez en España un total de 2.301.559 empresas clasificadas según actividad económica principal, según condición jurídica, por intervalos según número de asalariados, etc. Geográficamente existen desgloses provinciales (de este cómputo se excluye la agricultura, ganadería, pesca, las administraciones públicas, las actividades de comunidades de propietarios, el servicio doméstico y los organismos extraterritoriales). En el DIRCE se procesan anualmente del orden de seis millones de registros, tarea que sólo se puede ejecutar utilizando medios informáticos. Además el DIRCE se basa en registros administrativos ligados principalmente a la Administración Tributaria y a la Seguridad Social y si éstos no estuviesen operativos informáticamente, no hubiese sido posible la creación del DIRCE. Es evidente que, si existe un registro administrativo, aunque esté gestionando con puntualidad, si no está total o parcialmente informatizado, no ofrecerá la opera-tividad que es necesaria para su uso estadístico.

En esta situación la operatividad informática de los registros administrativos pasa a ser un elemento clave. El DIRCE es un marco esencial para las encuestas económicas oficiales, como por ejemplo la Encuesta de Salarios en la Industria y los Servicios, con periodicidad trimestral desde 1963, y que proporciona los datos básicos de ganancias por trabajador y hora trabajada, así como las horas trabajadas en promedio por cada trabajador. Las Encuestas Industriales, de Comercio Interior o la de Estructura de las Explotaciones Agrícolas son otros ejemplos de encuestas oficiales que utilizan como marco idóneo el DIRCE.

Otro ejemplo de marco informatizado acompañado de cartografía y planimetría es el referente a la división del territorio español en 40.000 secciones censales, de entre las cuales se seleccionan aproximadamente 3.000 para poder realizar cualquier tipo de encuestas oficiales que vaya dirigida a viviendas familiares o a las personas o grupos familiares que las habitan. Este marco es el utilizado por el INE en el diseño muestral de la Encuesta General de Población (EGP), a partir del cual se hace posible la realización de cualquier encuesta para obtener datos asociados a la población de viviendas de uso familiar o de los grupos humanos que en ellas residen. Sobre este diseño muestral se han podido construir en España, desde 1964, todas las Encuestas de Población Activa, las Encuestas de Presupuestos Familiares, Equipamiento y Nivel Cultural de las Familias, etc., y en general las encuestas del INE dirigidas, a viviendas familiares o a personas que las habitan. En términos comparativos podríamos decir que el DIRCE supone, respecto a las encuestas dirigidas a empresas e instituciones, el mismo avance que supuso el diseño de la Encuesta General de Población para las encuestas dirigidas a viviendas familiares y personas que las habitan, basado en el marco obtenido por la división del territorio nacional en secciones estadísticas.

Diseño de la muestra

Para los propósitos de la selección de la muestra debe ser posible dividir la población en lo que se ha denominado ***unidades de muestreo*** de forma no ambigua. Todo elemento de la población debe pertenecer a una sola unidad de muestreo.

Si, por ejemplo, la unidad es la familia, debe definírsela de tal forma que una persona no pertenezca a dos familias diferentes ni debe dejarse fuera a cualquier persona que perte-nezca a la población. Ésta no es una tarea fácil, ya que siempre hay casos dudosos, y deben establecerse algunas reglas arbitrarias para manejarlos. Una vez clarificadas sin ambigüedades las unidades de muestreo, los problemas técnicos que recibirán la más cuidadosa atención serán la forma en que se seleccionará la muestra y la estimación de las características de la población y de su margen de incertidumbre a partir de la misma. Estas cuestiones forman el núcleo central de la teoría del muestreo, que es el tema principal de este libro.

Son puntos importantes del diseño de la muestra los siguientes:

- Especificación de las unidades de muestreo
- Métodos estadísticos para la depuración del marco
- Posible utilización de la información complementaria
- Análisis y determinación del tamaño de la muestra
- Método de selección de la muestra, esto es, tipo de muestreo a utilizar
- Fórmulas para los estimadores a utilizar
- Fórmulas para la estimación de los errores de muestreo
- Métodos estadísticos para el tratamiento de la falta de respuesta
- Control de otros errores ajenos al muestreo

En lo que se refiere al uso de la informática para la elaboración de diseños muestrales óptimos, es evidente que la grabación de ingentes cantidades de datos procedentes de cuestionarios muestrales, la imputación de datos faltantes y el cálculo de estimadores para elevar los datos de la muestra a la población, en multitud de ocasiones de gran complejidad matemática, se han visto radicalmente potenciados respecto de lo que ocurría en la etapa preinformática. Pero lo importante no es la existencia o no de la informática, sino su creciente utilidad, versatilidad, facilidad de uso y creciente capacidad de proceso y, todo ello, con equipos mucho más económicos. Como ejemplo de uno de los primeros diseños muestrales informatizados puede citarse el realizado por el INE en el año 1976 relativo a la Encuesta Permanente de Consumo, en el que se trabajó con fichas perforadas que ocupaban casi la superficie de un amplio despacho y en el que el tiempo de respuesta fue elevado dado el estado del arte de la tecnología informática en aquella época. Los diseños muestrales tenían que basarse excesivamente en la intuición profesional por cuanto era prohibitivo pedir al ordenador central el estudio de diversas alternativas para la elección del diseño muestral óptimo. Hoy, la gran facilidad de manejo de los ordenadores, su capacidad de proceso y en particular la aparición de potentes microordenadores, permiten hacer multitud de estudios en torno a la elección óptima del diseño muestral de un encuesta. Como ejemplo de un diseño muestral actualizado tenemos el ya citado diseño muestral de la Encuesta General de Población (EGP), a partir del cual se realizan las encuestas oficiales del INE sobre población y hogares, como la EPA (Encuestas de Población Activa), la EPF (Encuestas de Presupuestos Familiares), y en general las encuestas del INE dirigidas a viviendas familiares o a personas que las habitan.

Trabajo de campo

Se consideran trabajos de campo aquellos que consisten en la obtención de las medidas de las variables objeto de estudio, asociadas a las unidades de la población sobre las que se realiza la medición. Para introducir, de forma somera, la complejidad que pueda suponer la realización de los trabajos de campo, vamos a analizar sintéticamente los elementos que participan en dichos trabajos de campo. Los elementos que consideramos en la realización de los trabajos de campo son los siguientes:

- Las unidades a medir
- Las variables objeto de medida
- El instrumento de medida
- La realización de la medida y la instrumentalización necesaria

- ***Las unidades a medir***

Cuando se realizan encuestas es necesario tratar de aplicar con el adecuado rigor las líneas de actuación que presiden la teoría del muestreo. Ello supone determinar de manera previa, a priori, la unidad informante sin que quepa ninguna arbitrariedad o indeterminación. Esto no sucede así, en general, en la mayoría de las encuestas de opinión y sociológicas, dirigidas a personas y hogares, con las que está familiarizada la mayoría de la gente y que, además, se toman como referencia de los trabajos estadísticos por muestreo. En ellas la unidad informante se selecciona, en la mayoría de los casos, siguiendo un criterio opinático dentro de unas determinadas restricciones generales (edad, sexo, zona de residencia...) y, por tanto, la elección no se hace estrictamente a priori, lo que supone una facilidad mucho mayor de localización y elección de las unidades informantes dispuestas a colaborar.

Sin embargo, una encuesta seria supone localizar inequívocamente las unidades informantes (hogares, empresas, instrucciones, etc.), dando lugar a dificultades añadidas. Enumeramos por ejemplo las siguientes:

- Visitas reiteradas a los hogares, las que sean necesarias, ante casos de ausencia en el momento en el que se verifica la visita para la entrevista.
- Si la ausencia del hogar es prolongada, la sustitución, siguiendo una norma rigurosa, por otro hogar, identificado también a priori.
- La localización de los hogares en lugares de población dispersa.
- La búsqueda de establecimientos o empresas cuya ubicación no es fácilmente asequible (polígonos industriales, diseminados, búsqueda incluso, en otra provincia, porque el titular informante reside en distinto lugar de aquel en el que se encuentra la empresa...).
- Establecer contacto con el informante idóneo en una gran empresa (un jefe de producción no es lo mismo que un jefe administrativo...).

Nos ocuparemos ahora de las variables objeto de medida.

- ***Las variables objeto de medida***

El rigor estadístico hay que trasladarlo también a la definición de las variables objeto de estudio para que la toma del dato esté correctamente acotada sin la menor indeterminación. Como ejemplo sencillo, supongamos que deseamos medir en un hogar la variable cualitativa tener o no teléfono.

Pudiera pensarse que esta variable no necesita ninguna explicación complementaria y, sin embargo, dentro del rigor al que aludimos señalamos como posibles alternativas la situación de propiedad (el teléfono es propiedad del hogar que lo utiliza y está dentro de la vivienda), situación de disponibilidad dentro de la vivienda (el teléfono está disponible en el hogar y está dentro de la vivienda, y es compartido por dos o más hogares que conviven en la misma vivienda aunque no sea de su propiedad), situación de disponibilidad fuera de la vivienda (por ejemplo, en la tienda situada en la planta baja aunque la vivienda esté en la primera planta).

Naturalmente lo anterior es un ejemplo sencillo de definición de variables en lo que se refiere al necesario grado de especificación. Podemos fácilmente imaginar la complejidad existente en definiciones asociadas a variables de tipo económico que, por otra parte, exigen gran especialización para definirlas. Como una muestra cualquiera reseñamos la descripción de lo que se entiende por prendas de uso masculino y de uso femenino. Prendas de uso masculino son las que, teniendo una abertura delante, se cierran superponiendo el lado izquierdo sobre el derecho. Cuando dicha abertura se cierra o se superpone el lado derecho sobre el izquierdo son de uso femenino. Si la prenda carece de abertura por delante pero el corte indica manifiestamente que ha sido diseñada para uno u otro sexo, se clasificarán en el uso para el que fue diseñada. Las prendas no identificables como prendas de uso masculino o femenino se clasifican en femeninas.

Son evidentes las horas de trabajo, reflexión y discusión que hay que utilizar para determinar cientos y cientos de variables como las apuntadas, y no es un lujo excesivo del trabajo estadístico el proceder con el grado de rigor y meticulosidad que se desprende de las mismas. Lo que sucede es que la realidad ofrece para su estudio un inmensa riqueza de matices distintos y todos tienen que ser recogidos por las variables elegidas para representar tal realidad. De aquí que las definiciones de las variables han de estar muy bien delimitadas porque, en caso contrario, al no diferenciarlas y acotarlas debidamente, correremos el riesgo de agrupar datos heterogéneos.

- ***El instrumento de medida***

El instrumento de medida es el elemento que se utiliza en las investigaciones por muestreo para recoger el valor de las variables investigadas asociadas a la unidad muestral sujeta a medición. El instrumento de medida habitual en las encuestas es el cuestionario, que contiene las variables cuyo valor han de cumplimentar las unidades muestrales informantes, normalmente personas, hogares, empresas o instituciones.

El cuestionario es el medio de comunicación entre el encuestador y la unidad informante. Es además el instrumento de trabajo para la posterior codificación de la información. Ha de estar, por tanto, estructurado convenientemente en secciones y preguntas para que sea fácilmente manejable y codificable informáticamente. Además, es conveniente que el cuestionario mantenga en todo momento el interés del encuestado, siendo el vocabulario utilizado adecuado a su nivel. Por otra parte, el cuestionario ha de diseñarse para que la entrevista no supere la duración de una hora.

A pesar de las indicaciones anteriores, es de destacar que en muchas de las estadísticas oficiales los cuestionarios habituales suelen ser muy extensos, no son de inmediata cumplimentación y exigen, en la mayoría de los casos, costosas y laboriosas elaboraciones añadidas. No son, pues, como los cuestionarios de opinión, donde la respuesta puede ser directa y al momento. Como ejemplo podemos citar la Encuesta Continua de Presupuestos Familiares que realiza trimestralmente el INE. En ella cada hogar que forma parte de la muestra tiene que cumplimentar tres cuestionarios individuales, de tantos miembros como existan en el hogar de catorce años o más, excepto el ama de casa. No obstante, sin esa información primaria tan exhaustiva no existirían las radiografías que constituyen las estadísticas o serían de mala calidad y no podrían tomarse las adecuadas decisiones políticas sociales y económicas que se realizan sobre ellas.

- ***La instrumentalización de la medida***

Evidentemente, la realización de la medida requiere la oportuna instrumentalización asociada a su ejecución. Sucede que los esfuerzos necesarios a realizar para lograr la correcta medida de las variables de estudio, asociadas a las unidades informantes, lógicamente se multiplican geométricamente en función de las dificultades ya apuntadas al hablar de la determinación y localización a priori y sobre el terreno de las unidades informantes, de la dificultad para especificar las variables objeto de observación y de los extensos cuestionarios y de su laboriosa cumplimentación. La instrumentalización aludida se materializa en:

- Formación de presupuestos y su realización y control
- Determinación del método idóneo de recogida de los datos (entrevistador, teléfono, servidor vocal, fax, correo, ordenador portátil, métodos mixtos)
- Elaboración de manuales de instrucción, generalmente extensos y detallados, dada la amplia casuística que suele presentar la recogida de datos
- Diseño e impresión del material de trabajo como el cuestionario y el resto de la documentación de control de trabajo de campo
- Diseño de propaganda y su contratación para motivar a los informantes (radio, televisión, prensa especializada o no)
- Diseño de múltiples visitas (para explicación, recordatorio o ayuda) a las unidades informantes para lograr la correcta elaboración
- Preparación de cuadros y tablas referentes a la información a recoger.
- Selección y adiestramientos de agentes y supervisores (ciclos de conferencias, clarificación de la documentación y sobre todo del cuestionario y sus fines, etc.).

Nos ocuparemos ahora de un tema tan importante como es la utilización de la informática dentro de las tareas del trabajo de campo.

- ***La informática en el trabajo de campo***

Respecto a la positiva interacción del desarrollo informático con los trabajos de campo de las encuestas, hay que señalar que esta interacción se concreta en el desarrollo de aplicaciones microinformáticas que favorecen notablemente tanto la gestión de la recogida de los datos como la grabación y depuración de la información. Así, una aplicación microinformática puede desarrollar módulos de gestión que incluyan:

- El control de estado de la colaboración de la unidad informante (cuestionario enviado, recibido, reclamado, proceso de sanción...)
- Asignación de trabajo para cada agente entrevistador
- Obtención de indicadores ligados a la recogida de datos (de unidades recibidas en plazo, fuera de plazo, unidades ausentes, negativas, fuera de ámbito, ilocalizables...)
- Altas, bajas y modificación de unidades para actualizar la base de datos de las unidades muestrales
- Inclusión de variables testigo que permitirán detectar dónde hay que concentrar más los esfuerzos en la última fase de la recogida

También la aplicación microinformática puede contener módulos de grabación y depuración de datos, que permitirán a un agente entrevistador aumentar su eficacia con menús muy asequibles para facilitar su manejo, grabando la información directa-mente según la recibe de la unidad informante y depurando, es decir, corrigiendo posibles datos erróneos detectados, con ayuda del programa informático, en el momento de la interacción agente entrevistador-unidad informante.

Aplicaciones informáticas, en el sentido apuntado, vienen desarrollándose en los últimos años en el Instituto Nacional de Estadística con utilidades cada vez más crecientes y con mejores prestaciones. Como ejemplo del inicio de estas actividades informatizadas tenemos las experiencias piloto de la Encuesta de Población Activa (EPA), en el sentido de sustituir o complementar el clásico cuestionario con un ordenador portátil que con un software adecuado facilite la toma de datos. Esto mismo es aplicable a las encuestas dirigidas a empresas que disponen también de aplicaciones informáticas más o menos sofisticadas para facilitar los trabajos de campo. Actualmente, los agentes entrevistadores de la EPA se equipan con ordenadores portátiles que implementan como aplicación el cuestionario, que es rellenado directamente sobre el ordenador utilizando dispositivos de entrada como el lápiz óptico, que mejoran el clásico e incómodo teclado. De esta forma, los datos del cuestionario se almacenan automáticamente en ficheros de los ordenadores portátiles que luego son descargados sobre ficheros del ordenador central. De esta forma se elimina el costoso trabajo de grabación que en lo referente a la EPA suponía un 20% de la grabación total en el INE.

Es necesario mejorar las aplicaciones informáticas para, de manera creciente, ir facilitando la gestión del encuestador en la recogida del dato y en el control inmediato de este trabajo.

Sería ideal hacer interactiva la ejecución del diseño muestral con los trabajos de campo para alguna o algunas de las variables básicas de la encuesta que se realiza, de modo que según se vaya recogiendo más muestra en campo sepamos cómo va la calidad de la estimación en cada estrato y globalmente. De este modo podríamos reasignar los esfuerzos de recogida allá donde más resentida pueda estar.

Encuesta piloto

Cuando se realizan encuestas de gran dimensión suele ser muy útil seleccionar una pequeña muestra para una prueba piloto. Esta prueba piloto puede ser crucial, ya que permite probar en campo el cuestionario y otros métodos de medición, calificar a los encuestadores y verificar el manejo de las operaciones generales de campo. De la encuesta piloto también se pueden obtener estimaciones de determinadas características poblacionales que pueden utilizarse posteriormente en cálculos sobre tamaños muestrales y estimaciones de los errores de muestreo. Los resultados de la encuesta piloto siempre sugieren modificaciones en la planificación de la encuesta general que van a mejorar la calidad de los resultados de la encuesta a escala completa. Podríamos señalar como características críticas de una encuesta piloto las siguientes:

- Ensaya el cuestionario en condiciones reales.
- Pone a prueba los aspectos fundamentales de la encuesta principal
- Contrasta la idoneidad del marco
- Resalta la variabilidad de determinados caracteres
- Permite intuir la tasa esperada de falta de respuesta
- Comprueba la idoneidad del método de recogida de datos
- Aporta datos sobre el probable coste y duración de la encuesta principal

Procesamiento de los datos

Las grandes encuestas generan gran cantidad de información, por lo que su planificación ha de recoger necesariamente el apartado de procesamiento de los datos. Dicho procesamiento ha de realizarse de modo automatizado utilizando en la mayor medida posible las prestaciones que ofrecen las nuevas tecnologías de la información y la comunicación. Entre las tareas más importantes que abarca este apartado, y que necesariamente se realizarán mediante medios informáticos, tendríamos las siguientes:

- Proceso y depuración automática de cuestionarios
- Imputación de información faltante
- Ajuste de la no respuesta
- Cálculo de estimaciones y sus errores
- Tabulación de los datos
- Análisis de resultados mediante técnicas avanzadas de análisis multivariante implementadas en la diversidad de software estadístico existente actualmente

El procesamiento de la información se optimizaría acercando lo más posible la grabación y depuración de los datos al momento de la obtención del dato mientras se está en campo, pues a posteriori se hace mucho más difícil volver a contactar con la unidad informante. Esto exige el desarrollo de sofisticados programas informáticos, idóneos para cada encuesta y de fácil manejo, para facilitar la grabación y posterior depuración del dato primario por el propio encuestador.

También es muy interesante el desarrollo de aplicaciones informáticas más sofisticadas que permitan la integración de recogida de información a través de fax automático asociado al ordenador (para la recogida de datos por fax), servidor vocal (recogida de datos a través del teléfono, con reconocimiento de voz), y por correo (usando, en lo posible, programas informáticos de reconocimiento de caracteres). Es decir, tratando de trasvasar los datos primarios recogidos de las unidades informantes, lo más directamente posible, a una base de datos, con el objetivo de una más rápida operatividad de control del dato y elaboración última del mismo.

En cuanto a la imputación informatizada podemos decir, en sentido amplio, que se trata de obtener estimaciones que permitan completar las tabulaciones sin dejar huecos, ya que omitir en las tablas los datos faltantes supondría aceptar que la distribución de los datos omitidos coincide con la de los datos presentes. Desde que Fellegi y Holt publicaron en 1996 su trabajo sobre corrección e imputación automatizada, se han venido desarrollando diferentes métodos sobre esta materia, cada vez más sofisticados y precisos, adaptándose a los considerables avances en el campo de las nuevas tecnologías.

Evaluación de resultados

Después de obtener los primeros datos relativos a una encuesta, es necesario proceder a su evaluación con la finalidad de ***contrastar la calidad de la encuesta*** antes de proceder a la presentación y difusión de resultados. Entre los puntos más importantes que se persiguen con la evaluación destacan los siguientes:

- Contrastar las discrepancias entre el diseño teórico y el aplicado
- Evaluar los errores ajenos al muestreo y los debidos al muestreo
- Analizar los costes
- Comparar los resultados con los de otros diseños alternativos
- Contrastar los resultados con los de fuentes externas para una encuesta similar

Presentación de resultados

Una vez obtenidos los resultados de una encuesta, la mera publicación de los mismos no dice nada respecto del trabajo realizado para obtenerlos. Es muy necesaria una presentación ordenada y lo suficientemente documentada de los resultados que permita conocer la calidad de los mismos y medir de alguna forma la confianza a depositar en las estimaciones resultantes.

Según indicaciones de la Conferencia de Estadísticos Europeos, suele ser habitual presentar dos tipos de informes sobre los resultados, el informe técnico y el informe resumido. El ***informe técnico*** puede publicarse de forma irregular y ser puesto al día cuando se estime conveniente. Dicho informe suele ir dirigido a personal especializado y ha de contener como mínimo información sobre las fuentes de los datos, conceptos, definiciones, clasificaciones y metodología. El ***informe resumido*** va enfocado hacia el usuario general y debe presentarse en cada difusión primaria de los datos de una encuesta. Dicho informe ha de contener como mínimo la referencia al informe técnico detallado, información básica sobre la fuente de los datos, definiciones, cobertura de la encuesta, idoneidad del marco, métodos de selección de la muestra y estimación, errores de muestreo, tasas de respuesta y comparación de resultados con los de fuentes externas.

Difusión de resultados

Una vez finalizada una encuesta es necesario trazar un plan de difusión de los resultados de la misma que divulgue lo suficiente la información obtenida. En esta fase hay que tener muy en cuenta los diferentes soportes de difusión de la información que la técnica aporta en el momento actual, y en especial todos aquellos medios novedosos de último momento. Actualmente la difusión de los resultados de una encuesta debe contemplar como mínimo las siguientes características:

- Difusión en soporte papel de modo resumido de resultados referidos a las variables más importantes de la encuesta.
- Difusión en soporte magnético del grueso de la información de la encuesta. En soporte magnético la información no ocupa lugar y los medios actuales de almacenamiento como el CD-ROM permiten difundir gran cantidad de información de forma barata.
- Difusión de la información más importante de la encuesta vía INTERNET
- Publicación de avances previos a los resultados finales
- Difusión a medida de la información, con la finalidad de realizar explotaciones de los microdatos que permitan obtener resultados muy específicos previa petición de usuarios especializados.

CONVENIENCIA Y LIMITACIONES DEL MUESTREO

Ya históricamente existió discrepancia entre los estadísticos defensores de los ***métodos representativos*** (obtención de información poblacional a partir de muestras que representen a toda la población) frente a los ***métodos exhaustivos*** (obtención de la información poblacional solo a partir de censos que analizan exhaustivamente todas las unidades de la población). En el caso de la utilización de los métodos representativos, puesto que la inferencia supone riesgo, es útil indicar en qué casos conviene o no obtener muestras en lugar de censos o investigaciones exhaustivas.

Conveniencia del muestreo

Aunque el objetivo óptimo en muestreo, al igual que en otras muchas disciplinas, consiste en emplear recursos mínimos para obtener determinada información, o bien en conseguir máxima información con recursos prefijados, existen unos criterios generales para el uso de las técnicas de muestreo que pueden resumirse en los siguientes puntos:

- Se empleará muuestreo cuando la población sea tan grande que el censo exceda de las posibilidades del investigador.

- Se tomarán muestras cuando la población sea suficientemente uniforme como para que cualquier muestra dé una buena representación de la misma.

- Se tomarán muestras cuando el proceso de medida o investigación de los caracteres de cada elemento sea destructivo (consumo de un artículo para juzgar su calidad, determinación de una dosis letal, etc.).

- Se utilizará muestreo cuando se observe desagrado de las personas de las que se requiere información con el fin de disminuir el número de elementos de la población que van a ser encuestados.

- Se utilizarán técnicas de muestreo cuando ello suponga una reducción de costes, considerando tanto el coste absoluto como el coste relativo (coste en relación a la cantidad de información obtenida). Este criterio suele conocerse con el nombre de ***criterio de economía***.

- El muestreo es conveniente cuando la acuracidad (ajuste del valor estimado al valor real de la característica en estudio) resulta ser muy buena. Este criterio suele conocerse con el nombre de ***criterio de calidad***.

- El muestreo es conveniente cuando la formación del personal y la intensidad de los controles y supervisión son altos.

- El muestreo será conveniente en general cuando constituya la ***solución de mayor eficiencia en el sentido del coste-beneficio***.

Limitaciones del muestreo

Al igual que existen determinadas situaciones en las que es evidente la ventaja de utilizar muestreo, existen otras en las que el muestreo no es muy conveniente. Podríamos citar las siguientes:

- Cuando se necesite información de cada uno de los elementos poblacionales.
- Cuando sea difícil superar la dificultad que supone el empleo de un instrumento delicado y complejo como la teoría del muestreo.
- El muestreo exige menos trabajo material que una investigación exhaustiva, pero más refinamiento y preparación (base adecuada de los diseñadores y preparación de los entrevistadores, inspectores y supervisores), lo que puede suponer en muchos casos una limitación a su utilización.
- Cuando el coste por unidad, que es mayor en las encuestas que en los censos, aconseje desestimar los métodos de muestreo.

CARACTERÍSTICAS DESEABLES DE UNA INVESTIGACIÓN POR MUESTREO

Hemos visto que el muestreo tiene sus limitaciones y sus ventajas. Sin embargo, es deseable que las investigaciones por muestreo se ajusten lo mejor posible a unas características determinas, consideradas como óptimas, y que podríamos resumir como se indica a continuación:

Acuracidad: proximidad al valor verdadero de las características poblacionales estimadas.

Pertinencia: capacidad de los resultados estadísticos obtenidos con la investigación por muestreo para completar ciertas lagunas en el resultado de un fenómeno.

Oportunidad: utilidad de un resultado estadístico en función de su disponibilidad en el tiempo (puntualidad, rapidez y actualidad). En el caso de censos y grandes encuestas es aconsejable la publicación de avances provisionales de resultados basados en muestras o submuestras.

Accesibilidad: aunque se disponga de un banco de datos informatizado pueden surgir dificultados legales para utilizarlo (protección de la intimidad, secreto estadístico, LORTAD y Ley de la Función Estadística Pública). La información obtenida por muestreo ha de ser totalmente accesible, así como tener presente desde el diseño la legalidad vigente.

Detalle y cobertura: la producción de datos extensos y profundos puede llevar a complementar una investigación exhaustiva con una muestra

Economía: las consideraciones sobre costos en las diferentes etapas de planificación, recogida y procesamiento de datos, evaluación, análisis y publicación pueden mostrar en algunos casos la no conveniencia de una investigación exhaustiva. Luego este criterio ha de tenerse siempre presente a la hora de planificar una investigación por muestreo.

***Integración**:* hay que obtener buena concepción global de la información y buena comparabilidad. La información obtenida en la investigación por muestreo ha de ser integrable y comparable con otras informaciones ya existentes o futuras.

ERRORES EN LAS ENCUESTAS POR MUESTREO

En las encuestas por muestreo puede definirse el «error» de una determinada estimación como la diferencia entre el valor observado $\hat{\theta}$ y el valor desconocido de la característica poblacional θ que tratamos de estimar (error $=|\hat{\theta}-\theta|$). El significado de la palabra «error» no equivale en Estadística, necesariamente, a equivocación, sino más bien a un indicador del margen esperado de incertidumbre. Los errores se deben a causas diversas, pudiendo clasificarse en ***errores de carácter aleatorio*** y ***errores de carácter sistemático o sesgos***. Como ejemplo de los primeros citaremos el originado por la variabilidad de los valores obtenidos en el proceso de muestreo, y entre los segundos el producido por un método tendencioso de medición.

Pueden originarse errores en los resultados de una muestra particular debido a los respondientes, entrevistadores, codificadores, etc., así como a la posible interdependencia entre ellos. Así, por ejemplo, los entrevistados pueden no comprender bien las preguntas, no conocer las respuestas, o quedar influenciados de algún modo por el entrevistador. Los errores de carácter aleatorio y los de carácter sistemático (sesgos) tienen, en general, distintas fuentes, efectos y métodos de medida. La reducción de los errores aleatorios requiere hacer «más de algo» como, por ejemplo, aumentar el tamaño de la muestra, mientras que la reducción de los errores sistemáticos requiere hacer «algo más» como, por ejemplo, una supervisión o un programa de control.

Otra clasificación muy útil es la que distingue entre ***errores de muestreo*** (que son los originados por la variabilidad de los valores obtenidos en el proceso de muestreo y que por lo tanto son de carácter aleatorio) y ***errores ajenos al muestreo*** (que se producen por causas ajenas al muestreo en sí, es decir, por causas no probabilísticas) y que por lo tanto pueden considerarse errores de carácter no aleatorio, sistemáticos o sesgos.

Un primer carácter diferencial entre estos tipos de error es que mientras los errores de muestreo decrecen al aumentar el tamaño de la muestra, los errores ajenos al muestreo suelen crecer con el tamaño de la investigación, o en cualquier caso no suelen decrecer. Un segundo carácter diferencial es que los errores de muestreo se estiman con los datos de la muestra, mientras que los errores ajenos al muestreo suelen requerir para su estimación datos extramuestrales.

Una clasificación inicial de estos tipos de error podría ser la siguiente:

$$\begin{cases} \text{Errores de muestreo (carácter aleatorio)} \begin{cases} E(\hat{\theta}-\theta)^2 = ACURACIDAD \\ E(\hat{\theta}-E(\hat{\theta}))^2 = VARIANZA \\ \sigma(\hat{\theta}) = ERROR\ DE\ MUESTRE\ 0 \end{cases} \\ \text{Errores ajenos al muestreo (carácter sistemático)} \begin{cases} ERRORES\ DE\ COBERTURA \\ ERRORES\ DE\ RESPUESTA \\ FALTA\ DE\ RESPUESTA \end{cases} \end{cases}$$

Errores de cobertura

La población marco ha de cubrir lo mejor posible a la población objetivo. La falta de cobertura de la población objetivo por la población marco produce, en general, una subestimación cuya importancia depende de las características de las unidades omitidas. Si una misma unidad es considerada más de una vez, en la población marco el efecto será una estimación por exceso. La población marco debe constituir una colección actualizada y exhaustiva de las unidades de muestreo, sin solapamientos, con límites bien definidos y fácilmente identificables, sin duplicaciones, sin omisiones y sin unidades extrañas ni vacías. Los errores de cobertura son difíciles de estimar, y requieren investigaciones especiales o la utilización de fuentes externas a la encuesta. En este capítulo se abordarán algunas técnicas sobre tratamiento de marcos imperfectos.

Estos errores pueden estimarse mediante el ***método de reenumeración***, que consiste en volver a enumerar las unidades en una submuestra de pequeñas áreas que figuren en la encuesta principal. Se establece una correspondencia unidad a unidad entre los listados obtenidos en ambas ocasiones con objeto de encontrar unidades omitidas o duplicadas. En la segunda enumeración se deberían utilizar agentes con mejor adiestramiento. Una ventaja de este método es que permite identificar la naturaleza del error de cobertura. Así, por ejemplo, puede encontrarse que la omisión de una persona se debe a la omisión de su vivienda, o a que ha sido omitida dentro de la vivienda, o a un error en el proceso de los datos. Entre los inconvenientes mencionaremos que la reenumeración puede a su vez introducir nuevos errores de cobertura.

También pueden estimarse los errores de cobertura mediante el ***método de las principales componentes demográficas***, que consiste en el conocimiento para toda la población en estudio de valores teóricos relativos a ciertos caracteres como sexo, edad, nacimientos, defunciones y migraciones etc., basados en los datos de censos anteriores, y comparar esos resultados con los obtenidos para nuestra muestra. Este método proporciona un indicador de la inconsistencia entre dos conjuntos de datos, pero sin identificar en qué conjunto se encuentra el error.

Errores de respuesta

Toda encuesta o censo puede considerarse como conceptualmente repetible en condiciones generales análogas. La respuesta dada por la unidad *i* en la realización *t* de una encuesta o censo es una variable aleatoria cuyos valores en distintas realizaciones no están correlacionados.

Las condiciones generales de una encuesta o censo incluyen los conceptos y definiciones, el cuestionario, el método de recogida de datos, la selección, adiestramiento y control de los agentes entrevistadores, la supervisión del trabajo de campo, el procesamiento de la información y, en el caso de encuestas, la estrategia muestral, etc. Para controlar todos estos aspectos es necesario tener presente que la varianza total de un estimador consta de las siguientes componentes aditivas: varianza total de respuesta, varianza de muestreo y la covarianza entre desviaciones de respuesta y desviaciones de muestreo. A su vez, la ***varianza total de respuesta*** recoge el efecto conjunto de las siguientes componentes: la ***varianza simple de respuesta***, que mide la variabilidad de las respuestas dadas en sucesivas realizaciones conceptualmente posibles dividida por el tamaño de la muestra y la ***componente correlacionada***, que recoge el efecto añadido de una posible influencia de los entrevistadores, codificadores, etc. sobre las respuestas. Esta segunda componente no se reduce al aumentar el tamaño de la muestra.

A la diferencia entre el valor esperado del estimador, sobre todas las realizaciones conceptualmente posibles y sobre todas las unidades de la población, y el valor «objetivo» se le denomina ***sesgo de respuesta***. Se suele llamar error total a la varianza total más el cuadrado del sesgo.

Para estimar el error de respuesta puede utilizarse el ***método de la reentrevista***, que consiste en la realización de nuevas entrevistas a una submuestra de entrevistados en la encuesta principal. La reentrevista a una unidad debería hacerse bajo las mismas condiciones generales y dentro de un lapso de tiempo no demasiado largo, con objeto de evitar el olvido de los datos correspondientes a la fecha de referencia, ni demasiado corto, con el fin de limitar en lo posible el efecto del factor memoria. Así, por ejemplo, una parte de la posible discrepancia en los resultados de las entrevistas a una misma unidad realizadas por agentes con un grado similar de adiestramiento, puede ser debida a diferencias entre los entrevistadores. Si estas diferencias son de carácter aleatorio y corresponden a errores de respuesta no correlacionados, dan lugar a la ***varianza simple de respuesta***, que puede ser estimada por el método de reentrevista. Para estimar el error de respuesta puede utilizarse también el ***método de las submuestras interpenetrantes.***

Falta de respuesta

En una encuesta puede no disponerse de información para todas o algunas de las preguntas que figuran en el cuestionario correspondiente a una unidad de la muestra. En el primer caso diremos que la falta de respuesta, para la unidad de la muestra, es total, y en el segundo caso que es parcial. La falta de respuesta puede ser debida a diversas causas, como por ejemplo:

a) Imposibilidad de identificar la unidad sobre el terreno o de acceder a la misma.
b) Incapacidad para contestar por parte del entrevistado.
c) Ausencia temporal del entrevistado.
d) Negativa a cooperar en la encuesta por parte del entrevistado.
e) Pérdida de información.

Método de Hansen y Hurwitz
Método de Politz y Simmons
Modelo de Deming
Modelos de respuesta aleatorizada. Modelo de Warner

MÉTODOS DE AJUSTE DE FALTA DE RESPUESTA Y EQUILIBRADO DE MUESTRAS. PONDERACIONES

Con el propósito de disminuir el posible sesgo introducido por la falta de respuesta se suelen utilizar varios tipos de ajuste. Entre ellos, mencionaremos:

a) Ajuste sobre el terreno, es decir, en la fase de la recogida de datos

El entrevistador recibe el listado de las unidades de la muestra que ha de visitar, por ejemplo, viviendas, y una lista de "suplentes" elegidos con el mismo mecanismo aleatorio que las anteriores. Si después de agotar el número establecido de visitas a una vivienda no consigue realizar la entrevista, por ausencia o negativa, la sustituye por la primera vivienda de la lista de suplentes. las sustituciones realizadas han de tenerse en cuenta a la hora de calcular la tasa de respuesta. Por supuesto que los datos obtenidos, con las sustituciones, siguen perteneciendo al estrato de los que contestan y por lo tanto no añaden información alguna sobre el estrato de los que no responden ni reducen el posible sesgo.

b) Ajuste utilizando ponderaciones

Si se dispone de información suplementaria sobre la proporción de unidades para ciertas clases de la población, por ejemplo las unidades urbanas, y debido a la falta de respuesta esa clase está representada por defecto, se puede estratificar a posteriori la muestra y utilizar las mencionadas proporciones como ponderaciones.

Si no existe información suplementaria se pueden ponderar los datos de la muestra para clase o subclase utilizando como pesos las inversas de las tasas de respuesta.

c) Ajuste para la falta de respuesta parcial

Cuando un cuestionario no está completo, se puede realizar una imputación para los datos que faltan basada en la posible correlación entre el dato omitido y el resto de los datos disponibles. El procedimiento más utilizado es el fichero caliente (*hot deck*), que en esencia consiste en las siguientes fases:

1. Se establecen una serie de caracteres que se suponen correlacionados con el que pretendemos imputar. Por ejemplo, sexo, educación y grupos de edad en relación con la situación laboral.

2. Se introducen en el ordenador unos valores iniciales (fichero frío, *cold deck),* obtenidos de encuestas anteriores.

3. Si en la primera ficha de la encuesta falta el dato, se imputa el correspondiente al "fichero caliente". Si por el contrario la ficha está completa se actualiza el correspondiente dato en el "fichero frío", y así sucesivamente.

Se ha demostrado que el procedimiento "fichero caliente" produce un incremento en la varianza del estimador que no se refleja en los métodos de estimación disponibles hasta ahora.

NOTAS HISTÓRICAS

Las técnicas de muestreo estadístico en poblaciones finitas son bastante recientes. Dichas técnicas vinieron originadas por necesidades prácticas relativas a censos, recuentos, juegos de azar y en general problemas de inferencia inductiva basados en datos empíricos. Como anécdota puede citarse que una de las primeras aplicaciones de la selección aleatoria la constituyó el diezmado de unidades militares como castigo.

A continuación se expone la evolución histórica de las técnicas de muestreo estadístico, distinguiendo entre trabajos preliminares, primeros trabajos específicos, consolidación de los textos generales sobre muestreo y evolución del muestreo en las décadas de los años setenta y ochenta. Esta distinción viene marcada por la propia evolución cronológica de las técnicas sobre muestreo estadístico en poblaciones finitas. Dentro de los primeros trabajos específicos se realiza una agrupación según los diferentes tipos de muestreo.

Trabajos preliminares

Entre los ***trabajos que podríamos considerar preliminares*** a la teoría del muestreo merecen destacar los siguientes:

1895- Kiaer (Director de la Oficina Central de Estadística en Noruega). En su publicación *Observations et experiences concernant les denombrements representatifs* hace una defensa de los ***métodos representativos*** (obtención de información poblacional a partir de muestras que representen a toda la población) frente a los ***métodos exhaustivos*** (obtención de la información poblacional sólo a partir de censos que analizan exhaustivamente todas las unidades de la población).

En esta época los métodos exhaustivos fueron defendidos por el alemán ***Von Mayr*** y otros estadísticos oficiales temerosos de que las muestras pudieran llegar a sustituir a los censos. También en esta época los métodos representativos fueron defendidos en Estados Unidos por ***C.D. Wright***, fundador del *Bureau of Labor Statistics en Massachussetts*, en Inglaterra por ***Arthur Bowley*** y en Rusia por ***A. Kaufmann*** y ***A. Chuprov***.

1906 - Bowley. Aplica la teoría de la inferencia a encuestas por muestreo, y en concreto aplicó el teorema central del límite para evaluar la precisión de las estimaciones obtenidas con grandes muestras aleatorias de grandes poblaciones finitas.

1923 - El ruso ***A. A. Chuprov*** escribe un artículo con fórmulas sobre teoría del muestreo de poblaciones finitas sin reposición.

1924 - El también ruso ***A. J. Kowalsky***, en su libro *Basic Theory of sampling Methods*, escribe ampliamente sobre teoría del muestreo de poblaciones finitas sin reposición.

1924 - En el marco de las reuniones del ***Instituto Internacional de Estadística (ISI)*** se nombra una comisión para el estudio de los métodos de muestreo formada por ***Jensen, Bowley, Gini*** y otros.

1927 - ***Tippet*** publica la primera ***tabla de números aleatorios*** para la obtención de muestras probabilísticas

Primeros trabajos específicos sobre muestreo

En cuanto a los ***primeros trabajos ya específicos sobre muestreo*** agrupados según los diferentes tipos podríamos considerar los siguientes:

- ***Muestreo estratificado***

1934 - Jerzy Neyman publica en la *Royal Statistical Society* de Londres el primer trabajo considerado como científico sobre muestreo en poblaciones finitas cuyo título es *On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection*. Neyman estableció que la selección

aleatoria era la base de una teoría científica que permitía predecir la validez de las estimaciones muestrales. Neyman se planteó medir el grado de incertidumbre y regularlo al actuar con observaciones afectadas de una cierta variabilidad. Dicho grado se midió por ***intervalos de confianza*** y se reguló por ***criterios de eficiencia como el de minimización de la varianza para tamaño de muestra fijo***. Neyman fue el primero en presentar conceptos básicos de muestreo en poblaciones finitas, proporcionando base científica sobre selección de unidades de muestreo, métodos de estimación, uso de información complementaria para estratificar y afijación óptima.
1935 - Yates y Zacopanay amplían el criterio de afijación de mínima varianza de Neyman a la afijación de mínima varianza para un coste fijo (afijación óptima).

1938 - Neyman considera el ***muestreo doble o bifásico*** para el caso en que no se conozcan los tamaños de los estratos pero puedan estimarse mediante una muestra aleatoria simple preliminar extensa y barata. La característica en estudio se estimará utilizando una submuestra de la muestra extensa.

1941 - Stephan considera la ***estratificación con caracteres económicos*** y ***King*** y ***McCarthy*** consideran la ***estratificación con caracteres agrarios***.

1942 - Jessen considera la ***estratificación geográfica*** basada en que las unidades adyacentes son en general más parecidas que las unidades lejanas.

1943 - Hansen y Hurwitz realizan trabajos sobre estratificación considerando la ***necesidad de homogeneidad dentro de los estratos y la heterogeneidad entre ellos***, hasta llegar al extremo de considerar una población tan homogénea que constituya un solo estrato.

Existen trabajos importantes sobre muestreo estratificado posteriores en el tiempo como ***Thionet*** en ***1953***, que aborda el problema del uso para la estratificación de una variable correlacionada con la variable a estimar, el de ***T. Dalenius*** en ***1950*** con título *The problem of optimum stratification*, el de ***H. Ayoma*** en ***1954*** con título *A study of the stratified random sampling*, el de ***D. Raj*** en ***1957*** con título *On estimations parametric functions in stratified sampling designs*, el de ***G. Ekman*** en ***1959*** con título *An approximation useful in univariate stratification* y el de ***T. Dalenius*** y ***J. L. Hodges*** en ***1959*** con título *Minimum variance stratification*.

- ***Muestreo por conglomerados***

Se trata de una técnica muestral que fue estudiada a partir de los años cuarenta, basada en el precedente de un trabajo de ***F. Smith*** en ***1938,*** de título *An empirical law describing heterogeneity in the yields of agricultural crops.*

1942- Hansen y Hurwitz utilizan por primera vez la palabra conglomerado para designar un grupo de elementos que constituye una unidad de muestreo. Estos

autores introdujeron el muestreo con reposición y probabilidades desiguales y el concepto de ***coeficiente de correlación intraconglomerados***. El muestreo por conglomerados se originó debido a la imposibilidad de disponer en muchos casos de listas de unidades elementales de muestreo.

1942- Jessen sostiene que la media cuadrática entre los elementos dentro de un conglomerado es una función monótona creciente del tamaño del conglomerado. En este año Jessen publicó la obra *Statistical investigation of a sample survey for obtaining farm facts*.

1943 - M. H. Hansen y W. N. Hurwitz tratan el muestreo por conglomerados en la obra *On the theory of sampling from finite populations*.

1949 - M. H. Hansen y W. N. Hurwitz tratan el muestreo por conglomerados en la obra *On the determination of the optimun probabilities in sampling*.

1950 - Ante problemas de costo derivados de las visitas a todas las unidades elementales de los conglomerados elegidos para la muestra se considera el ***muestreo con submuestreo***, que aparece tratado en la monografía *A Chapter in Population Sampling* del ***Bureau of the Census*** de Estados Unidos. ***P. G. Gray y T. Corlet*** en ***1950*** publicaron la obra *Sampling for the social survey* que trata el muestreo por conglomerados con submuestreo. También ***Yates*** en ***1950*** y ***Sukhatme*** en ***1950*** realizaron estudios relativos a muestreo con submuestreo.

1951- Sukhatme y Pense realizan estudios sobre muestreo polietápico. También ***Sukhatme y Narain*** en ***1952***, ***Sukhatme*** en ***1953*** y ***Thionet*** en ***1953*** realizaron trabajos sobre muestreo polietápico estimando diversas características poblacionales y sus varianzas.

1951 - Narain estudia el muestreo con reemplazamiento y publica la obra *On sampling without replacement with varying probabilities*

1951 - N. Keyfitz considera el muestreo con probabilidades proporcionales a los tamaños de los conglomerados y publica la obra *Sampling with probabilities proportional to size adjustment for changes in the probabilities*.

1951 - H. Midzumo considera el muestreo con probabilidades proporcionales a los tamaños de los conglomerados y publica la obra On the s*ampling system with probabilities proportional to sum of sizes.*

1952 - D. G Horvitz y D. J. Thompson estudian el muestreo de conglomerados sin reemplazamiento en la obra *A generalization of sampling without replacement from a finite universe*.

1953 - F. Yates y P. M. Grundy estudian el muestreo sin reemplazamiento en la obra *Selection without replacement from within strata with probability proportionale to size*.

1953 - W. N. Hurwitz y W. G. Madow publican el libro *Sample survey methos and theory* que contempla el muestreo por conglomerados.

1953 - J. Durbin estudia la selección con probabilidades desiguales mediante la publicación *Some results in sampling theory when the units are selected with unequal probabilities*.

1954 - D. Raj escribe sobre el muestreo con probabilidades proporcionales a los tamaños en la obra *On sampling with probabilities proportionate to size*.

1956 - D. Raj trata el muestreo de conglomerados sin reemplazamiento en la publicación *Some estimators in sampling with varying probabilities without replacement*.

1958 - D. Raj presenta un método de estimación de la varianza en muestreo con probabilidades proporcionales a los tamaños en la obra *On the estimate of variance in sampling with probabilities proportionate to size*.

1962 - Hartley y Rao escriben sobre muestreo con probabilidades desiguales y sin reemplazamiento en el artículo *Sampling with unequal probabilities and without replacement.*

- ***Muestreo sistemático***

1942 - J. G. Osborne considera la existencia de correlaciones internas en el muestreo sistemático (***coeficiente de correlación intramuestral***) y publica la obra *Sampling errors of sistematic and random surveys of cover-type areas.*

1944 - W. G y L. H. Madow en la obra *On the theory of systematic sampling* investigan formalmente por primera vez el muestreo sistemático, que pasó a usarse intensivamente a partir de 1944 debido a estudios sucesivos de estos autores en 1944, 1946 y 1949.

1946 - W. G. Cochran realiza estudios sobre muestreo sistemático en su obra *Relative accuracy of systematic and estratified random samples for a certain class of populations*. También ***Yates*** tiene un artículo al respecto publicado en el año 1946 de

título *A review of recent statistical developements in sampling and sampling surveys*. Este mismo autor en 1949

1949 - Yates contempla desarrollos sobre muestreo sistemático en la obra *Sampling methods for census and surveys*.

1950 - Das compara el muestreo sistemático con el estratificado.

1963 - Brewer, K.R.W. contempla el muestreo sistemático con probabilidades desiguales en la obra *A model of Systematic Sampling with Unequal Probabilities* (Austral. Jour. Statist).

- ***Estimaciones de razón, regresión y diferencia (estimación indirecta)***

1950 - H. Midzumo realiza estudios sobre métodos de estimación indirecta, tratando las estimaciones por razón y regresión, y publica la obra *An outline of theory of sampling systems*.

1951 - D.B. Lahiri publica la obra *A method of sample selection providing unbiased ratio estimates*, que trata la estimación insesgada de la razón.
1953 - Hansen, Hurwitz y Madow proponen el método de estimación indirecta por diferencia y publican la obra *Sample survey methods and theory*.

1954 - Hartley y Ross obtienen un método de estimación insesgada de la razón y publican la obra *Unbiased ratio estimator*. También en 1954 **D. Raj** publica un trabajo sobre estimación de la razón titulado *Ratio estimation in sampling with equal and unequal probabilities.*

1958 - Olkin publica la obra *Multivariate ratio estimation for finite populations.*

- ***Trabajos sobre errores ajenos al muestreo***

1938 - Mahalanobis indica la necesidad de evaluar, adicionalmente a los errores muestrales, los errores ajenos al muestreo (desviaciones de aleatoriedad introducidas por el personal de campo al no identificar bien las unidades muestrales, errores e imprecisiones en los cuestionarios, falta de respuesta por ausencias y negativas a contestar, etc.). En ***1946*** el propio Mahalanobis diseña la técnica de las submuestras interpenetrantes para tratar la falta de respuesta.

1940 - Sthepan y Hansen escriben artículos sobre la falta de respuesta.

1944 - Deming diseña el método que lleva su nombre para el tratamiento de la falta de respuesta. y publica el artículo *On errors in surveys*. Dicho método fue perfeccionado por el propio Deming en 1953, fecha en que publica el artículo *On a*

probability mechanism to attain an economic balance between the resultant error of response and the bias of non-response (JASA).

1946 - Hansen y Hurwitz diseñan el método que lleva su nombre para el tratamiento de la falta de respuesta, recogido en la publicación *The problem of non-response in sample surveys* (JASA).

1949 - Politz y Simmons diseñan el método que lleva su nombre para el tratamiento de la falta de respuesta. Estos autores publican la obra *An attempt to get the not at homes into the sample without callbacks* (JASA)

1954 - Durbin analiza el coste de las visitas repetidas y sus consecuencias. Publica el artículo *Non response and callbacks in surveys* (Boletín del Instituto internacional de estadística).

1954 - Durbin y Stuart publican el artículo *Callbacks and clustering in sample Surveys: An experimental study* (JRSS).

1954 - Simmons publica el artículo sobre la falta de respuesta titulado *A plan to account for not at homes by combining weighting and callbacks* (Journal of Marketing).

1955 - Dalenius propone un método para obtener información de las unidades que no han respondido antes de finalizar la encuesta.

1959 - Kish y Hess publican la obra sobre el sesgo de respuesta de título *A replacement procedure for reducing the bias of non response* (The American Statistician).

1961 - Hansen, Hurwitz y Bershad diseñan el método que lleva su nombre para el tratamiento de la falta de respuesta mediante la publicación *Measurements errors in censuses and surveys* (Bull. Int. Stat. Inst.- Boletín del Instituto Internacional de Estadística).

1965 - Warner publica el artículo sobre respuesta aleatorizada y sesgo de respuesta titulado *Randomized response: A survey technique for eliminating evasive answer bias* (JASA).

1967 - Horvitz, Shah y Simmons publican el artículo sobre el modelo de respuesta aleatorizada titulado *The unrelated question randomized response model* (American Statistician Association).

Consolidación de los textos generales sobre muestreo

En cuanto a la ***consolidación de los textos generales sobre muestreo***, que empieza a producirse a finales de los años cuarenta y que cobra su mayor fuerza en las décadas de los años cincuenta y sesenta, podríamos destacar los trabajos siguientes:

1949 - F. Yates publica el texto *Sampling methods for censuses and surveys*, cuya cuarta edición apareció en 1981 (Griffin).

1950 - W. E. Deming publica el texto *Some theory of sampling* (Wiley).

1950 - P. G. Gray *y* ***T. Corlett*** publican la obra *Sampling for the social survey* (Journal of the Royal Statistics Society A, en abreviatura J.R.S.S. A).

1953 - W. G. Cochran publica el texto *Sampling Tecniques*, que fue mejorado en su edición del año 1977 (Wiley).

1953 - M . H. Hansen, W. N. Hurwitz y W. G. Madow publican el texto *Sample survey methos and theory* (Wiley).

1954 - P.V. Sukhatme publica el texto *Sampling theory of surveys with applications* (FAO, Roma. Traducido al castellano por Fondo de Cultura Económica).
1955 - V.P. Godambe publica la obra *A unified theory of sampling from finite populations* (J.R.S.S. B).

1962 - M.R. Sampford publica el texto *An introduction to sampling theory, with aplicattions to agriculture* (Oliver and Boid).

1963 - M. N. Murthy publica la obra *Some recent advances in sampling theory* (Journal of the American Statistical Association, en abreviatura JASA).

1965 - L. Kish publica la obra *Survey sampling* (JASA).

1967 - M. N. Murthy publica la obra *Sampling theory and methods* (JASA).

1967 - J. Durbin publica el libro *Design of multistage surveys for the stimation of sampling errors* (Aplied Statis.).

1968 - D. Raj publica el libro *Sampling theory* (McGraw Hill).

1968 - R.M. Royal publica el texto *An old approach to finite population sampling theory* (JASA).

Trabajos sobre muestreo en las décadas de los setenta, ochenta noventa y tendencias actuales

En cuanto a los ***trabajos sobre muestreo en las tres últimas décadas*** podríamos considerar los siguientes aspectos:

- Reuniones de la ***Asociación Internacional de Estadísticos de Encuestas***. Se presentan desarrollos teóricos y prácticos de muestreo en poblaciones finitas.

- Trabajos sobre modelos de error total en encuestas y censos, cálculo de errores de muestreo y ajenos al muestreo.

- ***Diseño total de encuestas (D.T.E.).*** Se trata de la fijación de un estándar de encuestas con el que se persigue distribuir los recursos ejerciendo un control sobre las distintas componentes del error para minimizar el error total. Destacan las contribuciones de Nathan en 1972, Lessler en 1974, Fellegi en 1974, Nisselson y Bailar en 1976 y Bailar en 1976 y 1979.

- ***Sistema de información para el diseño por muestreo (SIDEM)*** impulsado por Horvitz en 1978 y que busca mejorar la calidad de las encuestas por muestreo. Los usuarios del sistema tenían acceso al mismo para diseñar sus encuestas y a la vez contribuían a su enriquecimiento. Este sistema propuso una estandarización de términos y definiciones, facilitó la aplicación del concepto de diseño total y proporcionó estándares para la comparación de errores

- En la década de los años ochenta se han intensificado los trabajos sobre ***control de calidad y muestreo***, aplicándose fuertemente las técnicas de muestreo en el control de calidad industrial.

- También en esta época se ha puesto énfasis en el desarrollado de aplicaciones del muestreo para su implantación en todo tipo de ***auditorías***.

- Actualmente se siguen aplicando las técnicas de muestreo en campos tan importantes como la biología, economía, agricultura, comercio, transporte de mercancías, procesos de simulación y técnicas de investigación de mercados (marketing). En cuanto a las estadísticas oficiales, tanto en las encuestas sobre población y hogares como en las encuestas sobre empresas e instituciones, se utilizan los diseños muestrales adecuados, ya más refinados y mejor documentados que en épocas anteriores. Asimismo, actualmente se dispone de marcos más depurados cuya obtención no ha sido tarea fácil, pero que facilitan sobremanera el diseño muestral y reducen los errores.

- Asimismo es necesario destacar que la mayor transformación se ha producido últimamente en la aplicación de las nuevas tecnologías de la información y la comunicación a las diferentes fases de la elaboración de encuestas, tal y como se ha indicado ya al analizar las etapas de una investigación por muestreo. La idea general es extender la informatización al mayor número posible de etapas de una encuesta, especialmente en la entrada de datos, codificación de respuestas, verificación, control e imputación, cálculo de estimadores y sus errores y análisis de resultados.

- El análisis de datos y la gestión de bases de datos, que solían considerarse pertenecientes respectivamente a dos campos distintos, el de los estadísticos y el de los informáticos, constituyen en la actualidad un área de interés común para unos y para otros.

- El uso de las técnicas informáticas ha permitido resolver la mayoría de los problemas relativos a la clasificación, almacenamiento y recuperación de microdatos y de macrodatos, posibilitando también la integración de la metainformación como puente entre la información almacenada y el usuario.

- ***HISTORIA***
  - ***Aleatorio simple***
    - ***Kiaier, Wright.. (representativistas)***
    - ***Bowley → TCL en grandes muestras***
    - ***Chuprov → Fórmulas en m.a.s.s.r.***
    - ***Jensen, Bowley, Gini → ISI***
    - ***Tippet → Tabla de números aleatorios***
  - ***Estratificado***
    - ***Neyman → Varianza mínima para tamaño fijo, bifásico***
    - ***Yates y Zacopanay → varianza mínima y coste fijo***
    - ***Stephan, King, Jesen → economía, agricultura, geograf***
    - ***Hansen y Hurwitz → Homeg. dentro y heter. entre***
  - ***Conglomerados***
    - ***Hansen y Hurwitz → Reposición y prob. desiguales***
    - ***Horvitz y Thompson → Sin repos. y prob. desiguales***
    - ***Yates y Grundy → Varianza mejorada (sin repos.)***
    - ***Hurwitz y Madow → Conglomerados sin y con rep.***
    - ***Sukhatme, Yates, Gray, Corlet, B.C. → Submuestreo***
    - ***Sukhatme, Pense, Narain, Thionet → Polioetápico***
    - ***Midzumo, Durbin, D. Raj, Lahiri, Brewer, Ikeda → ppt***
  - ***Sistemático***
    - ***Madow, Cochran, Yates → Muestreo sistemático formal***
    - ***Das → Muestreo sistemático y estratificado***
    - ***Brewer → Sistemático y prob. desiguales***
  - ***M. Indirectos***
    - ***Midzumo → Estimaciones por razón y regresión***
    - ***Hansen, Hurwitz y Madow → Estim. por diferencia***
    - ***Hartley, Ross y Lahiri → Estim. insesgada de razón***
  - ***Errores ajenos M***
    - ***Mahalanobis → Submuestras interp. falta respuesta***
    - ***Falta de respuesta***
      - ***Deming, Hansen y Hurwitz,***
      - ***Politz y Simmons, Kish y Hess .***
      - ***Hansen, Hurwitz y Bershad***
    - ***Sthepan y Hansen, Durbin, Dalenius, → Trabajos***
  - ***Textos***
    - ***Yates, Deming, Gray y Corlet, Cochran, H. H. y Madow, Sukhatme,***
    - ***Godambe, Murthy, Kish, Durbin, D. Raj, Lethonen y Pahkinene***
  - ***Actualidad***
    - ***Asociación Internacional de Estadísticos de encuestas (AIEE),***
    - ***Diseño Total de Encuestas (DTE), Sistema de Información***
    - ***Diseño por Muestreo (SIDM), Control de calidad, Auditoría,***
    - ***Biología, Economía, Agricultura, Comercio, Marketing, Simul.***
    - ***NTIC → Entrada datos, codificación, imputación, cálculos...***
    - ***Análisis de datos y Gestión de bases de datos (Data Mining)***

LA EVOLUCIÓN DEL MUESTREO EN ESPAÑA

La historia del muestreo en España se desarrolla en torno al Instituto Nacional de Estadística, que viene desarrollando eficientemente su labor de organismo oficial de la estadística española.

Primeros trabajos de muestreo en España

La primera aplicación en España de la teoría del muestreo fue con ocasión de los ***Censos de Edificios, Población y Viviendas de 1950***. Dadas las dificultades que suponía en aquella época procesar, por métodos casi manuales, el cien por cien de los cuestionarios censales, se optó, con encomiable espíritu innovador, por utilizar los métodos que proporcionaban las recién nacidas técnicas de muestreo. Así, ya entonces, se realizó un diseño muestral basado en un muestreo estratificado aleatorio, muestreando un 10 por ciento del total de cuestionarios censales, que se seleccionó en cada estrato aleatoriamente y sin remplazamiento y calculándose estimaciones para las características censales objeto de estudio, proporcionando además los correspondientes errores de muestreo.

Se abrió en España, con este primer trabajo, la aplicación de las técnicas del muestreo. Este trabajo poseía las condiciones ideales para la utilización de estas técnicas, puesto que los datos de infraestructura, materializados en el cien por cien de los cuestionarios censales, eran accesibles sin problemas como directorio de base. Este modo de proceder se siguió en los sucesivos censos decenales de Edificios, Población y Viviendas proporcionando en primer lugar una pequeña muestra avance del 1 por ciento o el 2 por ciento y después, en la mayoría de las veces, una explotación más amplia con una muestra en torno al 25 por ciento, e incluso encuestas para evaluar la calidad de los datos recogidos.

Sin embargo, la primera aplicación de las nuevas técnicas de muestreo a la realización de una encuesta propiamente dicha tiene lugar en el INE al realizarse la ***Encuesta sobre Cuentas Familiares*** 1958. Esta encuesta, pionera de las investigaciones por muestreo en el INE y en nuestro país, dio además unos resultados valorados como muy positivos a tenor de los escasos medios existentes y del carácter innovador que suponía la implantación de técnicas de trabajo de reciente aparición. Su diseño muestral, apoyado en los datos de infraestructura (el directorio que coyunturalmente le proporcionó el Censo Electoral de 1955 y su actualización a 31 de diciembre de 1957) consistió, ya con cierta sofisticación, en un muestreo en dos etapas en el que en la primera etapa se seleccionaron municipios y en la segunda familias, con estratificación de las unidades de primera etapa y con un tamaño muestral de 4.192 familias.

Esta primera Encuesta sobre Cuentas Familiares 1958, que medía los gastos de las familias, sirvió presumiblemente con cierta valentía a los funcionarios de entonces en el INE para encarrilar y ganar confianza respecto de la utilización de las nuevas técnicas de trabajo que proporcionaba la teoría del muestreo. En ella no ha habido problemas de infraestructura estadística aunque, de seguro, muchos de los cálculos necesarios se habrán tenido que realizar a mano y con máquinas clasificadoras de datos.

A medida que avanzan los años y la infraestructura estadística aumenta con el consiguiente aumento de las disponibilidades de marcos de los que se pueden seleccionar las muestras, se llega a una diferenciación clara entre las problemáticas de infraestructura estadística existentes en el campo de las encuestas de población y hogares y posteriormente en el campo de las encuestas de empresas e instituciones.

Infraestructura en Encuestas de Población y Hogares

El gran salto en la producción estadística por muestreo, en lo que a encuestas dirigidas a hogares se refiere, se produce con ocasión de generar la infraestructura estadística que supuso en 1963 la división en secciones estadísticas de todo el territorio nacional y que permitió de inmediato realizar dos grandes encuestas: la Encuesta de Población Activa, iniciada en 1964 y sin interrupción hasta la actualidad, y la Encuesta de Presupuestos Familiares, cuya primera versión data de 1964 y con posteriores repeticiones en 1967, 1973, 1980 y1990, utilizando en lo sucesivo una infraestructura análoga de secciones estadísticas. Además permitió sentar bases sólidas en las que fundamentar con rigor cualquier encuesta por muestreo dirigida a viviendas familiares y/o a los hogares o personas que las ocupan.

La división administrativa de España comprende la provincia, el municipio y el distrito municipal. A partir de ahí el INE introduce una división más fina, para usos exclusivamente estadísticos, denominada sección estadística; y respaldada por los correspondientes croquis, mapas de localización y callejeros. El resultado final es la división de España en aproximadamente 40.000 secciones estadísticas. Es fácilmente comprensible el enorme trabajo de infraestructura que esta división en secciones estadísticas supuso y su posterior mantenimiento.

Evidentemente, para diseñar encuestas dirigidas a hogares, lo ideal sería tener la lista de todos los hogares españoles con su correcta dirección postal y a ser posible datos del sustentador principal y algunos datos socioeconómicos de cada hogar para que el diseño muestral pueda ser más útil y potente. Además se haría necesario que tal lista estuviese actualizada en todo momento para que fuese operativa a la hora de tener que realizar una encuesta dirigida a los hogares.

Como puede comprenderse esto es prácticamente una ficción, porque si bien cuando se realiza un Censo de Población estos datos están disponibles, sin embargo, al poco tiempo ofrecerían variaciones sensibles, constituyendo un problema que de no resolverse arrastraría la imposibilidad de realizar encuestas fiables dirigidas a los hogares fuera de los momentos censales.

Los arduos trabajos de infraestructura que supuso parcelar España en secciones estadísticas, permitieron solucionar el problema aludido en dos pasos: primero se obtendría un subconjunto mucho menor de secciones estadísticas, una muestra de secciones estadísticas, y después tan sólo en estas secciones estadísticas de la muestra se procedería a su actualización, incluyendo las viviendas de nueva construcción con datos de sus ocupantes y también de otras viviendas que en el momento censal podían estar vacías y posteriormente ocupadas.

Los trabajos de estratificación en secciones estadísticas realizados en 1963 tuvieron su culminación en 1969 con la formación de un diseño muestral maestro, es decir de múltiples usos.

Nos referimos al diseño muestral de secciones estadísticas denominado ***Encuesta General de Población (EGP)*** y al que dedicaremos unos párrafos porque abre, de manera absolutamente rigurosa, la puerta a la posibilidad, como ya hemos indicado, de poder realizar cualquier tipo de encuesta que vaya dirigida a viviendas familiares o a las personas o grupos familiares que las habitan.

El diseño de la EGP constituye una muestra de aproximadamente 3.000 secciones estadísticas a imagen y semejanza de las aproximadamente 40.000 secciones estadísticas en que está dividida España. Para construir esta muestra, imagen de la población de hogares que había de representar, se utilizó la información de los Censos de Población, relativa a las características socioeconómicas que tenían las personas que residían en cada sección censal. Así se pudo construir la muestra-maqueta, de 3.000 secciones estadísticas, representativa de la población integrada por las 40.000 secciones estadísticas.

Para elevar los datos de la muestra a la población se utilizaron los datos de número de habitantes dados según los censos o padrones de población o, en los años intercensales, por las proyecciones de población que proporcionan los modelos de evolución demográfica.

Según lo anterior, se disponía entonces con la EGP de un diseño maestro a partir del cual se hacía posible la realización de cualquier encuesta para obtener datos asociados a la población de viviendas de uso familiar o de los grupos humanos que en ellas residen. Sobre este diseño muestral se han podido construir, desde 1964, todas las Encuestas de Población Activa, las Encuestas de Presupuestos Familiares, Equipamiento y Nivel Cultural de las Familias, y en general las encuestas del INE dirigidas a viviendas familiares o a personas que las habitan.

Comentados los elementos de infraestructura estadística que posibilitaron llevar a cabo encuestas por muestreo dirigidas a hogares se enfoca el mismo tema para encuestas económicas dirigidas a empresas e instituciones.

Infraestructura de las Encuestas de Empresas e Instituciones

En el campo de las encuestas dirigidas a empresas la principal limitación proviene, como en el campo de los hogares, de la existencia de infraestructura estadística. Resumiendo, podemos decir que allá donde existían directorios o registros administrativos en los que apoyarse como marco, las encuestas se pudieron llevar a cabo. Ahora bien, de forma general, hay que señalar determinadas dificultades añadidas:

En primer lugar los registros administrativos no son en general totalmente idóneos para fundamentar en ellos la investigación por muestreo. Ello es debido, en origen, a una incorrecta armonización de usos en lo que a aplicaciones estadísticas se refiere. Además, en muchos casos, adolecen de una incorrecta actualización, lo que invalida la representatividad de las muestras que sobre ellos se seleccionan.

En segundo lugar no existe paralelismo con la utilidad que tienen los Censos de Población y Padrones para el uso de la investigación por muestreo, porque los Censos económicos no tienen la tradición de los Censos de Población y además, lo que es un condicionante básico, no existe el equivalente a las proyecciones demográficas de población que permita rellenar las lagunas de información intercensales.

Con las dificultades anteriores se comprende que las encuestas económicas se fueran ofreciendo más lentamente, con menos garantías de pervivencia continuada y con enormes dificultades técnicas, debido a la no idoneidad, en la mayoría de los casos, de los directorios de base o marcos en el uso estadístico. Aunque se puede apreciar un aprovechamiento, hasta el último resquicio, de las posibilidades que los directorios existentes pudieran ofrecer dentro de sus deficiencias.

Un ejemplo tipo de lo que acabamos de señalar es la actual Encuesta de Salarios en la Industria y los Servicios, con periodicidad trimestral desde 1963, y que proporciona los datos básicos de ganancias por trabajador y hora trabajada, así como las horas trabajadas en promedio por cada trabajador. Esta encuesta empezó con una muestra opinática, por tanto no sujeta a la metodología que preside la teoría del muestreo, de aproximadamente quinientas empresas. En 1963 se mejora ostensible-mente la Encuesta al aplicar un muestreo aleatorio estratificado y se le da rigor científico a las cifras ofrecidas. El directorio se formó con las listas de establecimientos facilitados por el Ministerio de Trabajo obtenidas por las mutualidades laborales. Sin embargo, el diseño de 1963 de la Encuesta de Salarios tuvo que modificarse en 1977, porque durante el período 1963-77 pudieron observarse problemas como el deterioro en el directorio, debido a las altas y bajas de empresas y a los cambios de rama de actividad y estrato de tamaño, las alteraciones en los factores de elevación que se producen como consecuencia del problema anterior y de la no respuesta, y las fuertes oscilaciones en las estimaciones de un trimestre a otro.

El diseño muestral de 1977 de la Encuesta de Salarios se apoya en varios directorios: Directorios del Ministerio de Industria para la minería e industrias manufactureras, Directorios del Ministerio de Obras Públicas en lo que se refiere a transportes de mercancías y viajeros por carretera, y directorios del Ministerio de Trabajo, en el resto de actividades, según listas obtenidas de los datos de las mutualidades laborales.

En 1981, sin embargo, se lleva a cabo una modificación al diseño muestral del año 1977 de la Encuesta de Salarios. Se alude, como motivación de esta motificación, que se han venido observando variaciones excesivas en las estimaciones mensuales de ganancias medias, número de horas trabajadas y número de trabajadores por rama de actividad, señalando como principales causas la utilización de diferentes directorios para las distintas ramas de actividad que no se ajustan a los mismos criterios de definición en todos los casos y que adolecen de cualquier actualización y el solapamientos entre directorios y actualizaciones dispares en forma y tiempo que dificultan el tratamiento operativo de la Encuesta.

Como primer paso para dar solución a los inconvenientes observados se propone disponer de un directorio único para todas las ramas de actividad encuestadas, utilizando el Directorio de Unidades de Cotización proporcionando por el Ministerio de Sanidad y Seguridad Social, pues aunque no se ajusta a la unidad de observación de la Encuesta (el establecimiento), ya que se refiere a centros de cotización, tiene como grandes ventajas el ser actualizado cada año y el ser un directorio homogéneo para todas las ramas de actividad de la Encuesta.

Las explicaciones anteriores, realizadas con cierto detalle, ilustran sobre las dificultades técnicas habidas con la Encuesta de Salarios hasta su fundamentación en 1981 sobre la base sólida de un directorio convenientemente gestionado y actualizado. Hay que señalar, que las dificultades de infraestructura estadística encontradas en la Encuesta de Salarios se han podido sobrellevar, con mayor o menor acierto, gracias al hecho de que las estimaciones básicas que ofrecen, salario/hora y ganancia por trabajador, son estimaciones de promedios y estas estimaciones son más fáciles de lograr y más estables frente a los clásicos defectos que pueden ofrecer los directorios. No sucede así con las estimaciones de nivel en donde determinados defectos de los directorios pueden imposibilitar su obtención.

Similarmente a como hemos hecho con la Encuesta de Salarios, podríamos establecer una casuística para otras encuestas como las industriales, de comercio interior o las de estructura de las explotaciones agrícolas, en las que ha habido que sortear, con mayor o menor ingenio técnico e incluso con mayores costes económicos, las deficiencias de la infraestructura estadística contenida en los directorios sobre los que se apoyan. No han tenido estos problemas, por ejemplo, las encuestas de morbilidad hospitalaria, por la existencia de un libro de registro de ingresos y altas de enfermos en los hospitales (Real Decreto 1360/1978), y que constituye un ejemplo claro de posibilitar el trabajo estadístico por creación obligada de un registro en el que basarlo; las encuestas de gastos de enseñanza, porque se dispuso de correctos directorios de los centros de enseñanza, las encuestas de movimientos de viajeros en establecimientos turísticos, para las que la guía de hoteles ofrece un adecuado directorio, cuando ha estado bien gestionado, o las de transportes, que son posibles por la existencia obligada de tarjetas de autorización para el transporte que sirven como directorio.

En el sector servicios, exceptuando las estadísticas ya existentes de comercio, hostelería y transporte, es donde existió, hasta muy recientemente, lagunas de información estadística y que en los últimos años están empezando a cubrirse con diversas encuestas: Encuesta sobre la Estructura de las Empresas de Restauración 1989 y 1994, Encuesta Piloto de Estructura de Empresas Hoteleras 1992, Encuesta piloto de Agencias de Viaje 1993, Encuesta de Servicios Técnicos 1992, Encuesta de Empresas de Servicios Audiovisuales 1992, Encuesta de Coyuntura de Comercio al por Menor 1994, Encuesta de Empresas de Consultoría y Asesoramiento 1993, Encuesta sobre Actividades conexas al Transporte y Comunicaciones 1994 y Encuesta sobre las Empresas de Servicios de Alquiler de Maquinaria y Equipo 1994.

En el sector servicios todavía quedan parcelas por cubrir, pero en estos momentos se cuenta ya con la infraestructura estadística que va a posibilitar la próxima realización de encuestas que subsanen estas deficiencias, con la misma infraestructura que posibilitó las encuestas recientes del sector servicios. Esta infraestructura no es otra que la reciente creación del ***Directorio Central de Empresas***, en siglas ***DIRCE***, que por su importancia merece mención aparte.

El Directorio Central de Empresas (DIRCE)

En el acto de presentación del DIRCE, el entonces presidente del INE, José Quevedo, ilustraba la importancia que supone esta nueva herramienta del trabajo estadístico que facilitará las tareas, no sólo del INE, sino de todas las Instituciones del Sistema Estadístico Nacional, y que eliminará los sufrimientos profesionales que las deficiencias de los directorios económicos han acarreado. En términos comparativos podríamos decir que el DIRCE supone, respecto a las encuestas dirigidas a empresas e instituciones, el mismo avance que supuso el diseño de la Encuesta General de Población para las encuestas dirigidas a viviendas familiares y personas que las habitan, basado en la división del territorio nacional en secciones estadísticas.

El INE inició los trabajos de elaboración del DIRCE en 1987 y la sensibilidad existente a nivel nacional para potenciar su construcción entró en resonancia con análoga sensibilidad de la oficina estadística Europea (EUROSTAT) que, recogiendo similares necesidades en los distintos países de la Unión Europea, estableció un reglamento comunitario, en julio de 1993, obligando a los países miembros de la Unión Europea a tener disponible un directorio de empresas para usos estadísticos a 1 de enero de 1996 y un directorio de unidades locales a 1 de enero de 1997.

De manera sucinta diremos que el DIRCE trata de reunir en un directorio único todas las empresas españolas, siendo su objetivo básico hacer posible la realización de encuestas por muestreo dirigidas a empresas, pues, evidentemente, una empresa no podrá ser seleccionada en una encuesta si no figura reseñada en el directorio, debidamente actualizado. Es pues una pieza clave, columna vertebral de cualquier investigación por muestreo dirigida a las empresas. Los datos básicos que ofrece el DIRCE son: la identificación de la empresa, la localización, la clasificación por actividad económica principal según la Clasificación de Actividades Económicas 1993 (CNAE-93), y la clasificación por intervalos de asalariados.

Para obtener los datos del DIRCE se han utilizado diversas fuentes de entrada sometidas a fuertes tratamientos de armonización y depuraciones. Las fuentes más importantes de entrada son:

- Fuentes fiscales procedentes de la Agencia Estatal de Administración Tributaria y de la Comunidad Foral de Navarra
- Cuentas de cotización de la Seguridad Social
- Directorio de Locales del País Vasco
- Otras fuentes (Registro Mercantil, Censo de Locales 1990)

El DIRCE relaciona, por primera vez en España, un total de 2.301.559 empresas clasificadas según actividad económica principal, intervalos según número de asalariados, condición jurídica y geográficamente, en desgloses provinciales (de este cómputo se excluyen la agricultura, ganadería, pesca, las administraciones públicas, las actividades de comunidades de propietarios, el servicio doméstico y los organismos extraterritoriales).

Para dar idea final de los trabajos de formación del DIRCE, se señala que se tratan informáticamente, cada año, del orden de seis millones de registros provenientes de fuentes distintas y cuyas metodologías hay que armonizar (un millón de registros de la Seguridad Social, cuatro millones y medio de fuentes fiscales y quinientas mil de otras fuentes).

Señalar, por último, que como utilidad inmediata a la existencia el DIRCE, se realizaron trece encuestas de carácter económico dirigidas a empresas que no podrían haberse realizado sin su existencia, y se prestó apoyo sensible a otras encuestas que vienen realizándose. Por ello no han de regatearse esfuerzos para mantenerlo y mejorarlo, y de no ser así su deterioro y desactualización nos llevarían inevitablemente a épocas pasadas.

Debe aumentarse el grado de coordinación y sensibilización, en este terreno de la explotación de registros administrativos y su debida informatización, con las distintas administraciones y organismos públicos y, sin duda, el fruto será la posibilidad de obtención de nuevas estadísticas, mejora de muchas otras y quizás, en determinados casos, aliviar el peso creciente de colaboración que actualmente soportan las unidades informantes para cumplimentar las necesarias estadísticas oficiales por el aprovechamiento directo de registros administrativos.