Taller Regional sobre el uso de Muestreo en las Encuestas Agrícolas

METODOS DE MUESTREO

POR

MIGUEL GALMÉS

FOOD AND AGRICULTURE ORGANIZATION

UNITED NATIONS

MONTEVIDEO, URUGUAY, 20 - 24 DE JUNIO 2011

PREFACIO

Las notas que aquí se presentan tienen como objetivo servir de referencia a las presentaciones que sobre Métodos de Muestreo en Censos y Encuestas Agrícolas se realizan en el marco del "Taller Regional sobre el uso del muestreo en encuestas agrícolas", realizado en Montevideo, Uruguay entre el 20 y el 25 de junio de 2011.

Estas notas, son una actualización del documento preparado por el autor para un cursillo de 20 horas de duración dictado en el marco del "Programa Subregional de Capacitación en Censos y Encuestas Agrícolas" desarrollado en San Salvador, El Salvador en diciembre de 1996.

Pretenden ser un documento metodológico de referencia para quienes son responsables del diseño y de la implementación de censos y encuestas agrícolas por muestreo.

Luego de una introducción general sobre el papel del muestreo en los censos y encuestas agrícolas, presenta los principales aspectos metodológicos del diseño y se introduce luego en el desarrollo detallado de los métodos de muestreo más comúnmente utilizados. Se dan pautas generales para la elección de un diseño de muestreo, se presentan algunos ejemplos prácticos de aplicación y por último se dedica un capítulo al control de calidad de los datos.

Miguel Galmés

Consultor FAO

ÍNDICE

INDICE (CONT.)

MÉTODOS DE MUESTREO

I. INTRODUCCIÓN.

En el lenguaje corriente, un censo comprende la enumeración completa de la población de interés[1] . Así por ejemplo, en un censo de población se tomará información sobre cada una de las personas del ámbito correspondiente (país, región, etc.) o en un Censo Nacional Agropecuario se relevará el conjunto de explotaciones agrícolas del país. , la palabra "censo" se contrapone a "muestra" (o a "encuesta por muestreo") en el sentido que una "muestra" implica la enumeración de una parte de la población. Sin embargo, como veremos enseguida, esto no es estrictamente así' Sea cual sea la interpretación que se le dé a la palabra "censo", el muestreo juega un papel fundamental en los mismos.

1. En primer lugar y como aproximación muy general un censo - como enumeración completa - puede verse como una "encuesta por muestreo" con una tasa de muestreo del 100%, es decir con inclusión forzosa de todas las unidades de la población, de manera que todas las derivaciones hechas para las encuestas por muestreo son aplicables a los censos como caso particular de ellas.
2. En un sentido amplio, FAO define un censo agrícola como "una operación estadística en gran escala y periódica con el fin de recoger información cuantitativa sobre la estructura de la agricultura de un país", de manera que no se excluye, en esta definición que el propio censo agropecuario sea tomado sobre una base de enumeración muestral. En este caso, las variables estructurales de la agricultura se relevan a través de una gran encuesta nacional (lo que en programas anteriores se llamaba “censo por muestreo”).
3. El presente programa de la FAO (CAM 2010) asigna un papel fundamental al muestreo ya que en el enfoque modular se sugiere que los módulos suplementarios al módulo central se levanten por muestreo.
4. Incluso en los censos (o módulos censales) conducidos por enumeración completa, en diversas etapas de su ejecución deberían aplicarse procedimientos de muestreo, fundamentalmente en lo que hace referencia al control de calidad de la información.
5. Uno de los objetivos fundamentales de un censo agropecuario es proveer un marco para las encuestas agrícolas por muestreo, que son parte central del programa de estadísticas agropecuarias continuas de cualquier país. Por lo tanto el conocimiento de los métodos de muestreo resulta crucial para la construcción de marcos adecuados a partir del censo.

[1] Así el Diccionario de la Academia define "censo" como 'padrón o lista de los componentes de una población"

II. ASPECTOS METODOLÓGICOS DEL DISEÑO.

En cualquier investigación estadística, y en particular, en una investigación de la escala de un censo agrícola, existen muchos problemas metodológicos a considerar. Como ya se ha visto al estudiar el "Programa para el Censo Agrícola Mundial 2010", estos aspectos tienen que ver con: los objetivos y alcance del censo, la unidad estadística, la cobertura, la forma de enumeración, la periodicidad, el período de referencia, el período de enumeración, el marco censal, etc.

Para la consideración de los métodos de muestreo y su aplicación, es importante detenerse en los aspectos metodológicos que hacen al diseño. Su definición, las formas de enumeración censal, consideraciones generales sobre los diferentes tipos de diseño de muestreo para los censos agrícolas, y otros usos de técnicas de muestreo en la ejecución censal.

1. Diseño.

Deben distinguirse dos tipos de diseño: **el diseño de muestreo** y el **diseño de la encuesta**. Por **diseño de muestreo** de una encuesta o de un censo agrícola conducido por enumeración muestral se entiende al conjunto de técnicas para seleccionar una muestra probabilística y los métodos de estimación empleados para inferir los valores de las características bajo estudio a partir de la muestra seleccionada. Así por ejemplo cuando se habla de una Muestra Aleatoria Estratificada se está haciendo referencia al diseño de muestreo, de la misma manera que cuando se expresa que los estimadores usados fueron de razón al tamaño, por ejemplo. Por su parte el **diseño de la encuesta** se refiere a las definiciones, métodos y procedimientos concernientes a todas las fases necesarias para conducir la operación: el diseño de muestreo, la selección y entrenamiento de personal, la organización de la logística para la entrega y recepción de los materiales, los procedimientos de recolección y procesamiento de datos, y los análisis de datos necesarios para brindar los resultados finales.

Un censo agrícola puede tener diferentes diseños de encuesta. Por ejemplo, puede ser ejecutado mediante distintos procedimientos de enumeración: entrevista directa con formulario de papel, entrevista directa con PDA, encuestas por correo o por teléfono, mediciones objetivas, etc.; puede ser por enumeración completa o por muestreo o una combinación de ambos; puede utilizar distintos tipos de unidades muestrales o distintos métodos de selección de la muestra o procedimientos de estimación y así sucesivamente.

2. Formas de enumeración censal.

Un censo agrícola puede ser conducido a través de enumeración completa, muestreo una combinación de ambos. La **enumeración completa** implica obtener información de

todas las explotaciones agrícolas del país, mientras que la **enumeración por muestreo** obtiene información sólo de un número predeterminado de unidades[2].

Como consecuencia de las anteriores definiciones, en el caso de un censo por enumeración completa (generalmente llamado simplemente "censo") los valores totales para las diferentes características se obtienen mediante agregación de los valores individuales para cada explotación agrícola del país. De manera que los resultados censales no incluyen "errores de muestreo". Ello no implica que no incluyan errores, pero éstos serán todos "errores no muestrales" como se verá más adelante. Por su parte, los censos agrícolas basados en enumeración por muestreo, son encuestas agrícolas por muestreo (encuestas particulares donde las variables a relevar son "variables censales", es decir las que hacen a las características estructurales de la producción agropecuaria). Ello significa que, en este caso, una muestra probabilística de explotaciones agrícolas es seleccionada y que los totales poblacionales se obtendrán infiriéndolos de los resultados muestrales mediante los estimadores que se definan en el correspondiente diseño de muestreo. Por lo tanto en estos casos es posible determinar la precisión estadística de las estimaciones obtenidas.

Un censo conducido sobre la base de enumeración completa permite:

a. Obtener resultados a nivel de pequeñas áreas o unidades administrativas. Esta información es útil para la planificación local, para estudios de proyectos de inversión, para análisis agroecológicos, etc.

b. Obtener información confiable para aquellos cultivos que se siembran en áreas pequeñas o en zonas muy localizadas y que pueden ser de gran importancia económica (igualmente algunas especies animales "raras").

c. Construir el marco de lista primordial para las encuestas agrícolas "continuas". Estas encuestas pueden ser planeadas mucho más eficientemente si los resultados censales están disponibles para pequeñas áreas y además los resultados censales pueden ser usados para mejorar el diseño de muestreo y los procedimientos de estimación.

d. Utilizar personal estadístico menos calificado que el necesario en el caso de una enumeración por muestreo. Este aspecto es importante cuando no se tiene suficiente personal con experiencia en conducir encuestas por muestreo.

e. Procesar los datos censales más fácilmente ya que el procesamiento es la simple agregación de los resultados por cuestionario. En cambio en el caso de enumeración por muestreo, la expansión de los resultados y el cálculo de las varianzas para estimar los errores muestrales requieren más labor de programación.

[2] En programas anteriores al del 2000, FAO utilizaba el término "censos por muestreo" para este tipo de enumeración. Para precisar el lenguaje y evitar confusiones semánticas, el CAM 2000 introdujo el término "enumeración por muestreo" para referirse a este tipo de investigaciones.

Por otra parte, un censo por enumeración completa, tiene los siguientes inconvenientes frente a la enumeración por muestreo:

a. Es más caro y lleva más tiempo.

b. Requiere más personal de campo (enumeradores y supervisores) y muchas veces no se posee la cantidad necesaria con las habilidades requeridas.

c. La cantidad de datos a ser procesados es mucho mayor que en una enumeración por muestreo.

d. Por la extensión del trabajo y por las limitaciones que implica entrenar a un conjunto muy grande de enumeradores y supervisores los errores de relevamiento son más difíciles de controlar.

Esta es la lógica que está detrás del enfoque modular del CAM 2010: para construir buenos marcos y tener la información estructural básica se tiene al módulo central por enumeración completa y para complementar y profundizar la información sobre aspectos de interés se propone la enumeración por muestreo de las variables que integran los módulos suplementarios.

3. Otros usos del muestreo en los censos agrícolas.

Además de la recolección de datos sobre la base de una muestra de explotaciones para obtener inferencias " sobre estructura y producción agrícola, el muestreo tiene otras importantes aplicaciones en el desarrollo de los censos agrícolas. Algunos de estos usos son:

1. En algunos censos la recolección de datos se organiza en dos partes:

   A) un cuestionario simple aplicado a todas las explotaciones agrícolas (censo de enumeración completa). Este cuestionario comprende las variables del módulo central en el caso del CAM 2010. Este cuestionario puede ser levantado por enumeradores menos calificados;
   B) uno o varios cuestionarios más complejos para investigar en profundidad algunas variables, aplicados a una muestra de las explotaciones del país. Estos cuestionarios corresponden a las variables de los módulos suplementarios. Estos cuestionarios pueden ser responsabilidad de enumeradores más calificados.

2. Un censo puede combinar la enumeración completa de explotaciones grandes con la enumeración por muestreo del resto.

3. Para checar la calidad de una lista de explotaciones agrícolas o de hogares que sirva de marco al censo. La existencia de un marco apropiado es fundamental para el éxito de la enumeración censal ya sea ésta efectuada mediante enumeración completa o por muestreo. Cuando el marco de lista original es viejo se corre el

riesgo de que esté desactualizado, entonces se utilizan procedimientos de muestreo para evaluar su cobertura extensión de las omisiones, etc. y decidir en definitiva si puede o no ser usado.

4. Para seleccionar las explotaciones o las áreas donde aplicar el censo piloto. El censo piloto es una actividad esencial en la conducción de un censo agrícola. El mismo permite reproducir "en pequeño" la operación censal en todos sus aspectos: desde el entrenamiento del personal hasta la producción de tabulados finales. La selección de las zonas donde aplicar el censo piloto (o aún de las explotaciones agrícolas a quienes .se le aplicará) puede hacerse por muestreo a fin de cubrir las diferentes características del país de manera lo más objetiva posible. Una actividad previa al censo piloto es el "pre-test" de los cuestionarios y otros materiales. Este pre-test se hace a una escala reducida, generalmente al comienzo de las tareas preparatorias del censo, también este pre-test puede hacerse por muestreo (sobre todo en los países grandes).

5. Para organizar la supervisión del trabajo de campo, una submuestra de áreas y de explotaciones es seleccionada. Esta forma de conducir la supervisión de los enumeradores combina el elemento sorpresa con la posibilidad de establecer el tipo y extensión de los errores que están siendo cometidos con el fin de corregirlos para el futuro y dar factores de corrección para los casos ya relevados con error.

6. Para organizar el trabajo de campo como un conjunto de muestras independientes de tal manera que cada una por separado puede dar estimaciones confiables de los. totales de las variables bajo estudio. La consideración conjunta de estos resultados dan una medida de la confiabilidad de los datos censales.

7. Para diseñar encuestas por muestreo concomitantemente con el censo a fin de estudiar características especiales, por ejemplo: existencia de ganado o aves de corral en zonas urbanas no cubiertas por el censo; o tomar medidas objetivas de ciertas variables.

8. Para diseñar las encuestas de post-enumeración. Estas encuestas son muy importantes a fin de evaluar la calidad de los datos censales. Esta calidad de los datos hace referencia tanto a la bondad de la información recogida como a la cobertura del censo. En el primer punto se tiene la evaluación de los errores de respuesta y en el segundo, la evaluación de la completitud del censo. Más adelante volveremos sobre estas encuestas de post-enumeración al considerar los controles de calidad

9. Para preparar resultados anticipados del censo. Se toma una muestra cuestionarios que se procesa antes que el resto (incluso puede procesarse manualmente) a fin de dar resultados adelantados de los principales totales del censo.

10. Para realizar controles de calidad de la edición, codificación, entrada de datos y procesamiento, Sobre este punto también volveremos más adelante al considerar los controles de calidad.

11. En algunas situaciones donde los recursos no son suficientes para el análisis de todos los datos recogidos, las tabulaciones finales pueden obtenerse sobre la base de una muestra como una solución de emergencia (no deseable por cierto).

4. Principales tipos de diseños de muestreo para los censos y encuestas agrícolas. Generalidades.

Más adelante, en este trabajo, se analizarán con detalle los principales diseños de muestreo para las investigaciones agrícolas. Se considera útil, sin embargo, a fin de situar al lector, brindar aquí un resumen de los principales tipos de diseños muestrales para los censos agrícolas conducidos por enumeración por muestreo y para las encuestas agrícolas.

En lo que sigue se hará referencia al "marco" utilizado, es importante conceptualizar qué se entiende por **marco**.

El marco puede definirse como los materiales, procedimientos y mecanismos que identifican, distinguen y permiten acceder, a los elementos de la población objetivo. El marco se compone de un conjunto finito de unidades a las cuales se les aplica el esquema de muestreo. Las reglas o procedimientos para ligar las unidades del marco con las de la población objetivo son una parte fundamental del marco. (Por ej. si la población objetivo está constituida por explotaciones agrícolas y se tiene un listado de viviendas rurales, este listado y las reglas definidas para identificar explotaciones a partir de las viviendas forman parte del marco). El marco también incluye información auxiliar usada tanto para el diseño (en el ejemplo, el número de personas por vivienda por ejemplo) como para la estimación (variables auxiliares para estimadores de razón).

La construcción del marco muestral es una de las tareas más difíciles que debe encarar el diseñador de encuestas. Idealmente las características y naturaleza de la población objetivo deberían determinar el tipo de marco a usar; en la práctica, sin embargo, el marco disponible, determina - muchas veces - a la población objetivo a considerar.

En una enumeración por muestreo, ya se trate de un censo o una encuesta, pueden distinguirse básicamente dos tipos de marcos: **muestreo basado en marcos de**

lista y **muestreo basado en marcos de área.** A veces estos tipos se combinan dando lugar a un tercer tipo, conocido como **muestreo con marcos múltiples**.

4.1. Muestreo probabilístico con marcos de lista.

Son los más comúnmente usados de los procedimientos de muestreo. Usualmente las unidades que se seleccionan finalmente (unidades de última etapa) son las explotaciones agrícolas o la dirección del productor. La lista que sirve de marco para la selección de las muestras se obtiene de un empadronamiento previo en el caso de los censos por enumeración por muestreo o del propio censo levantado por enumeración completa, en el caso de las encuestas posteriores al censo. Otras listas utilizadas como marco son: los censos de población y vivienda, listados de asociaciones de productores, catastros, registros de contribución territorial, listados de beneficiarios de reforma agraria, etc. Las explotaciones agrícolas listadas pueden estratificarse, conglomerarse, puede ser seleccionadas en una o en varias etapas, pueden seleccionarse mediante muestreo aleatorio simple o mediante muestreo con probabilidad proporcional a una medida de tamaño, etc. porque, generalmente, la información necesaria se encuentra en el listado.

El principal problema de los diseños basados en los marcos de lista es su desactualización ya que la explotación agrícola es una entidad dinámica y por lo tanto a medida que pasa el tiempo, la población listada tiende a apartarse de la población objetivo de la encuesta. Además las listas a veces son incompletas o inexactas, contienen duplicaciones o tienen omisiones. Estos marcos de lista deberían actualizarse periódicamente pero ello es costoso y generalmente inabordable en nuestros países. A menudo este problema se soluciona tomando la muestra en dos etapas: se seleccionan primero conglomerados (por ejemplo distritos o segmentos censales) con algún procedimiento de muestreo (generalmente con probabilidad proporcional a alguna medida de tamaño) y para los conglomerados seleccionados, se listan las explotaciones que contienen y, en la segunda etapa, se seleccionan explotaciones. Esto permite actualizar el marco sólo en aquellas unidades seleccionadas en la primera etapa.

En encuestas con estos diseños, el encuestador usualmente completa un cuestionario para cada explotación seleccionada mediante una entrevista con el productor. A veces esta forma de recolección se complementa con mediciones objetivas sobre el terreno.

4.2. Diseños de muestreo de área.

Un diseño de muestreo de área es un método de muestreo probabilístico en el cual las unidades de última etapa son áreas de tierra llamadas *segmentos* y las probabilidades de selección son proporcionales a su superficie. Toda el área a cubrir

se divide en segmentos no superpuestos que constituyen las unidades de muestreo. Usualmente estos segmentos se estratifican de acuerdo a características del uso del suelo.

Tres tipos de segmentos se utilizan para diseños de muestreo de encuestas sobre estructura y producción agrícola (incluyendo los censos conducidos por enumeración muestral):

a. Segmentos con límites físicos reconocibles. Los límites de los segmentos son caminos, ríos, canales, vías de ferrocarril, divisorias de agua, etc. fácilmente reconocibles en el terreno y dibujados en mapas. En este caso, para cada estrato, los segmentos se definen de aproximadamente el mismo tamaño y se utiliza un factor de expansión constante para cada estrato (como se analizará más adelante al estudiar el muestreo estratificado). El diseño puede verse como una muestra estratificada de *tramos.* Los tramos son partes de explotaciones agrícolas incluidas dentro de los límites del segmento, o las tierras del segmento que no pertenecen a ninguna explotación agrícola. Así un *tramo* está determinado por los límites de un segmento y por las explotaciones con tierra en el segmento. Un tramo dentro de un segmento se subdivide a menudo en *campos* que tienen límites reconocibles y con diferentes usos de la tierra. Una explotación se compone de uno o más tramos. En este tipo de muestras, la recogida de datos se hace mediante un cuestionario que se completa para cada trozo de terreno de la unidad de cada segmento. Esta forma de recoger los datos puede acompañarse de medidas objetivas utilizando fotos aéreas o fotos de satélite. Para ello, para cada tramo de un segmento muestreado, los enumeradores delinean sobre la foto del segmento los límites del tramo y los límites de todos los campos incluidos en el tramo y verifican los cultivos plantados y otros usos de la tierra para cada campo y también la información que les brinda el productor. Estas áreas agrícolas identificadas son luego medidas en la oficina sobre la foto, usando un planímetro o un sistema de computación gráfica y luego se expanden a fin de tener una estimación objetiva del área agrícola.

b. Segmentos cuadrados o rectangulares. Los segmentos están definidos por líneas rectas que se cortan en ángulos rectos en puntos establecidos por coordenadas geográficas en mapas. En este caso se utilizan procedimientos de muestreo con cuadrículas. Usualmente se selecciona una muestra Estratificada de segmentos cuadrados. La recolección de datos en este caso es similar a la descrita en el punto a. el problema radica en la dificultad de obtener de los productores datos confiables para los tramos de un segmento que no puede ser observado sobre el terreno.

c. Segmentos que coinciden con la tierra de las explotaciones agrícolas. En este caso una cuadrícula es superpuesta a un mapa de los estratos y una muestra de puntos es seleccionada (cruce de líneas de las cuadrículas). Los puntos obtenidos son identificados sobre el terreno y las explotaciones agrícolas que los contienen forman la muestra de áreas. Por la forma de selección, es un muestreo de unidades agropecuarias seleccionadas con probabilidad proporcional a su tamaño (ver muestreo "ppt" más adelante). Los encuestadores tomarán la información de las explotaciones así seleccionadas. Si se requieren mediciones objetivas deberán hacerse, en este caso midiendo los campos de las explotaciones seleccionadas lo que resulta más engorroso que la medición hecha en la oficina sobre la foto aérea.

4.3. Diseños basados en marcos múltiples.

Un diseño que combina un diseño de muestreo de áreas con uno de lista es llamado **diseño de muestreo de marcos múltiples**. Los estimadores, en este caso, combinan los de la muestra de áreas con los de la muestra de lista para cada variable investigada. Se hacen estimaciones separadas para cada diseño y luego se suman. Por ejemplo, de una lista proveniente de un censo por enumeración completa, se seleccionan las explotaciones mayores a 20 hectáreas, esto conforma el "marco de lista" para un diseño de muestreo (por ejemplo estas explotaciones se encuestan mediante Muestreo Aleatorio Estratificado). Luego, se confecciona un marco de áreas donde las explotaciones incluidas en el marco de lista son eliminadas, es decir: en la muestra seleccionada de segmentos no habrá ningún tramo correspondiente a explotaciones incluidas en el marco de lista. Este "cribado" de las áreas requiere mucho cuidado y una inversión importante de recursos. Lo usual es que se elabore un listado de "explotaciones especiales" (por ejemplo, grandes fundos, explotaciones avícolas intensivas, explotaciones porcinas intensivas, explotaciones con cultivos "raros" de alto valor económico, etc.) y esta lista se enumera completamente. Los resultados de esta enumeración se adicionan luego a las estimaciones obtenidas a partir de una muestra de áreas donde se han eliminado previamente las explotaciones especiales incluidas en la lista.

5. Comparación entre diseños.

Un breve resumen de la comparación entre estos tipos de diseño realizada en el documento de FAO "Encuestas agrícolas con múltiples marcos de muestreo" establece:

- Los diseños basados en marcos múltiples son preferibles a los diseños basados sólo en área ya que el trabajo adicional involucrado no es significativo y brinda mejores estimaciones para items importantes;

- El insesgamiento que establece la teoría puede perderse por el uso de marcos de lista que se desactualizan fácilmente o son incompletos. Este problema no se tiene

en los marcos múltiples ya que proveen cobertura completa del área de interés;

- Para la estimación del área agrícola, el método basado en marcos múltiples que combina muestreo de área de segmentos con límites reconocibles con un marco de lista de explotaciones especiales es más eficiente que el diseño basado sólo en marcos de lista;

- En un diseño basado en marcos de áreas, los errores debidos a medidas de área o declaraciones incorrectas de los productores son menores ya que se realizan mediciones sobre fotos aéreas a fin de checar áreas reportadas del campo. La implementación de mediciones objetivas en encuestas basadas en marcos de lista es mucho más complicada, ya que requiere la medición en el momento de visitar el campo y si la explotación tiene parcelas distantes, estas mediciones se tornan casi imposibles.

- Para encuestas periódicas, un diseño basado en áreas es más estable que uno de lista que se desactualiza rápidamente. La única forma que pierde actualidad un marco de áreas es cuando la agricultura se desplaza o se extiende sobre nuevas zonas no cubiertas por el marco, pero cambios en el uso del suelo, en el número y ubicación de las explotaciones, no introducen sesgo en un muestreo basado en marcos de áreas.

- Un muestreo basado en áreas da mejores medios para estimar rendimientos y su pronóstico ya que permiten mediciones objetivas (de corte y pesaje) en los propios campos seleccionados;

- La implementación de un muestreo basado en áreas requiere más experiencia técnica que la de uno basado en marcos de lista;

- La implementación de un muestreo basado en marcos de áreas requiere mapas muy precisos para identificar y medir las áreas.

- Puede no ser posible utilizar marcos de área en países con grandes dificultades de terreno o donde un volumen importante de las variables investigadas corresponde a explotaciones cuyos productores viven lejos de sus explotaciones o son difíciles de ubicar;

- Un diseño basado en muestreo de áreas es más costoso que uno de lista;

- A veces hay problemas para identificar los límites de los segmentos a partir de fotos aéreas, imágenes de satélite o de los mapas. esto es particularmente cierto en países donde por razones climáticas se cultiva por turnos en ciertas áreas o el terreno es cubierto con selva y los límites identificables se pierden;

- Por último, los avances tecnológicos en el manejo de imágenes por computadora hace que los métodos de muestreo de áreas puedan utilizar imágenes de satélite o imágenes digitalizadas como parte de sistemas de información geográfica (GIS) y otros procedimientos automáticos y técnicas para selección de muestras y análisis de los datos.

III. MÉTODOS DE MUESTREO

A continuación, se presentan los principales métodos de muestreo aplicables para la estimación de características de la estructura y de la producción agrícola. Estos métodos generales son aplicados en las diferentes circunstancias enumeradas anteriormente.

Algunas definiciones previas y acuerdos sobre la notación son necesarios.

1. CONCEPTOS PREVIOS.

1. 1. Poblaciones y muestras.

El muestreo consiste en seleccionar parte de una población y - a partir de lo observado en esa parte - inferir a la población. Esta inferencia (extrapolación) es científica (estadística) cuando va acompañada de una medida del error cometido por el hecho de tener información parcial para generalizar (error de muestreo).

Una **población** es un conjunto de elementos. Estos elementos pueden ser muy variados: explotaciones agrícolas, hogares, personas, segmentos de área, etc.. Estos elementos son, generalmente, la unidad de información, o están estrechamente vinculados a ella. La población consiste de un número **N** de elementos que llamaremos **unidades**. Por ejemplo en una población de 53,426 explotaciones agrícola, N=53,426 y cada explotación agrícola es una unidad. Las unidades de la población pueden identificarse con los números 1,2,3,...N. Cada unidad tiene asociados valores de **variables** de interés (en nuestro caso: el área total, el área cultivada, el número de personas que viven en la explotación, el número de cabezas de ganado, una variable que valga 1 si el productor reside en la explotación y 0 en caso contrario, el porcentaje de tierras con montes y bosques, son algunos ejemplos de variables de interés). El valor que toma cada variable para cada unidad de la población es considerado un número fijo (no aleatorio) y por supuesto, desconocido a priori. Llamaremos **y_i al valor que toma la variable y en la i-ésima unidad de la población (i=1,2,..N)**. Por ejemplo si y = área cultivada y la explotación que en la lista anterior aparece con el número 56 tiene 4.5 hectáreas cultivadas, entonces: $y_{56} = 4.5$.

Una **muestra** consiste en una parte de la población. Sólo una muestra es seleccionada y los valores de cada variable en las unidades pertenecientes a la muestra son registrados. Entonces, para cada variable se tiene un **conjunto de observaciones.** Llamemos **n** al

número de elementos de la muestra. Una vez seleccionada la muestra y registrados los n valores observados para cada variable (y) tendremos el conjunto de observaciones: $O = \{y_i : i = 1,2, ..., n\}$. Una precisión en la notación usada es necesaria: hemos designado con el mismo subíndice "i" a los elementos de la población y a los de la muestra, en realidad los elementos de la muestra son algunos de la población y por lo tanto lo correcto sería utilizar un doble subíndice donde el primero indicaría el elemento de la población y el segundo el de la muestra (por ejemplo si en una muestra el quinto elemento de la población es seleccionado en primer lugar en la muestra, el valor correspondiente de la variable y debería anotarse $y_{5,1}$ y no y_1), sin embargo para evitar las complicaciones de una notación que ya de por sí se torna pesada, utilizaremos el mismo subíndice para ambos casos, indicando cuándo se trata de valores poblacionales y cuándo de muestrales.

Como ya vimos, se entiende por **diseño de muestreo** al conjunto de técnicas para seleccionar una muestra probabilística y los métodos de estimación empleados para inferir los valores de las características bajo estudio a partir de la muestra seleccionada. Cada diseño muestral queda definido cuando a cada posible muestra (M) se le asigna su probabilidad de selección (**P(M)**).

El problema es aproximar (estimar) algún parámetro de la población (por ejemplo el número total de cabezas de ganado o el área bajo riego) a partir sólo del conjunto O. Para ello es necesario definir alguna función útil de las observaciones ($T = t(y_1, y_2,..y_n)$) llamada un **estimador** y calcularla numéricamente para la muestra observada para lo cual es preciso que T no contenga valores no observables (es decir parámetros desconocidos luego de extraer la muestra), ese valor calculado para T (t) es una **estimación**. Esta estimación será el valor inferido del parámetro poblacional.

Conviene acá distinguir diferentes **niveles de poblaciones**. Hay una **población encuestada** es la población representada por la muestra realmente observada. Hay una **población marco** que es la población de la cual se extrajo efectivamente la muestra. Esta población marco es mayor que la encuestada porque incluye las *no respuestas* (informantes ausentes, rechazos, etc.). Hay una **población objetivo** que es diferente a la población marco pues incluye los elementos que por error de cobertura no fueron incluidos en la población marco (*Subcobertura del marco*) y excluye a los erróneamente incluidos (*sobrecobertura*). De manera que ascendemos desde la muestra a la población encuestada, luego al marco y de ésta a la población objetivo. Pero las inferencias también se hacen desde las poblaciones objetivo a una variedad más amplia de otras poblaciones: por ejemplo, a partir de las estadísticas de un año particular se hacen inferencias para el futuro (usando modelos econométricos por ejemplo), este conjunto más amplio al cual pueden dirigirse las inferencias se denomina **población inferencial**.

1.2. Errores muestrales y no muestrales.

Al realizar el proceso de inferencia, como sólo se observa una parte del todo (n de N), estamos cometiendo un error, necesariamente. Dicho de otra manera, si observáramos los N elementos (censo) y no hubiera errores de observación o de otro tipo, teóricamente, conoceríamos el valor exacto del parámetro poblacional. Como se observa una parte hay un error implícito en el procedimiento provocado por la falta de información sobre el resto. Este error se conoce como **error de muestreo**. El tamaño de este error dependerá del diseño de muestreo usado (incluido el tamaño de la muestra). Si definido un tamaño (n) extrajéramos muchas muestras con el mismo diseño, obtendríamos diferentes conjuntos O y por lo tanto diferentes estimaciones del mismo valor poblacional. Si estas estimaciones difieren poco entre sí es de presumir que nuestra aproximación es buena. **Lo importante es que podemos calcular el error de muestreo observando una sola muestra**. Esto es así porque el error lo calcularemos en términos probabilísticos y dependerá de la distribución de la variable T en el conjunto de las muestras. Esta distribución, si bien es difícil en la mayoría de los casos calcularla exactamente podrá aproximarse con bastante grado de certidumbre (el Teorema Central del Límite juega un papel clave acá y los métodos de simulación Montecarlo serán sumamente útiles cuando aquel no pueda aplicarse o haya dudas sobre el cumplimiento de sus hipótesis). La preocupación por obtener buenas estimaciones ha derivado en un sinnúmero de diseños de muestreo y en nosotros está el ser capaces de adoptar el más eficiente (en términos de precisión vs. costo).

El supuesto básico es que cada unidad es observada sin error y que el único error final proviene del error de muestreo como se indicaba antes. Sin embargo, en la realidad aparecen **errores no muestrales**: errores de registración de datos, errores de cobertura, errores de respuesta, personas que no responden, personas que no se encuentran, imposibilidad de llegar al lugar seleccionado, errores de procesamiento, etc.. Más adelante nos ocuparemos de estos errores dentro del capítulo sobre controles de calidad de los datos.

1.3. Muestreo asistido por modelos.

En el enfoque anterior, la probabilidad entra sólo en la selección de la muestra, es decir a través del diseño. Pero en algunas situaciones es realista y práctico suponer un modelo de probabilidad para la población misma. El modelo puede basarse en el conocimiento de los fenómenos naturales que influyen en la distribución de los elementos de la población o también en modelos que resuman algunas características básicas. Por ejemplo puede usarse un modelo econométrico para relacionar, por ejemplo, el rendimiento de un cultivo con características agrometeorológicas y este modelo puede usarse tanto en el diseño de muestreo como en los estimadores a utilizar. Por ejemplo, se puede utilizar un modelo para detectar si hay correlación entre los valores de una variable determinada para diferentes explotaciones agrícolas y la distancia a que estos se encuentran uno de otro, por

ejemplo a menor distancia los valores son más parecidos y a mayor distancia menos, en ese caso un diseño sistemático de selección puede ser más eficiente que uno aleatorio simple.

2. PRINCIPALES DISEÑOS DE MUESTREO.

2.1. Muestreo Aleatorio Simple.

2.1.1. Definición.

El Muestreo Aleatorio Simple (M.A.S.) es el diseño de muestreo en el cual **n** unidades distintas son seleccionadas de las **N** de la población de tal manera que cada uno de los posibles subconjuntos de n elementos tomados de los N tiene igual probabilidad de selección. A partir de N elementos pueden formarse $\binom{N}{n} = \frac{N!}{(N-n)!\,n!}$ subconjuntos de tamaño n . Por ejemplo de una población de 50 elementos pueden obtenerse 10,272,278,170 subconjuntos diferentes de 10 elementos cada uno. Si tenemos un procedimiento de selección que garantice que la probabilidad de cada subconjunto M (muestra) es de $p(M) = \frac{1}{\binom{N}{n}}$ entonces estamos seleccionando mediante M.A.S. Puede probarse que un procedimiento que garantiza dicha probabilidad de selección para cada muestra consiste en extraer la muestra a partir de n selecciones de un elemento cada una de tal manera que en cada paso, cada unidad de la población no seleccionada anteriormente tiene igual probabilidad de selección. Esto es equivalente a realizar una secuencia de extracciones independientes de la población total con igual probabilidad de selección en cada extracción, descartando las selecciones repetidas y continuando hasta que **n** unidades diferentes hayan sido seleccionadas.

Con M.A.S. la probabilidad de que la i-ésima unidad de la población sea incluida en la muestra es: $p_i = \frac{n}{N}$ de manera que la probabilidad de inclusión es la misma para cada unidad. Hay otros diseños que dan igual probabilidad de inclusión a cada elemento pero *sólo el M.A.S. asigna igual probabilidad de selección a cada una de las muestras posibles.*

En la práctica la selección se hace numerando los elementos de la población de 1 a N y seleccionando mediante algún mecanismo de azar (bolillero, tablas de números aleatorios o generador de números aleatorios por computadora) n unidades distintas (de a uno y descartando las repetidas). Paquetes estadísticos de computación como el SPSS (Statistical Package for Social Sciences) o el SAS (Statistical Analysis System), mediante un simple comando extraen muestran aleatorias simples.

2.1.2. Estimadores, intervalos de confianza y tamaños de muestra.

a) Estimación de medias y totales.

a1) Estimadores.

En el M.A.S. la media muestral es un estimador insesgado de la media poblacional. Por estimador "insesgado" entendemos un estimador que si es calculado sobre todas las muestras posibles y todas las estimaciones obtenidas son promediadas, este promedio coincide con el valor poblacional, en nuestro caso, con la media poblacional (μ).

La media muestral se computa como:

[1] $$\bar{y} = \frac{1}{n}(y_1 + y_2 + \ldots + y_n) = \frac{1}{n}\sum_{i=1}^{n} y_i$$

A su vez, como el total poblacional Y = Nμ entonces:

[2] $$\hat{Y} = N\bar{y} = \frac{N}{n}\left(\sum_{i=1}^{n} y_i\right)$$

es un estimador insesgado de Y. Obsérvese que en este estimador la suma de los valores de la muestra (la suma de los elementos del conjunto O de observaciones) se expande multiplicándola por N/n que es el inverso de la fracción de muestreo. A este número (N/n) se le llama, por ello **factor de expansión**.

Con M.A.S. también la varianza muestral (s^2) es un estimador insesgado de la varianza de la población (σ^2) definidas como:

[3] $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \bar{y})^2$$

[4] $$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N} (y_i - \mu)^2$$

Se demuestra que la varianza del estimador $\bar{y}$ con M.A.S. es:

[5] $$var\ (\bar{y}) = \left(\frac{N-n}{N}\right)\frac{\sigma^2}{n}$$

y un estimador insesgado de esta varianza se obtiene simplemente sustituyendo la varianza poblacional por s^2 :

[6] $$\hat{var}\ (\bar{y}) = \left(\frac{N-n}{N}\right)\frac{s^2}{n}$$

La raíz cuadrada de la varianza del estimador es su error estándar. El error estándar estimado no es - en general - un estimador insesgado de σ.

La cantidad $\frac{N-n}{N}$ se denomina **factor de corrección por población finita.** Obsérvese que dicha cantidad es igual a: $1 - \frac{n}{N}$; (es decir: 1 menos la fracción de muestreo). Si la población es grande con respecto a la muestra (lo que es usual en las poblaciones utilizadas para encuestas agrícolas) dicha cantidad es cercana a 1 y la varianza del estimador es próxima a σ^2/n. Si por simplicidad de cálculo se omite el factor de corrección por población finita la ligera sobreestimación de la varianza verdadera que se obtiene no afecta significativamente los resultados. Obsérvese, por último que si n se aproxima a N la varianza del estimador se aproxima a cero. En el caso particular de n=N (censo) esta varianza es igual a cero.

Como el estimador insesgado del total es $\hat{Y} = N\bar{y}$, su varianza es:

[7] $$var\ (\hat{Y}) = N^2\ var\ (\bar{y})$$

y un estimador insesgado de la misma:

[8] $$\hat{var}\ (\hat{Y}) = N^2\ \hat{var}\ (\bar{y})$$

a2) Intervalos de confianza

Una vez realizadas las inferencias a la población mediante el uso de los estimadores anteriores es necesario dar una medida de la precisión de las estimaciones realizadas. Lo interesante de la teoría del muestreo es que observando **una sola muestra** de todas las posibles, se puede dar esa medida de precisión. Generalmente esta medida de la precisión se brinda construyendo un intervalo de confianza dentro del cual es "casi seguro" que estará el verdadero valor poblacional. "Casi seguro" significa "con una alta probabilidad" (generalmente 90, 95 o 99%). Un intervalo de confianza para la media es del tipo: $\bar{y} \pm \epsilon$ donde ϵ se denomina "error".

Para el M.A.S. se demuestra que, aproximadamente:

[9] $$\epsilon = t\ \sqrt{\hat{var}\ (\bar{y})}$$

donde t es el valor donde la distribución t-Student con n-1 grados de libertad, acumula el

$1-\alpha/2$ de probabilidad siendo $\alpha = 1$ - (probabilidad deseada para el intervalo). Por ejemplo si se quiere un intervalo al 95%, $\alpha = 0.05$.

Para tamaños de muestra mayores a 50 (lo usual en muestreos agrícolas) puede sustituírse el valor t anterior por el punto donde la distribución normal estándar acumula el $\alpha/2$ de probabilidad. En este caso los valores son:

- Para un intervalo de confianza al 90%, t=1.645
- Para un intervalo de confianza al 95%, t=1.96
- Para un intervalo de confianza al 99%, t=2.576

Si se está estimando un total poblacional, en lugar de la estimación de la varianza de la media en [9] se pondrá la estimación [8].

a3) Determinación del tamaño de muestra.

La primera pregunta que surge al planificar una encuesta es ¿cuán grande debe ser la muestra?. En el caso del M.A.S. la respuesta es sencilla. Hay dos posibles aproximaciones al tema:

i) fijando la precisión en términos absolutos;
ii) fijándola en términos relativos.

En el primer caso, se fija de antemano el nivel de error permitido (ϵ) para un determinado nivel de confianza ($1-\alpha$) y se despeja **n** en [9] (sustituyendo la varianza estimada por la poblacional [5]) si se está estimando μ ó en la fórmula correspondiente si se está estimando un total. Ello conduce a los siguientes resultados:

i.1) Para estimar μ con un error menor o igual a ϵ al $(1-\alpha)\%$ de confianza:

[10] $$n = \frac{N t^2 \sigma^2}{N \epsilon^2 + t^2 \sigma^2}$$

i.2) Para estimar el total $N\mu$ en las mismas condiciones:

[11] $$n = \frac{N^2 t^2 \sigma^2}{\epsilon^2 + N t^2 \sigma^2}$$

ii) Para estimar μ ó $N\mu$ con un error relativo $e = \frac{\bar{y} - \mu}{\mu}$ al $(1-\alpha)\%$:

[12] $$n = \frac{N t^2 \sigma^2}{N e^2 + t^2 \gamma^2}$$

donde γ es el coeficiente de variación poblacional: $\gamma = \sigma/\mu$

En la aplicación de las fórmulas anteriores hay dos problemas a destacar:

1. Es necesario conocer σ^2 que es un valor poblacional. No será nunca posible conocerla exactamente. Se utilizan aproximaciones dadas o bien por algún censo reciente ó por una encuesta previa o de datos históricos. El caso particular del muestreo para estimar proporciones se presenta más adelante.

2. En general en todas las encuestas se investigan, y se desea cuidar la precisión de varias variables. Como cada una de ellas tendrá varianzas diferentes, los tamaños de muestra necesarios serán distintos según la variable. En estos casos o bien se utiliza una variable altamente correlacionada con las variables de interés (área cultivada por ejemplo en investigaciones de producción de cultivos) o hay que arribar a algún tipo de compromiso entre los distintos tamaños posibles.

b) <u>Estimación de proporciones.</u>

b1) Estimadores

En el caso que la variable de interés (y) sólo pueda tomar dos valores (0 y 1), entonces su media no es más que la proporción de individuos con y=1. Es decir la proporción con la característica de interés. Por ejemplo, preguntas tales como si el productor vive o no vive en la explotación; si la explotación tiene o no riego; si los cultivos se fertilizan o no; si se usan pesticidas, etc. conducen a variables de este tipo (también llamadas variables dicotómicas). En este caso μ será la proporción de unidades de la población con la característica (ya que al sumar sobre todas las unidades los únicos que cuentan son los "unos") y lo mismo es válido para la media muestral. Si se denomina por P la proporción de individuos con la característica en la población y por p dicha proporción en la muestra y por Q y q respectivamente las proporciones de quienes no tienen la característica (obviamente P+Q=p+q =1)
puede verse muy fácilmente (sustituyendo en [3], [4], [5] y [6]) que:

[13] $$\sigma^2 = \frac{N}{N-1} PQ$$

[14] $$s^2 = \frac{n}{n-1} pq$$

y todos los resultados anteriores valen teniendo en cuenta, que en este caso particular: $\bar{y} = p$ y $n - \bar{y} = q$ siendo, por tanto p y q los estimadores de P y Q respectivamente. De manera que:

$$var\ (p) = \left(\frac{N-n}{N-1}\right) \frac{P\,Q}{n} \qquad [15]$$

$$\hat{var}\ (p) = \left(\frac{N-n}{N}\right) \frac{p\,q}{n-1} \qquad [16]$$

b2) Intervalos de confianza

Al tratarse de proporciones, se sustituirá $\hat{var}\ (\bar{y})$ en [9] por la fórmula [16]. Se obtendrá así un intervalo de confianza para P. En cuanto a la aplicación de la aproximación Normal, en este caso no es tan directa como en el caso general. La validez del uso de los puntos de la distribución Normal Estándar depende del verdadero valor de P y del tamaño de la muestra. Obviamente P es desconocido pero puede aproximarse por p para validar el uso de la distribución Normal. La tabla siguiente presenta criterios operativos para decidir si la aproximación normal es válida:

Si p vale:	n debe ser mayor o igual a:
0.5	30
0.4	50
0.3	80
0.2	200
0.1	600
0.05	1400

b3) Determinación del tamaño de muestra.

Se aplicará [10] o [12] según se quiera un error absoluto o relativo para P.

En el caso de las proporciones, el mayor valor que puede adoptar la varianza poblacional (σ^2) es 0.25 lo que ocurre cuando la población se divide "por mitades": P=Q=0.5. Por lo tanto si se tiene alguna idea de una cota superior para P, entonces puede acotarse σ^2 = PQ, cuando no hay información previa sobre una cota superior para P (ó para Q) se calcula n para la "peor" situación, es decir $\sigma^2 = 0.25$.

c) <u>Estimación de una razón.</u>

c1) Estimadores

Supongamos que en diversas parcelas se recoge información objetiva sobre la producción de algodón (y) y el área cosechada de algodón (x) a fin de estimar el

rendimiento de algodón y/x (por ejemplo si la producción se registra en kgs. y el área en hectáreas, se desea conocer el rendimiento en kilogramos por hectárea. Este rendimiento (r = y/x) es una razón. Para estimar esta razón usualmente se utiliza el siguiente estimador:

[17] $$\hat{r} = \frac{\bar{y}}{\bar{x}}$$

Como tanto el numerador como el denominador son variables aleatorias (toman diferentes valores según la muestra extraída) estos estimadores **no son insesgados bajo un diseño de M.A.S.**

Otro ejemplo de estimadores de razón es cuando se utiliza **información auxiliar**. Supongamos que de un censo reciente se sabe que el número de cabezas de ganado (X) en el momento del censo en determinada región geográfica era de 200,000. Se extrae una muestra de explotaciones de esa región y en las n explotaciones de la muestra se tiene que en el censo había, en promedio 23 cabezas de ganado ($\bar{x}$). Se realiza la encuesta y resulta que al momento de la encuesta, el número promedio de cabezas de ganado por explotación se estima en 24.5 ($\bar{y}$) entonces una estimación del total de cabezas de ganado en la región será:
213,047 = 200,000.(24.5/23). Simplemente se calcula la razón entre la estimación actual y el valor del censo y esa razón (en el ejemplo, un 6.5% de crecimiento) se aplica al total censal: $\hat{Y} = X\hat{r} = X\left(\frac{\bar{y}}{\bar{x}}\right)$.

El estimador de razón de la media de la población (con información auxiliar) es:

[18] $$\bar{y}_r = \hat{r}\,\mu_x$$

donde μ_x es la media verdadera de la variable auxiliar (en nuestro caso el promedio censal de cabezas de ganado por explotación).

Como el estimador [18] es sesgado, en lugar de la varianza el estadístico correcto para medir la dispersión será el Error Cuadrático Medio (ECM) definido como la varianza más el sesgo al cuadrado. En el caso del estimador de razón el sesgo disminuye rápidamente al aumentar el tamaño de la muestra y para muestras moderadas ya resulta pequeño con respecto a la varianza, de manera que una aproximación al ECM la da la varianza del estimador:

[19] $$ECM(\bar{y}_r) \approx var\,(\bar{y}_r) \approx \left(\frac{N-n}{N}\right)\frac{\sigma^2{}_r}{n}$$

siendo σ^2_r : $\sigma^2_r = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - R x_i)^2$ donde R es la razón en la población (desconocida a partir de la muestra, en nuestro ejemplo anterior sería el total de cabezas de ganado en toda la región en el momento de la encuesta (valor desconocido) dividido por 200,000). Un estimador de [19] es:

[20] $$\hat{var}(\bar{y}_r) = \left(\frac{N-n}{N}\right) \frac{s^2_r}{n}$$

siendo s^2_r : $s^2_r = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{r}\, x_i)^2$

c2) Intervalos de confianza

Un intervalo aproximado al 100 (1-α)% de confianza es:

[21] $$\bar{y}_r \pm t_{n-1}(\alpha/2) \sqrt{\hat{var}(\bar{y}_r)}$$

donde $t_{n-1}(\alpha/2)$ es el valor de la distribución t- Student con (n-1) grados de libertad para el punto (1-α/2). Con n mayor a 50 el valor t puede sustituirse por los resultantes de la aplicación de la aproximación Normal como se vio en a2).

Si se está estimando un total, como el estimador del total es $\hat{Y} = N\bar{y}_r$ el intervalo para el total se computa como:

[22] $$N\bar{y}_r \pm t_{n-1}(\alpha/2) \sqrt{N^2\, \hat{var}(\bar{y}_r)}$$

c3) Determinación del tamaño de muestra.

Para determinar el tamaño de muestra se procederá como fue explicado en a3) utilizando ahora σ^2_r en lugar de σ^2.

2.2. <u>Muestreo Aleatorio Estratificado</u>.

2.2.1. <u>Definición.</u>

De las fórmulas presentadas para el M.A.S. resulta claro que cuanto mayor es la varianza poblacional se requieren tamaños mayores de muestra para mantener un nivel de

precisión predeterminado o equivalentemente el nivel de precisión es menor para igual n. Estas consideraciones llevan a pensar que si la población pudiera ser agrupada en subconjuntos internamente homogéneos con respecto a la(s) variable(s) de interés y heterogéneos entre ellos y se tomaran muestras independientes dentro de cada grupo quizás la suma de los diferentes tamaños fuera menor que el tamaño requerido por una M.A.S. para el mismo nivel de precisión total. Esto es particularmente cierto en poblaciones muy asimétricas, como son generalmente las que se encuentran en las estadísticas agrícolas. Pocas explotaciones grandes concentran una alta proporción de la tierra mientras muchas explotaciones chicas acumulan un pequeño porcentaje del total. Entonces si estamos interesados en estimar variables relacionadas con el área (por ejemplo: áreas con determinados cultivos, producción, rendimientos, etc.) parece razonable tomar una fracción de muestreo alta entre las explotaciones grandes (es decir para las que contribuyen con mucho al total) a fin de estimar estos subtotales con un error pequeño, y una fracción de muestreo pequeña para el subconjunto de explotaciones chicas (si el error aquí es grande no importa porque contribuyen con una proporción mínima al total). Esa es la idea que está detrás del Muestreo Aleatorio Estratificado (M.A.E.),

La población de N elementos se divide en L grupos llamados **estratos**. Se toman luego M.A.S. dentro de cada estrato. Se computan las estimaciones para cada estrato y luego se combinan ponderándolas adecuadamente para formar la estimación global.

2.2.2. <u>Estimadores, intervalos de confianza y tamaños de muestra</u>.

a1) Estimadores

Llamemos N_h al tamaño de la población del estrato h y n_h al tamaño de la muestra (obtenida mediante M.A.S.) del mismo estrato. $N = \sum_{h=1}^{L} N_h$ y $n = \sum_{h=1}^{L} n_h$.

Anotemos por y_{hi} al valor de la variable y, en la i-ésima unidad del estrato h.

Como dentro de cada estrato la selección es con M.A.S. la media del estrato h (μ_h) es insesgadamente estimada por la media muestral del estrato: $\bar{y}_h = \frac{1}{n_h} \sum_{h=1}^{L} y_{hi}$. Análogamente el total del estrato h: $N_h\mu_h$ es estimado por $N_h\, \bar{y}_h$.

Ahora bien, el total poblacional $N\mu$ es la suma de los totales de cada estrato: $N\mu = \sum_h N_h\, \mu_h$ de donde $\mu = \sum_h \frac{N_h}{N}\, \mu_h$ que no es más que el promedio ponderado de las medias por estrato. De lo anterior surge que los estimadores insesgados de la media y

del total de Y para toda la población son:

[23] $$\bar{y}_{est} = \sum_h \frac{N_h}{N} \bar{y}_h$$

[24] $$\hat{Y}_{est} = \sum_h N_h \bar{y}_h$$

Dentro de cada estrato se definen s_h^2 y σ_h^2 como en [3] y [4] y la varianza del estimador y un estimador de esta varianza resultan:

[25] $$var\ (\bar{y}_{est}) = \sum_h \frac{N_h}{N^2} (N_h - n_h) \frac{\sigma^2_h}{n_h}$$

[26] $$\hat{var}\ (\bar{y}_{est}) = \sum_h \frac{N_h}{N^2} (N_h - n_h) \frac{s^2_h}{n_h}$$

Análogamente se obtienen las correspondientes fórmulas para la varianza del estimador $N\bar{y}_{est}$ del total Y. Los desarrollos realizados en el caso del M.A.S. para el muestreo de proporciones y para los estimadores de razón valen aquí dentro de cada estrato y luego los totales correspondientes se computarán agregando los totales por estrato.

a2) Intervalos de confianza

Lo ya desarrollado para el M.A.S. vale aquí sustituyendo $\bar{y}$ por $\bar{y}_{est}$.
En el caso del M.A.E., la aplicación de la aproximación normal no es tan directa como en el caso del M.A.S porque juega no sólo el tamaño de la muestra total sino también el de las muestras por estrato. Sin embargo para los tamaños con que usualmente se trabaja en muestreos agrícolas, en general la aproximación es correcta.

a3) Determinación del tamaño de muestra y su distribución entre estratos.

En el M.A.E. se debe determinar no sólo el tamaño total de la muestra (n) sino también los n_h. Hay diferentes procedimientos para distribuir la muestra entre los estratos. Aquí consideraremos los dos más utilizados: la distribución proporcional y la distribución óptima.

a3.1) Asignación proporcional.

En la asignación proporcional, simplemente se distribuye **n** entre los estratos proporcionalmente a los tamaños de la población en cada estrato, es decir:

[27] $$n_h = n \frac{N_h}{N}$$

en este forma de distribuir la muestra entre los estratos sólo se tienen en cuenta los tamaños de cada estrato. En general en los muestreos agrícolas, no respeta el criterio expuesto en 2.2.1. en el sentido de tomar tasas altas de muestreo para aquellos estratos que contribuyen más a los totales que se quieren estimar. Esta forma de asignación de la muestra, es útil por ejemplo, cuando se están estimando variables altamente correlacionadas con el tamaño de los estratos (población rural por ejemplo, donde la mayor parte de la población reside en las explotaciones pequeñas o, en muchos países no ganaderos las existencias de ganado que se concentran en explotaciones chicas o sin tierra).

A fin de determinar n, si se va a distribuir con asignación proporcional procederemos de la siguiente manera: si se desea un error total de ϵ al $(1-\alpha)\%$ de confianza para la estimación de la media poblacional de determinada variable, plantearemos (en forma análoga a lo hecho en el caso del M.A.S.): $\epsilon = t \sqrt{var\ (\overline{y}_{est})}$ donde la fórmula para la varianza del estimador está dada por [25] . Ahora bien, si en la fórmula [25] sustituimos n_h por el valor dado en [27] se obtiene:

[28] $$n = \frac{N t^2}{N^2 \epsilon^2 + (\sum_{h=1}^{L} N_h \sigma^2_h)\ t^2}$$

donde t surgirá de la tabla Normal. Este uso de la tabla Normal es correcto si la variable de interés se distribuye Normal o aproximadamente Normal en la población. El uso generalizado de los valores normales proviene de la aproximación Normal para muestras grandes.

Nótese también que en este caso, es necesario tener alguna idea previa de L varianzas poblacionales (las σ^2_h). Las observaciones realizadas para el M.A.S. valen aquí.

a3.2) Asignación óptima.

Como se hacía notar anteriormente, el problema de la asignación proporcional es que no tiene en cuenta la contribución de cada estrato al total que se quiere estimar. Más aún, como dentro de cada estrato las muestras se seleccionan con M.A.S. es razonable que si el estrato es más variable internamente el tamaño de muestra allí deba ser mayor. La llamada asignación óptima (o asignación de Neyman) logra este propósito. La asignación es óptima en el sentido que minimiza la varianza del estimador para un costo de relevamiento dado (con algunos supuestos sobre la forma de la función de costo) o minimiza el costo para una varianza dada. La asignación óptima es la siguiente:

[29] $$n_h = n \frac{N_h \sigma_h}{\sum_{h=1}^{L} N_h \sigma_h}$$

El tamaño total de muestra para un error de ϵ predeterminado para la media de la variable de interés al $(1-\alpha)\%$ de confianza se demuestra que se obtiene de la siguiente manera:

[30] $$n = \frac{t^2 \left(\sum_{h=1}^{L} N_h \sigma_h\right)^2}{N^2 \epsilon^2 + t^2 \sum_{h=1}^{L} N_h \sigma^2{}_h}$$

donde t es, como antes, el valor correspondiente de la tabla Normal para una probabilidad acumulada de $1 - \alpha/2$.

La observación hecha anteriormente sobre la necesidad de conocer aproximadamente las varianzas poblacionales también es válida aquí.

2.2.3. Uso de la estratificación en muestreos agrícolas.

Por las características señaladas de las poblaciones que usualmente aparece en los muestreos agrícolas, la estratificación de las mismas juega un papel fundamental. Más adelante consideraremos muestreos en varias etapas, muy utilizados en las estadísticas agrícolas. Generalmente en una o en varias de las etapas la estratificación de las unidades aumenta considerablemente la precisión de las estimaciones para una muestra de tamaño dado. En los diseños de muestreo de áreas, como ya se señaló en 3.2. los segmentos generalmente se estratifican de acuerdo a las características de uso del suelo ya que las variables agrícolas serán más homogéneas en su comportamiento cuando se refieren a suelos similares.

Algunos problemas interesantes del muestreo estratificado y que aparecen inmediatamente que uno se plantea un diseño de este tipo, en particular en los muestreos agrícolas (pero estos problemas son generales) son: a) la determinación de la o las variables de estratificación (el criterio de estratificación); b) la determinación del número de estratos (L); c) la determinación de los puntos de corte de los estratos; d) el tamaño final de muestra para controlar varias variables.

En general no hay una solución única para los puntos anteriores, entre otras cosas porque depende de las variables a investigar y de los ámbitos en que se desee dar la

información. Por ejemplo sería deseable que una encuesta agrícola pudiera brindar información desagregada en los mismos tramos de tamaño en que se clasificó la información censal para las mismas variables: menores de 0.5 hectáreas, de 0.5 a 1 hectárea, de 1 a 2 etc. etc.. En ese caso habría que estudiar si no es conveniente estratificar la muestra de acuerdo al área total en esos tramos aunque no sea la solución teóricamente mejor. Otra solución es post estratificar (ver punto 2.2.4). En la práctica o bien se selecciona una única variable de estratificación que esté altamente correlacionada con las variables investigadas (por ejemplo el área agrícola para las variables sobre área sembrada, cosechada, producción, rendimientos, pronósticos de siembras y cosechas; o número de árboles para una encuesta sobre producción de cultivos permanentes), o bien se analiza el comportamiento de diferentes tamaños de muestra bajo una misma asignación teniendo en cuenta diferentes variables y luego se llega a un compromiso entre ellas o se utiliza "análisis de clusters" para formar grupos homogéneos de unidades considerando diferentes variables y luego se toma cada conglomerado así formado como un estrato. Más adelante consideraremos algunos casos de encuestas agrícolas y se verán con más detalle estas aplicaciones.

Determinado el número de estratos, los puntos de corte pueden aproximarse dividiendo la población en un número "grande" de clases de acuerdo con valores de la variable de interés y acumulando la raíz cuadrada de las frecuencias absolutas observadas en cada clase y luego elegir los puntos de corte de tal manera que formen intervalos iguales en la escala de las raíces acumuladas. Por ejemplo, supongamos que 10,000 explotaciones agrícolas de la población se desean estratificar en 5 estratos de acuerdo con la variable "area agrícola". En primer lugar se clasifican las 10,000 explotaciones en intervalos (más bien "pequeños") de área agrícola. Luego se calcula la raíz cuadrada de la cantidad de explotaciones en cada clase y se computa el acumulado de dicha raíz; por fin, el acumulado final se divide por 5 y se tienen los puntos de corte aproximados. Los siguientes datos ilustran el método:

Intervalo de área agrícola (has)	Frecuencia observada (marco)	Raíz cuadrada de la frecuencia	Acumulado de la raíz cuadrada
Hasta 0.5	2,500	50	50
De 0.5 hasta 1	1,800	42.426	92.426
De 1 hasta 2	1,200	34.641	127.067
De 2 hasta 3	950	30.822	157.889
De 3 hasta 4	900	30	187.889
De 4 hasta 5	800	28.284	216.173
De 5 hasta 7	920	30.332	246.505
De 7 hasta 10	480	21.909	268.414
De 10 hasta 15	200	14.142	282.556
De 15 hasta 20	120	10.955	293.511
De 20 hasta 25	85	9.22	302.731
De 25 hasta 30	30	5.477	308.208
De 30 y más	15	3.873	312.081
TOTAL	10,000		

Con los datos anteriores los puntos de corte de los estratos serían aproximadamente:

- Estrato 1: Hasta 0.5 has. de agricultura
- Estrato 2: De 0.5 has. y hasta 2 has. de agricultura
- Estrato 3: De 2 has. y hasta 4 has. de agricultura
- Estrato 4: De 4 has. y hasta 7 has. de agricultura
- Estrato 5: De 7 has. de agricultura y más.

Ello es así porque 312 dividido por 5 (número predefinido de estratos) es 62.4 y el extremo de clase más cercano es 0.5 has (acumulado = 50), luego 62.4 por 2 (segundo punto en la escala de las raíces cuadradas acumuladas) es 124.8 y el valor acumulado más cercano es 127.067 correspondiente a 2 has. y así sucesivamente. Esta forma de construir los estratos ha demostrado ser altamente eficiente en la mayoría de los casos.

Tanto el número de estratos, como los puntos de corte así como juegos de variables de estratificación, hoy en día pueden simularse rápidamente por computadora y luego comparar la eficiencia lograda bajo cada alternativa para elegir la más conveniente.

2.2.4. Post-estratificación

A veces se deben clasificar las unidades de una muestra en estratos y usar un estimador estratificado aun cuando la muestra haya sido seleccionada por M.A.S.. Por ejemplo uno puede tomar una M.A.S. de explotaciones agropecuarios y luego desea clasificarlas en tres tramos de tamaño e inferir para cada estrato (post-estrato pues fueron formados luego de extraer la muestra) y para el total. Lo mismo puede suceder dentro de los estratos de un muestreo estratificado, por ejemplo dentro de la clase de 7 has. agrícolas y más del ejemplo anterior, luego de extraer la muestra estamos interesados en saber qué pasa con las de 7 a 10 has, de 10 a 20 has, de 20 a 30 has y de 30 y más. Como la muestra dentro del estrato de 7 y más fue extraída mediante M.A.S. estamos en una situación como la anterior.

Supóngase que de una población de N unidades se extrajeron n con M.A.S. y luego de observada la muestra se decide post-estratificar la población en L grupos. La principal diferencia con el M.A.E. es que ahora los n_h es decir la cantidad de unidades que cayeron en la clase h no son fijos de antemano sino que son variables aleatorias (serán diferentes según la muestra extraída). Las estimaciones de los parámetros de interés se hacen como si fuera M.A.E. El problema se presenta con el cálculo de las varianzas. Intuitivamente es claro que en la clase h esperaremos que hayan aproximadamente $n\frac{N_h}{N}$ unidades de la muestra (es decir si, por ejemplo el grupo h tiene el 20% de las unidades de la población, esperaremos que una M.A.S. extraída de toda la población contenga en dicha grupo aproximadamente el 20% de las unidades). Esta intuición es ratificada por la teoría. En efecto, se demuestra que el valor esperado de n_h es precisamente $n\frac{N_h}{N}$. Por tanto, la muestra tiende a asignarse en forma aproximadamente proporcional. La varianza del estimador de la media [23] es ahora mayor que la varianza que se hubiera obtenido aplicando M.A.E. proporcional. Sin embargo para <u>estimar</u> la varianza es conveniente utilizar directamente el estimador [26].

A fin de obtener las estimaciones es necesario conocer las cantidades $\frac{N_h}{N}$, estas cantidades, en este caso, de post estratificación muchas veces son desconocidas y se estimarán utilizando una muestra diseñada para ello (muestreo doble).

2.3. Muestreo por conglomerados.

2.3.1. Definición.

Al igual que en el muestreo estratificado, la población es dividida en grupos ("conglomerados"). Cada conglomerado constituye una unidad de muestreo y está formada por unidades finales. Los conglomerados se llaman **unidades primarias** y las unidades finales, **unidades secundarias**. Por ejemplo los sectores de empadronamiento de un censo agropecuario pueden ser las unidades primarias y las explotaciones dentro de ellos las secundarias. En un muestreo de áreas los segmentos pueden ser unidades primarias y los tramos las secundarias. En una cuadrícula, cada cuadrado de la cuadrícula puede ser un conglomerado (unidad primaria) y las explotaciones con parcelas total o parcialmente incluidas en la cuadrícula, unidades secundarias.

En el muestreo por conglomerados se seleccionan mediante algún diseño (M.A.S., M.A.E., u otro) unidades primarias y luego se enumeran **todas** las unidades secundarias dentro del conglomerado seleccionado.

Desde el punto de vista de su uso, el principio es el contrario al del muestreo estratificado: en aquél cuanto más homogéneos fueran los estratos internamente y más heterogéneos entre ellos, mejor; en el muestreo por conglomerados, cuanto más heterogéneo sea el conglomerado a su interior y más homogéneos sean entre ellos será más eficiente el diseño pues cada conglomerado con su heterogeneidad interna aportará mucha información (si un conglomerado fuera tan heterogéneo como toda la población "contaría toda la historia").

En realidad, como todas las unidades secundarias de los conglomerados seleccionados son encuestadas, el muestreo por conglomerados podría considerarse en los diseños ya vistos: el subíndice "i" se referiría al conglomerado y los valores de los y_i serían las sumas simples de los valores registrados en todas las unidades secundarias del conglomerado. Sin embargo hay dos razones por las cuales vale la pena considerar al muestreo por conglomerados como un caso especial:

- El tamaño del conglomerado puede servir como información auxiliar valiosa y puede ser usado para seleccionar los conglomerados son probabilidades desiguales (ppt) o para formar estimadores de razón.

- El tamaño y la forma de los conglomerados puede afectar la eficiencia.

Sea N el número de unidades primarias en la población y n el tamaño de la muestra de unidades primarias. Sea M_j el número de unidades secundarias en la j-ésima unidad primaria. El número total de unidades secundarias en la población será: $M = \sum_{j=1}^{N} M_j$. Sea

y_{jk} el valor de la variable de interés en la k-ésima unidad secundaria de la j-ésima primaria. El total de los valores de la variable en la j-esima unidad primaria los llamaremos y_j. Así el total poblacional no es más que $Y = \sum_{j=1}^{N} y_j$ y la media poblacional $\mu = Y/M$.

Distinguiremos dos casos: el primero, cuando las unidades primarias son elegidas mediante M.A.S. y el segundo cuando son elegidas con probabilidad proporcional al tamaño. No consideramos el caso de las unidades primarias elegidas mediante M.A.E. porque surge directamente del primero considerando las precisiones hechas para el muestreo estratificado.

2.3.2. Unidades primarias seleccionadas mediante M.A.S.

En este caso, un estimador insesgado de la media es:

[31] $$\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$$

y el del total $N\bar{y}$.

Su varianza y el estimador de la varianza, son respectivamente:

[32] $$var\ (\ \bar{y}\ ) = \frac{N-n}{N}\ \frac{\sigma^2{}_u}{n}$$

[33] $$v\hat{a}r\ (\ \bar{y}\ ) = \frac{N-n}{N}\ \frac{s^2{}_u}{n}$$

siendo σ^2_u y s^2_u las varianzas poblacional y muestral entre los totales de las unidades primarias. Las varianzas del total estimado y su estimador se obtienen simplemente multiplicando [32] y [33] por N^2. Nótese la diferencia con el muestreo estratificado: allí aparecen las varianzas <u>dentro</u> de los estratos y aquí sólo aparecen las varianzas <u>entre</u> los conglomerados.

Intervalos de confianza y tamaño de muestra de unidades primarias surgen como en el M.A.S. ya considerado.

En el caso que el total de las unidades primarias (y_j) está altamente correlacionado con el tamaño de la correspondiente unidad primaria (M_j) (por ejemplo el número de personas que viven en explotaciones agrícolas puede estar altamente correlacionado con la cantidad de explotaciones dentro de los conglomerados, porque el promedio de personas por explotación no es muy variable),un estimador de razón basado en el tamaño es generalmente muy eficiente.

En ese caso el estimador de razón del total de la población es:

$$[34] \qquad \hat{Y}_r = \hat{r}\ M = \frac{\sum_{j=1}^{n} y_j}{\sum_{j=1}^{n} M_j} M$$

En el ejemplo anterior es el total de personas viviendo en explotaciones en los conglomerados sorteados dividido por el total de explotaciones en los conglomerados sorteados (es decir el promedio por explotación) multiplicado por el total de explotaciones en la población. Este estimador es sesgado pero su sesgo disminuye rápidamente a medida que aumenta n.

Como estimador de la varianza del total puede utilizarse:

$$[35] \qquad \hat{var}\ (\ \hat{Y}_r\ ) = \frac{N\,(N-n)}{n\ (n-1)} \sum_{j=1}^{n} (y_j - \hat{r}\ M_j)^2$$

Para estimar la media por unidad primaria el estimador de razón es: $\bar{y}_r = \frac{\hat{Y}}{N}$. Claramente, el estimador de la media poblacional por unidad secundaria es $\hat{r}$.

2.3.3. Unidades primarias seleccionadas con probabilidad proporcional al tamaño.

Si las unidades primarias son seleccionados con probabilidad proporcional a su tamaño (ppt) es decir: Probabilidad de seleccionar la unidad j es: $p_j = M_j / M$, y con reemplazo, entonces un estimador insesgado del total es:

$$[36] \qquad \hat{Y}_{ppt} = \frac{M}{n} \sum_{j=1}^{n} \frac{y_j}{M_j}$$

donde cada observación se computa tantas veces como su unidad primaria fue seleccionada.

Su varianza y el estimador de la varianza son respectivamente:

$$[37] \qquad var\ (\hat{Y}_{ppt}\ ) = \frac{M}{n} \sum_{j=1}^{N} M_j\ (\bar{y}_j - \mu)^2$$

$$[38] \qquad \hat{var}\ (\hat{Y}_{ppt}\ ) = \frac{M^2}{n\ (n-1)} \sum_{j=1}^{n} (\bar{y}_j - \hat{\mu}_{ppt}\ )^2$$

donde $\bar{y}_j = \frac{y_j}{M_j}$ es decir el promedio de la variable de interés en el j-ésimo conglomerado y $\hat{\mu}_{ppt} = \frac{\hat{Y}_{ppt}}{M}$ es decir, la media estimada por unidad secundaria. En muchos casos, la selección mediante este diseño resulta más eficiente que mediante M.A.S.

2.4. Muestreo sistemático.

En el muestreo sistemático se forman conglomerados de unidades secundarias espaciadas de manera sistemática a través de la población. El diseño sistemático más común es el que consiste en seleccionar al azar una unidad de la población y luego incorporar a la muestra una unidad cada tantas hasta completar las n. Por ejemplo supongamos una población de 10,000 explotaciones agropecuarias de donde se quiere seleccionar una muestra sistemática de 500. Se elige una unidad al azar entre las primeras 20 (10,000/500) y luego una de cada 20. El muestreo sistemático es ventajoso cuando las unidades están ordenadas de tal manera que las unidades cercanas en la lista son "similares" en cuanto al comportamiento de la variable de interés y esta similitud desaparece a medida que las unidades se apartan en la lista. Por ejemplo, cuando se tienen explotaciones agrícolas listadas por número de cuestionario censal, en general, las cercanas en la lista son cercanas en el campo (ya que el encuestador siguió posiblemente el orden de los formularios que se le entregaron numerados en forma correlativa) y explotaciones cercanas tienen comportamientos similares (tipo de cultivos, tipo de prácticas agrícolas, etc.) que hacen que las explotaciones sean parecidas a sus vecinas y diferentes a las lejanas. El muestreo sistemático puede verse como un muestreo por conglomerados (en el caso del ejemplo, la población de 10,000 unidades se divide en 20 conglomerados de 500 unidades cada uno) de donde se selecciona al azar uno. Este procedimiento tiene el inconveniente que no puede obtenerse un estimador insesgado de la varianza. En la práctica es común que se utilice [6] como si la muestra hubiera sido seleccionada por M.A.S. Esta forma de proceder, en general **sobreestima** la verdadera varianza. Debe tenerse cuidado, también en que si N/n no es un número entero, la probabilidad de selección de las unidades puede no ser la misma (por ejemplo si en lugar de 10,000 unidades en el ejemplo anterior tuviéramos 10,003 y elegimos una de cada 20, las 3 últimas unidades tendrían probabilidad cero de ser elegidas). Esto se soluciona ordenándolas de manera "circular" (a la unidad 1 le corresponde el número 10,004, etc y elegir una de cada 21 "dando la vuelta en el círculo").

2.5. Muestreo en varias etapas.

Son los diseños más utilizados en las encuestas de estructura y de producción agrícola y combinan los diseños anteriormente estudiados. Los diseños en varias etapas tienen la ventaja del abaratamiento de los costos tanto de confección de marcos (es necesario listar sólo las unidades finales dentro de las seleccionadas en etapas anteriores) como de relevamiento ya que las unidades de última etapa seleccionadas se agrupan naturalmente

haciendo disminuir el costo de traslado entre unidades finales. Tienen el inconveniente que, en general son diseños complejos y el cálculo de lo errores así como algunos análisis de la información no son directos.

2.5.1. Definición.

La población se divide en grupos. Estos grupos (unidades primarias) contienen unidades secundarias. Si luego de seleccionar una muestra de unidades primarias (UPM) en lugar de encuestar a todas las unidades secundarias de las UPM seleccionadas (como en el muestreo por conglomerados) se extrae una muestra de unidades secundarias (USM) dentro de las unidades primarias sorteadas se dice que se tiene un muestreo en dos etapas. Si luego, una muestra de unidades terciarias es seleccionada dentro de las secundarias se dice que se tiene un muestreo en tres etapas y así sucesivamente. Por ejemplo si la población de explotaciones agropecuarias está dividida en los sectores de empadronamiento censal y se extrae una muestra de sectores y luego dentro de los seleccionados se extrae una muestra de explotaciones tendremos un muestreo en dos etapas. Si dentro de las explotaciones sorteadas se sortean, a su vez, parcelas se tendrá una muestra en tres etapas.

La contribución a la varianza muestral total tiene ahora varios componentes. Por ejemplo en dos etapas tendremos la varianza entre UPM y la varianza entre USM dentro de UPMs.

La notación usada para los valores poblacionales en el caso del muestreo por conglomerados es válida aquí.

Consideraremos sólo el caso de dos etapas y distinguiremos dos situaciones: la primera, cuando se utiliza M.A.S. en ambas etapas y la segunda cuando las UPM son elegidas con probabilidad proporcional al tamaño.

2.5.2. M.A.S. en ambas etapas.

En la primera etapa se eligen n UPM de las N de la población mediante M.A.S. y, en la segunda etapa se eligen también mediante M.A.S. m_i unidades secundarias de la i-ésima UPM seleccionada.

En este caso, un estimador insesgado de la media es:

[39] $$\bar{y} = \frac{1}{n}\sum_{j=1}^{n} \hat{y}_j$$

siendo $\hat{y}_j = M_j \bar{y}_j$ es decir el total estimado de la j-ésima UPM. Un estimador insesgado del total es $N\bar{y}$.

Su varianza y el estimador de la varianza, son respectivamente:

$$[40] \quad var(\bar{y}) = \frac{N-n}{N}\frac{\sigma^2_u}{n} + \frac{1}{Nn}\sum_{j=1}^{N} M_j (M_j - m_j)\frac{\sigma^2_j}{m_j}$$

$$[41] \quad \hat{var}(\bar{y}) = \frac{N-n}{N}\frac{s^2_u}{n} + \frac{1}{Nn}\sum_{j=1}^{n} M_j (M_j - m_j)\frac{s^2_j}{m_j}$$

siendo σ^2_u y s^2_u las varianzas poblacional y muestral entre los totales de las unidades primarias y σ^2_j y s^2_j las respectivas varianzas entre las unidades secundarias. Es interesante observar que los primeros términos de [40] y [41] son los mismos que [32] y [33] que es la contribución de la varianza **entre UPM** mientras que el segundo término se agrega por el hecho de **submuestrear** dentro de las UPM, es decir es la contribución de la varianza **entre USM.** Las varianzas del total estimado y su estimador se obtienen simplemente multiplicando [40] y [41] por N^2.

De manera similar a lo ya expresado para el muestreo por conglomerados, en el caso que el total de las unidades primarias (y_j) esté altamente correlacionado con el tamaño de la correspondiente unidad primaria (M_j), un estimador de razón basado en el tamaño es generalmente muy eficiente.

En ese caso y con este diseño en dos etapas, el estimador de razón del total de la población es:

$$[42] \quad \hat{Y}_r = \hat{r}\, M = \frac{\sum_{j=1}^{n} \hat{y}_j}{\sum_{j=1}^{n} M_j} M$$

Como estimador de la varianza del total puede utilizarse:

$$[43] \quad \hat{var}(\hat{Y}_r) = \frac{N(N-n)}{n(n-1)}\sum_{j=1}^{n} (y_j - \hat{r}\, M_j)^2 + \frac{N}{n}\sum_{j=1}^{n} M_j (M_j - m_j)\frac{s^2_j}{m_j}$$

Para estimar la media por unidad primaria el estimador de razón es: $\bar{y}_r = \frac{\hat{Y}}{N}$. Obviamente el estimador de la media poblacional por unidad secundaria es $\hat{r}$.

2.5.3. Unidades primarias seleccionadas con probabilidad proporcional al tamaño.

Si las unidades primarias son seleccionados con probabilidad proporcional a su tamaño (ppt) es decir: Probabilidad de seleccionar la unidad j es: $p_j = M_j / M$, y con reemplazo y una muestra de USM es seleccionada independientemente con M.A.S. (obsérvese que una misma USM podría aparecer más de una vez por la selección con reemplazo de UPMs), entonces un estimador insesgado del total es:

[44] $$\hat{Y}_{ppt} = \frac{M}{n} \sum_{j=1}^{n} \frac{\hat{y}_j}{M_j}$$

Su varianza y el estimador de la varianza son respectivamente:

$$var\ (\hat{Y}_{ppt}) = \frac{M}{n} \sum_{j=1}^{N} M_j\ (\mu_j - \mu)^2 + \frac{M}{n} \sum_{j=1}^{N} [\ \frac{(\ M_j - m_j\ )}{m_j\ (M_j - 1\ )} \sum_{k=1}^{M_j} (y_{ij} - \mu_j\ )^2]$$

[45]

$$v\hat{a}r\ (\hat{Y}_{ppt}) = \frac{M^2}{n\ (n - 1)} \sum_{j=1}^{n} (\bar{y}_j - \hat{\mu}_p)^2$$

[46]

donde $\bar{y}_j$ es el promedio de la variable de interés dentro de la j-ésima unidad primaria de la muestra y $\hat{\mu}_p = \frac{\hat{Y}_{ppt}}{M}$ es decir, la media estimada por unidad secundaria.

3. ELECCIÓN DE UN DISEÑO DE MUESTREO.

La elección del diseño más apropiado de muestreo para conducir una encuesta agrícola de estructura o producción agropecuaria es una de las tareas más delicadas por su proyección y por los costos que implica su implementación.

Las características del país, los recursos disponibles (no solo en dinero sino también en personal capacitado), los niveles de precisión necesarios , las capacidades computacionales disponibles, la información de base, la situación coyuntural del sector

agrícola, etc. son elementos a tener en cuenta a la hora de decidir el diseño de muestreo a utilizar. Este diseño debería ser suficientemente simple para que pueda ser entendido y trabajado por el personal disponible. Cuanto más complejo es un diseño, más difícil resulta ajustarlo cuando surgen problemas de relevamiento (errores de marco, cuestionarios perdidos, no respuesta, etc.). Si el diseño elegido implica gastos que van más allá de los presupuestados debería también ser corregido, explicándole a los decisores las consecuencias que estos cambios realizados por razones presupuestales tendrán en las estimaciones obtenidas. La determinación del tamaño de la muestra debe ser, también cuidadosamente estudiada. Como ya se vio, los niveles de precisión requeridos juegan un papel fundamental en el tamaño final de la muestra. También es fundamental el nivel de desagregación que se requiere para los datos. Si se quieren estimaciones precisas para ámbitos geográficos reducidos (a nivel de provincia, departamento, distrito, municipio) será necesario tomar muestras mayores. El costo y el presupuesto disponible juega también un rol clave. Muy comúnmente el tamaño total de la muestra posible viene dado de antemano en función de los recursos disponibles y es tarea del estadístico diseñar la muestra más eficiente compatible con los recursos disponibles, indicar, cuáles son, a priori los niveles de precisión esperados para que quien debe decidir la ejecución de la encuesta tenga todos los elementos para ello.

En general las encuestas realizadas en base a diseños en varias etapas son de costo menor que las apoyadas en diseños uni-etápicos. Esto es así porque un componente importante del costo de relevamiento viene dado por el traslado del personal en el campo. Cuando se tiene un muestreo en una sola etapa, las unidades a encuestar aparecen diseminadas (en forma aleatoria) en el ámbito geográfico bajo estudio. Por su parte, si el diseño fue en varias etapas, las unidades de última etapa son cercanas entre sí porque pertenecen a una misma unidad de la etapa inmediata anterior (por ejemplo si seleccionamos segmentos dentro de una UPM de área, los segmentos elegidos estarán próximos en el terreno por pertenecer a la misma UPM), entonces el costo mayor es el del traslado hacia la correspondiente unidad mayor (en el caso de dos etapas hasta la UPM elegida) y luego el traslado entre unidades finales es mucho menos oneroso.

El costo de la construcción de marcos de lista, es también menor en los muestreos en varias etapas, porque es necesario construir listas de unidades finales de muestreo sólo para las unidades mayores seleccionadas. Por ejemplo en un muestreo en dos etapas, se construirá un padrón de USM sólo para las UPM seleccionadas.

Como contrapartida a las consideraciones anteriores, en general los muestreos en varias etapas requieren tamaños de muestra mayores que los uni-etápicos para un mismo nivel de precisión y esto es así porque la variabilidad interna de las UPM generalmente no es tan grande como la observada en la población.

En general los muestreos basados en marcos de área son más económicos en cuanto

a desplazamientos en el campo que los basados en marcos de lista. Sin embargo requieren mucho más trabajo cartográfico previo y más labor de gabinete en mediciones sobre planos y fotos aéreas.
Todos esos elementos deben tenerse particularmente en cuenta a la hora de decidir un diseño muestral porque la eficiencia del mismo depende de la adecuada resolución de la ecuación: precisión de las estimaciones versus costo de implementación.

IV. ALGUNOS EJEMPLOS DE DISEÑOS.

A continuación se presentan algunos ejemplos de diseños de muestreo agrícola usados en América Latina y el Caribe, tanto para encuestas continuas como para enumeración censal por muestreo. Se brinda una breve descripción de los objetivos de la investigación y las características del diseño.

1. Ejemplo de CENSO AGRÍCOLA CON ENUMERACIÓN POR MUESTREO.

El caso de Dominica.

Es poco común encontrar censos levantados mediante enumeración por muestreo en América Latina o el Caribe. Una revisión metodológica realizada por FAO (y en vías de publicación) sobre la ronda de censos agropecuarios cubiertos por el Programa CAM 2000 (es decir los levantados entre 1996 y 2005) no identifica ningún censo levantado exclusivamente por muestreo en la región. En efecto, de 114 países que tomaron censos agropecuarios en dicha década, 16 corresponden a América Latina o el Caribe y todos ellos fueron relevados por enumeración completa (contra 17 por enumeración completa y 5 por enumeración por muestreo en África, 19 de enumeración completa y 3 por muestreo en Asia, 25 por enumeración completa y 2 por muestreo en Europa).

Un caso típico de levantamiento de un censo agrícola de enumeración por muestreo lo constituye el Censo Agrícola de Dominica (Caribe) de 1995. La metodología usada fue una combinación de enumeración completa y enumeración por muestreo. La primera "comprendió la obtención de información de un pequeño número de variables claves para todas las explotaciones agrícolas existentes en el país en áreas no urbanas a través de un padrón". Se estableció que se consideraban explotaciones agrícolas las que operaban 0.01 de acre o más, o las que no llegando a dicho límite inferior, poseían al menos una cabeza de ganado o aves). Al mismo tiempo que los enumeradores preparaban este padrón, todas las explotaciones agrícolas de 5 acres y más de superficie total fueron enumeradas mediante un cuestionario censal. Por lo tanto se realizó una enumeración completa de las de 5 acres y más. Inmediatamente después de completado el padrón, se extrajo una Muestra Sistemática de una de cada cinco (tasa de muestreo del 20 %) explotaciones listadas con superficie comprendida entre 0.01 de acre y 5 acres. Las "sin tierra", que sólo tenían alguna cabeza de ganado o de aves, fueron

enumeradas a través del padrón.

2. Ejemplos de encuestas agrícolas basadas en marcos de área.

El caso de El Salvador.

De acuerdo con la información manejada (FAO: Métodos nacionales de compilación de estadísticas agropecuarias) desde 1976 se realizaron encuestas agropecuarias basadas en un marco muestral de áreas, para el cual se utilizaron cartas topográficas y fotografías aéreas que permitieron identificar segmentos con límites físicos reconocibles. La muestra estaba compuesta de 553 segmentos distribuidos por todo el . país según muestreo por conglomerados. Se realizaban tres encuestas al año: una sobre pronóstico de siembra y existencia de cereales básicos en poder de los productores; las otras dos sobre siembras, cosechas, estado de los cultivos y existencias para aproximadamente 40 productos agrícolas y pecuarios aunque el objetivo principal lo constituía los cereales básicos. Para los productos agrícolas se utilizaban estimadores de "segmento cerrado" es decir se enumeran la tierras del segmento. Para las variables referidas a ganado, aves y otros bienes semovientes como personas, tractores y otros vehículos, se emplea el "segmento abierto" es decir que se enumera toda la explotación agropecuaria que tiene tierras en el segmento, sin importar si la totalidad de sus tierras están en él o no. Las principales variables relevadas comprenden: superficie, producción, rendimientos, variedad de cultivos, modalidades de siembra, utilización de insumos, productos existentes en la finca, destino de la producción, capacidad de abastecimiento, tamaño y composición del hato vacuno y porcino, existencias avícolas, producción de leche, huevos y otros productos agropecuarios.

El caso de Perú

Este caso se refiere al diseño y prueba de una encuesta en 1994 que iba a realizarse anualmente y extenderse a otros departamentos aunque luego por problemas presupuestales quedó trunca. El objetivo de la encuesta era producir información sobre área sembrada cosechada, producción, rendimientos, de cultivos transitorios y permanentes, existencia de ganado y de aves, uso de maquinaria agrícola, precios de los productos y de los insumos.

La población está formada por todas las unidades agropecuarias del Departamento que explotan sus tierras en valles o irrigaciones. El marco muestral lo constituyen los segmentos de área de aproximadamente 50 has. de extensión cada uno, identificados por límites reconocibles en el terreno en que se divide cada Departamento. Se utilizó para ello toda la cartografía existente, básicamente los planos de catastro rural a escalas 1:25,000 y 1:10,000 complementados con otros documentos cartográficos de referencia como cartas nacionales, mapas departamentales, mapas planimétricos de imagen satelital, mapas viales, planos de irrigaciones, etc.

En cuanto al diseño, el área total de cada valle o irrigación se divide en N segmentos de aproximadamente 50 Has. cada uno sin sobreposición y de tal manera que cubran toda el área agrícola de interés. Estos segmentos se estratifican de acuerdo con el uso actual del suelo. El número de estratos depende de las características de cada valle o irrigación. Se extrae luego una M.A.E. de segmentos determinando el tamaño de la muestra por asignación óptima. Cada segmento seleccionado es barrido entrevistando a todas las unidades agropecuarias con tierras en el segmento, también aquí se utiliza el concepto de "segmento cerrado" para las variables agrícolas y de "segmento abierto" para las pecuarias y otras que hacen a la explotación en general. Los estimadores usados son los ya vistos al considerar el M.A.E.

3. Ejemplo de encuesta agrícola basada en marco de lista.

El caso de Uruguay

3.1. Objetivo.

Producir información sobre área sembrada, cosechada, producción, rendimientos, pronósticos de siembra y de cosecha de los principales cultivos de cereales del país. Estas encuestas se realizan desde 1973. Aquí se presenta el diseño elaborado a partir del marco dado por el Censo Nacional Agropecuario de 1990.

3.2. Periodicidad.

Como el país tiene dos campañas agrícolas (invierno y verano) y los cultivos de invierno (trigo, avena y cebada cervecera) se siembran en junio-agosto y se cosechan en diciembre-enero, mientras que los de verano (maíz, girasol, sorgo y soja) se siembran en octubre-diciembre y se cosechan en abril-junio, las encuestas agrícolas se levantan tres veces por año de acuerdo con el siguiente esquema:

PERÍODOS DE RELEVAMIENTO E ITEMS RELEVADOS POR ENCUESTA

PERÍODOS DE RELEVAMIENTO	ITEMS			
	ÁREAS		PRONÓSTICOS	
	SEMBRADAS	COSECHADAS Y PRODUCCIÓN	DE SIEMBRA	DE COSECHA Y PRODUCC.
VERANO (Dic.-Ene.).	Cultivos de Verano	Cultivos de Invierno		Cultivos de Verano
OTOÑO (Mayo - Junio).		Cultivos de Verano	Cultivos de Invierno	
PRIMAVERA (Set.-Oct.).	Cultivos de Invierno		Cultivos de Verano	Cultivos de Invierno

3.3 Universo. (Población inferencial).

El universo a considerar está formado por todas explotaciones agrícolas que siembran, al momento del Censo, alguno de los cultivos de interés (15,003 explotaciones).

3.4. Población marco.

Los 15003 productores agrícolas se encuentran ubicados en 644 sectores censales. Por razones de eficiencia y teniendo en cuenta que las variables de interés se relacionan a volúmenes de producción y área, se eliminaron los sectores censales que en total acumulaban menos del uno por ciento del área de labranza total. Quedaron así 441 sectores censales que cubrían el 99% de la superficie dedicada a cultivos de labranza. Esta constituyó la población marco del estudio. Las explotaciones incluidas en esta población cubrían más del 98% del área para cada uno de los cultivos investigados.

3.5. Diseño de muestreo.

3.5.1. Características generales.

Teniendo en consideración razones de eficiencia, en cuanto a precisión de las estimaciones versus costos de relevamiento y en virtud de la experiencia de las encuestas anteriores, se decidió utilizar un diseño de muestreo bietápico. Al ser los sectores censales (sectores de empadronamiento censal) unidades contiguas de terreno, identificables claramente sobre el mismo, relativamente heterogéneas en su comportamiento agrícola y de un tamaño manejable, se consideró conveniente adoptarlas como UPM. Las (USM) son las

explotaciones agrícolas con chacra (con tierras de labranza). En la primera etapa, por tanto se muestrea un determinado número de UPM. seleccionadas de las 441 que componen la población. En la segunda etapa, dentro de las UPM seleccionadas, se toma un muestra de explotaciones (submuestra).

La heterogeneidad entre sectores censales con respecto a las variables de interés así como la heterogeneidad (altamente asociada al tamaño de chacra) entre las explotaciones, hacían aconsejable la adopción de un muestreo estratificado en ambas etapas. Es así que se conformaron estratos de UPM y luego dentro de las seleccionadas, estratos de explotaciones, a continuación se detallan los criterios de estratificación seguidos.

3.5.2. Estratificación de las UPM. (Sectores Censales).

En un diseño como el adoptado, los criterios de agrupamiento en estratos de las UPM juegan un papel fundamental ya que la homogeneidad dentro de cada estrato de UPM es difícil de lograr considerando una única variable (aunque sea resumen de otras). Utilizando técnicas de análisis de conglomerados se agruparon, en primer lugar, los 441 sectores censales aplicando la conglomeración sobre las ocho variables constituidas por las áreas sembradas de los cultivos de interés, previamente estandarizadas en sus valores. Se intentó también tratar de reducir el número de variables utilizando la técnica de los componentes principales sobre las variables originales, y para retener un 80% de la variación total observada se requirió de la utilización de los cuatro primeros componentes. Ambos métodos detectaron 24 sectores censales que eran observaciones atípicas con respecto a los grupos formados ("outliers"). Un análisis más detallado mostró que eran sectores con elevados valores de áreas sembradas de trigo y maíz unos y soja o girasol otros. Estos sectores se mantuvieron al margen del análisis y se procedió a reclasificar a los 417 sectores restantes mediante los dos métodos ya descritos. Se determinaron así 5 grupos de sectores censales que conformarían los estratos de UPM. Luego se clasificaron los 24 outliers dentro de estos 5 estratos utilizando análisis discriminante.

3.5.3. Estratificación de explotaciones (USM).

Las USM se estratificaron de acuerdo con el tamaño de chacra. A efectos de simplificar los cálculos posteriores así como el trabajo de preparación de las encuestas se decidió adoptar un criterio uniforme de estratificación aunque idealmente podrían haberse adoptado criterios distintos por sector lo que habría complicado enormemente el trabajo de procesamiento. De manera que, considerando las explotaciones en su conjunto se adoptaron los siguientes estratos de USM:

- **Estrato I**: Explotaciones con menos de 10 has. de área de chacra;
- **Estrato II**: Explotaciones con área de chacra entre 10 y 50 hectáreas de chacra;
- **Estrato III**: Explotaciones con más de 50 has de chacra.

3.5.4. Selección de las muestras.

Para las UPM, se decidió seleccionarlas mediante asignación óptima, fijando un nivel de error menor al 1% del área inferida para cada cultivo, con un nivel de confianza del 95%, como si se tratara de muestreo estratificado de conglomerados. Esta asignación arrojó un tamaño de muestra de 131 sectores censales.

Por restricciones presupuestales, se determinó que la muestra total de explotaciones debería ser de un tamaño comprendido entre 1200 y 1500. Tomando, en un primera instancia, 1200 productores se les repartió de manera proporcional entre los sectores sorteados y los tamaños resultantes se asignaron mediante distribución óptima entre los estratos de USM. Luego se aumentó a dos el tamaño de las muestras en aquellos casos en que la distribución mencionada conducía a 0 ó 1 (y la población lo permitía) (esto es importante para poder estimar varianzas) y se asignó una tasa de muestreo del 100% al estrato III de USM. El tamaño final de la muestra de explotaciones es de 1503.

El muestreo de UPM se hizo mediante muestreo aleatorio estratificado y el de USM mediante muestreo sistemático con arranque aleatorio independiente dentro de cada estrato.

3.6. Estimadores.

Se utilizó el estimador directo para la media y los totales de cada variable. Para las varianzas, el componente de la varianza debido al submuestreo de explotaciones, se estima como si la muestra hubiera sido aleatoria (y no sistemática) estratificada (al no poderse estimar la varianza con una única muestra sistemática.). De todas formas, por las características de la población se espera que el muestreo sistemático sea más eficiente que el M.A.S. por lo que la aproximación dada por la estimación de esta forma, podrá sobreestimar las varianzas verdaderas. Con respecto a las varianzas debe tenerse en cuenta que al error total de muestreo contribuirán las siguientes fuentes:

- error de muestreo al seleccionar explotaciones dentro de los sectores. Este es el error del muestreo estratificado y se estimará la varianza de esa manera. Esta fuente de variabilidad es "dentro de sectores" que a su vez se subdivide en "dentro de estratos" y "entre estratos" para tres estratos.

- error de muestreo al seleccionar sectores. Este es, nuevamente, el error del muestreo estratificado y es la fuente "entre sectores" que a su vez se subdivide en "dentro de estratos" y "entre estratos" para los cinco estratos de sectores.

Llamemos:

s^2_{til} a la varianza estimada entre explotaciones del til-ésimo subestrato (es decir dentro del-ésimo estrato de explotaciones (l=1,2,3), para el i-ésimo sector seleccionado del t-ésimo estrato de sectores (t=1,2,3,4,5).

s^2_t a la varianza estimada entre sectores del t-ésimo estrato de sectores.

La primera es la varianza usual ya que resulta de seleccionar m_{til} unidades de M_{til} y se computará dentro de cada estrato de explotaciones y dentro de cada sector como la varianza entre explotaciones seleccionadas en ese estrato.

La segunda es igual a:

$$s^2_t = \frac{1}{(n_t - 1)} \sum_{i=1}^{n_t} \left(\sum_{l=1}^{3} M_{til} \, \overline{y_{til}} - \overline{y_t} \right)^2$$

siendo $\overline{y_{til}}$ y $\overline{y_t}$ respectivamente las medias de las variables de interés calculadas en el estrato til-ésimo la primera y para todo el estrato de sectores censales la segunda, es decir: $\overline{y_{til}} = \frac{1}{m_{til}} \sum_{j=1}^{m_{til}} y_{tilj}$ y $\overline{y_t} = \frac{1}{n_t} \sum_{i=1}^{n_t} \sum_{l=1}^{3} M_{til} \, \overline{y_{til}}$

Luego la varianza estimada del total estimado de cada cultivo será la suma de las varianzas entre y dentro adecuadamente ponderadas, llamemos A_t y B_t a estas varianzas ponderadas:

$$A_t = N_t \left(\frac{N_t - n_t}{n_t} \right) s^2_t$$

y

$$B_t = \frac{N_t}{n_t} \sum_{i=1}^{n_t} \sum_{l=1}^{3} M_{til} \frac{(M_{til} - m_{til})}{m_{til}} s^2_{til}$$

y la varianza estimada buscada será:

$$\hat{v}(\hat{y}) = \sum_{t=1}^{5} (A_t + B_t)$$

donde:

M_t= Número de productores en el estrato t-ésimo de sectores censales en la población

3.7. Actualidad del marco.

El marco de UPM no cambia en el tiempo porque son segmentos reconocibles por

límites físicos en el terreno. El marco de USM es poco variable por las características del país. Sin embargo, cuando, en algún momento se detectaron problemas de desactualización dentro de alguna UPM seleccionada, se levantó un padrón actualizado de la misma y se volvieron a seleccionar USM para dicho ámbito.

V. CONTROL DE CALIDAD

1. Introducción.

Dada la magnitud del trabajo que implica la realización de un censo o una encuesta agrícola por muestreo, los errores pueden introducirse en muchas de las operaciones desde la organización inicial hasta las etapas finales de procesamiento y difusión de los resultados. Por dicha razón, deben extremarse los controles de calidad de los datos en las diferentes etapas censales. En general, cuando el muestreo es utilizado, los datos resultantes se presentan acompañados de una medida del error (por ejemplo un intervalo de confianza), este error es el **error muestral**, y como ya se vio, el mismo depende del diseño de muestreo. El error muestral es inherente al método: hay error porque se observa sólo una parte de una población mayor. Lo que nos ocupa ahora es el control de calidad para prevenir o corregir otro tipo de errores, los llamados genéricamente **errores no muestrales**.

Estos errores no muestrales provienen de múltiples factores debidos a deficiencias durante el desarrollo y ejecución de las tareas censales o de encuesta o también a la elección deliberada de algún método que se sabe que contiene error (por ejemplo entrevista personal en lugar de mediciones objetivas) y en este caso habrá que evaluar, medir y en última instancia controlar los errores no muestrales asociados con el método elegido. Como consecuencia de lo compleja que es la actividad de levantamiento de un censo o de una encuesta en gran escala estos errores son inevitables. Lo que debe hacerse es controlarlos, evaluarlos y tratar de disminuirlos lo más posible. En el caso de enumeración total, los errores no muestrales y en el caso de enumeración por muestreo, ambos tipos de errores deben ser controlados y reducidos a un nivel tal que su existencia no vicie los resultados finales.

A diferencia de lo que sucede con los errores muestrales, no existe una teoría comprensiva de los errores no muestrales por la naturaleza compleja de los mismos y por las innúmeras situaciones y formas en que pueden darse.

2. Fuentes de errores no muestrales.

Los errores no muestrales pueden ocurrir en cualquier etapa del desarrollo de una enumeración completa o por muestreo. Un listado no exhaustivo de fuentes de errores no muestrales es el siguiente:

- Definiciones poco claras o erróneas de los términos o redacción imprecisa o confusa de los manuales, es lo que se conoce como "instrumentos sesgados";

- Uso inadecuado de procedimientos de selección de las muestras o de medida o de estimación, o de procedimientos de entrevista, son los llamados "procedimientos sesgados".

- Omisión o duplicación de unidades debido a: definiciones imprecisas de los límites de las unidades de área en los mapas o en su descripción literal; uso de marcos; métodos erróneos de enumeración.

- Problemas por falta de personal responsable, adecuadamente entrenado y con experiencia. Los enumeradores, sobretodo cuando son pagados sobre la base de cantidad de cuestionarios completos, pueden tender a realizar el trabajo de manera descuidada, para completar la mayor cantidad de cuestionarios por día, duplicándolos o incluso inventándolos. Otros pueden no estar suficientemente motivados por el trabajo y no resuelven adecuadamente las situaciones más complicadas. Otros, por último, pueden haber sido mal entrenados y resolver erróneamente las situaciones que se le presenten.

- Dificultades en la propia recolección de datos por informantes ausentes, rechazo a brindar respuestas o errores, deliberados o no, de los propios informantes. Otro problema vinculado es el del uso de diferentes unidades de medida en distintas zonas del país o aún, por parte de distintos productores en la misma zona.

- Errores en el procesamiento de los datos. Estos errores pueden introducirse en la etapa de crítica y codificación de la información , en la entrada de datos o en el procesamiento propiamente dicho: programas erróneos, mal manejo de los archivos, etc..

- Errores en la presentación de los resultados: tablas con errores, gráficos erróneos, etc.

Las fuentes mencionadas en los dos primeros lugares, generalmente se producen en las etapas preparatorias del censo o de la encuesta y son básicamente errores de especificación. Las fuentes de error mencionadas en tercero a quinto lugar ocurren en la etapa de recolección, mientras que las dos últimas se refieren a errores de procesamiento.

3. Control de calidad de los datos.

Distinguiremos tres grandes tipos de controles de calidad de los datos, según cuando se realizan: supervisión; chequeo de los datos en la oficina y encuestas de post-enumeración (PES).

3.1. Supervisión

Una adecuada supervisión de todas las tareas censales o de planificación y ejecución

de una encuesta por muestreo es el método fundamental para preservar la calidad de los datos. La supervisión debe comenzar con las primera etapas de preparación y continuar durante todo el desarrollo de la actividad. En particular la supervisión del trabajo de campo incluyendo visitas por sorpresa a ciertas explotaciones, un permanente contacto con los enumeradores y la correcta revisión y corrección (incluyendo re-visitas) de los cuestionarios recibidos de los enumeradores son partes fundamentales para el control primario de la calidad de los datos relevados. Esta supervisión de campo tiene por objetivo fundamental mejorar la calidad del trabajo de los enumeradores. La supervisión de campo debe combinar varios criterios y métodos. Debe supervisarse, re-entrevistando a un número de productores aleatoriamente elegidos dentro del sector de enumeración. Una selección sistemática de productores a reentrevistar eligiendo, por ejemplo uno de cada diez, es un método útil. El punto de arranque y el criterio de selección de las reentrevistas debe ser desconocido por el enumerador, de lo contrario podrá inferir cuáles productores serán visitados por su supervisor y en ese caso la supervisión pierde sentido. En estas revisitas no es necesario que el supervisor vuelva a completar el trabajo censal, en primer lugar averiguará si la explotación fue visitada por un enumerador y luego completará algunas preguntas claves (como áreas, cabezas de ganado y número de personas que viven, así como aquellas preguntas que sean de más difícil llenado, por ejemplo el cuadro de uso del suelo). A estas revisitas sistemáticas agregará visitas a explotaciones alejadas o de difícil acceso si es que no fueron seleccionadas por ese procedimiento de azar. Además al momento de recibir los cuestionarios completos por el enumerador y realizar el primer chequeo de consistencia, a la par de evaluar el trabajo realizado podrá detectar explotaciones que deben ser revisitadas por problemas de llenado del cuestionario.

3.2. Chequeo de los datos en la oficina

El chequeo manual de inconsistencias (también llamado proceso de crítica o proceso de edición) trata de detectar omisiones, inconsistencias y errores obvios así como clarificar la escritura manual para facilitar y evitar errores en la entrada de datos. En esta etapa, cuestionarios incompletos o erróneos deberán ser enviados nuevamente al campo si no es posible corregirlos en la oficina. Esta etapa de crítica de la información se realiza, generalmente al mismo tiempo que la codificación de las respuestas que necesitan ser codificadas. Si bien es deseable poseer cuestionarios que estén pre-codificados en la mayor parte de las respuestas, pues esto elimina errores de relevamiento, codificación y entrada de datos; algunas respuestas, deben ser necesariamente codificadas (por ejemplo nombres de cultivos o de maquinaria que aparecen en "otros", o el nombre del Departamento, Provincia, Distrito, Municipio, etc. del cuestionario). Esta crítica de la información debería realizarse lo más cerca posible en el tiempo del proceso de relevamiento (por ejemplo terminado un sector censal y recibida la carga de trabajo del mismo, debería criticarse y codificarse a fin de corregir los errores detectados lo más pronto posible). La revisión de la crítica es importante para evitar omisiones en la misma o aun, la incorporación de errores por parte de los crítico-codificadores. El uso de los "lápices de colores" está bastante extendido: lápiz rojo para la primera crítica-codificación, lápiz verde para la revisión de la crítica y codificación y lápiz azul para el control de los procesos anteriores sobre una muestra de

cuestionarios. esto va acompañado de la instrucción precisa a los digitadores en el sentido de que hay una "jerarquía" en los colores.

El chequeo automático, por computadora, de inconsistencias es la etapa siguiente a la entrada de datos. Este chequeo busca detectar, datos faltantes, errores de entrada de datos e inconsistencias generales. Esta forma de checar los datos es complementaria del proceso de crítica manual y no debe jamás sustituirlo. La crítica automática de los datos puede hacerse interactiva al momento de la entrada de datos o a posteriori procesando los datos ya entrados en paquetes ("batches") o una combinación de ambos. El realizarlo en el momento de la entrada de datos tiene el inconveniente de enlentecerla y tiene la ventaja de que los datos entrados ya están depurados. Quizás lo más práctico es programar la entrada de datos de tal manera que rechace los errores "obvios" de digitación como por ejemplo: códigos fuera de rango, dejando el chequeo de las inconsistencias mayores para una etapa posterior. Un ejemplo- absolutamente parcial y meramente indicativo- de posibles inconsistencias que deben programarse a fin de que el computador emita un listado de "posibles cuestionarios con problemas" en un proceso donde el rango de los códigos ya fue controlado en la entrada de datos, es el siguiente:

(ejemplo)

CONDICIÓN	MENSAJE DE SALIDA DEL COMPUTADOR SI LA CONDICIÓN NO SE CUMPLE.
Si área total (Item 30)=0 entonces: Items 31 a 40 (usos de suelo, cultivos) deben estar en blanco	No. de cuestionario + "Verificar área total"
Si área total (Item 30)=0 entonces: Item 41 (uso de fertilizantes)=0.	No. de cuestionario + "Verificar fertilizantes"
Si área total (Item 30)=0 entonces: Item 42(riego)=0.	No. de cuestionario + "Verificar riego"
Si área total (Item 30)=0 entonces: Item 43 (rotación de cultivos).	No. de cuestionario + "Verificar rotación"
Área total (Item 30)=Suma de las áreas bajo cada forma de uso: Item 32.1+32.2+.....+32.7	No. de cuestionario + "Verificar usos"
Si Área total (Item 30) < 10 y stock vacuno (Item 38.1) > 20	No. de cuestionario + "Stock vacuno?"
Si Item 38.2 (número de cerdos) > 50	No. de cuestionario + "cerdos?"
Si Item 38.3 (número de aves) > 500	No. de cuestionario + "aves?"
Si Item 38.4 (número de cabras) > 100	No. de cuestionario + "cabras?"
Si Item (36.1) número de matas de banano > 10,000	No. de cuestionario + "Banano?"
Si Item (36.2) número de árboles de mango > 500	No. de cuestionario + "Banano?"

Obsérvese que hay inconsistencias lógicas (una explotación sin tierra no puede tener cultivos) y otras que pueden no ser errores pero que hay que verificarlos para descartar la existencia de errores (valores que parecen muy altos para algunas variables). La cantidad de inconsistencias que pueden preverse varía mucho y depende mucho de la imaginación de quien las concibe. Pero es un excelente ejercicio el tratar de prever la mayor cantidad posible de inconsistencias a ser checadas mediante el computador, porque ello evita muchos problemas posteriores, a la hora de obtener los tabulados. Es mucho mejor corregir los errores antes de procesar la información que volver a montar una nueva etapa de crítica cuando se tiene que producir la información final. Con carácter meramente indicativo, un cuestionario con 50 items que ya fue checado en la entrada de datos por errores de rango

de códigos, no debería tener menos de 70 consistencias cruzadas a realizar.

Una vez producidos los datos debe comprobarse su calidad final. La técnica más común para checar la calidad de los datos producidos es la **comparación** entre los datos producidos a partir del censo o de la encuesta y fuentes externas. Esta comparación debe realizarse cuando los primeros datos surjan. Por ejemplo, uno de los primeros datos que se tienen es el número total de explotaciones. Esta información puede corroborarse con los datos del último censo, con registros de titulación de tierras, con datos del censo de población, comparando por ejemplo con el número de viviendas en áreas rurales, con otras fuentes como registros de asociaciones de productores, etc. La información sobre población que vive en explotaciones agropecuarias puede checarse contra los datos de un censo de población reciente o con las estimaciones de población rural de la Oficina Central de Estadística y así sucesivamente. En estos chequeos debe tenerse especial cuidado en que el relevamiento de los datos siguió definiciones y conceptos comparables.

Otro chequeo fundamental en esta etapa y observando los datos ya tabulados es el de la **razonabilidad** de los datos. Por ejemplo es de esperar una distribución asimétrica y muy concentrada en los tramos más pequeños de área total, de las explotaciones agropecuarias o que los trabajadores agrícolas remunerados se concentren en explotaciones mayores, etc.. La distribución de las parcelas por forma de tenencia de la tierra puede checarse contra los registros territoriales de catastro o de Reforma Agraria. Esta revisión de los datos agregados, con juicio crítico es fundamental realizarla en las primeras etapas del proceso. No es extraño, observar que algunos usuarios calificados de la información (el Ministro de Agricultura, por ejemplo) descubren errores de este tipo cuando ya los datos están prontos para ser divulgados y es mucho más costoso corregirlos con la consecuencia posterior de demoras en la entrega de los resultados.

3.3. Encuestas de post-enumeración (PES).

Las encuestas de post- enumeración (PES por su sigla en inglés) son un componente fundamental del control final de la calidad censal. Buscan controlar tanto la cobertura del censo como la calidad de los datos relevados. Las encuestas de post-enumeración deben concebirse como parte del censo (y no como una tarea no censal) ello implica que debe estar incorporada al plan censal desde el comienzo y debe presupuestarse conjuntamente con el censo. Muchas veces sucede que la impaciencia por brindar los datos finales del censo así como la descarga de trabajo que significa el haber completado la enumeración censal llevan a que la PES no se realice, lo que constituye una carencia importante en la labor censal. La ejecución de la PES es la última tarea de campo a realizar. La PES debe realizarse lo más cercanamente posible en el tiempo a la finalización del relevamiento censal porque de esa manera se aprovecha el ambiente creado por el censo y asegura la colaboración de los productores. Para realizar la PES deberán elegirse los mejores encuestadores y se les asignarán área diferentes a las trabajadas por ellos durante el censo.

Las encuestas de post-enumeración se organizan en base a muestreo. Su propósito

es determinar una medida de la calidad de los datos censales y proveer tal información a los usuarios. Los datos recogidos en la PES **no deben ser usados para ajustar los resultados censales** ya que provienen de una muestra pequeña cuyo único objetivo es establecer la calidad de dichos datos.

3.3.1. Diseño de la PES.

El tamaño de la muestra y su distribución dependerán de los recursos disponibles. Como marco pueden tomarse los sectores de enumeración. Éstos tienen la ventaja de estar volcados en mapas, con límites reconocibles en el terreno. Estos sectores pueden dividirse sobre el mapa o fotos aéreas o fotos satélite, en segmentos que constituirían el marco de áreas para la PES o, pueden tomarse los sectores de enumeración como segmentos en sí mismos. Los segmentos deberían estratificarse de acuerdo a las condiciones agro-climáticas y sociológicas pues estos elementos pueden determinar niveles diferentes en la calidad de la información. Se seleccionarán segmentos que serán barridos listando las explotaciones con tierras en ellos. Durante este barrido se realizará un muestreo de explotaciones (por ejemplo de manera sistemática) donde se recogerá información sobre las variables censales más importantes. Si es posible, sería altamente deseable que en esta etapa se realizaran mediciones objetivas (por ejemplo mediciones de área o conteo de árboles o de cabezas de ganado). A partir de la información obtenida se estimará la cobertura y los errores de respuesta.

3.3.2. Análisis de los errores de cobertura y de no respuesta

La encuesta de post-enumeración permite detectar **errores de cobertura** y **errores de respuesta**. Los errores de cobertura, ya sean de omisión o de duplicación de explotaciones afectan los resultados censales más que cualquier otro tipo de error. Los errores de cobertura son, obviamente, más graves cuando se producen en explotaciones que concentran proporciones importantes de los valores de las variables de interés (por ejemplo explotaciones grandes para variables relacionadas al uso de la tierra). Los resultados de la PES deberían analizarse de tal manera que permitan extraer conclusiones referidas a ambos aspectos.

3.3.2.1. Análisis de los errores de cobertura.

Al realizarse el listado de todas las explotaciones dentro de los segmentos seleccionados para la PES y luego de comparar esta lista "correcta" de explotaciones con la efectivamente relevadas en el censo surgirán errores de cobertura que pueden ser de dos tipos: a) explotaciones omitidas en el censo y b) explotaciones erróneamente incluidas en el censo. Un modelo para evaluar ambos errores es el siguiente (Zarkovich 1966: Quality of Statistical Data, FAO):

Definamos, para cada explotación (j) , las siguientes variables aleatorias:

$$Z_j = \begin{cases} 1 & \text{si la } j\text{-ésima explotación fue censada.} \\ 0 & \text{en caso contrario.} \end{cases}$$

$$X_j = \begin{cases} 1 & \text{si la } j\text{-ésima explotación resultó ser un elemento de la población objetivo.} \\ 0 & \text{en caso contrario.} \end{cases}$$

Por ejemplo durante la PES se encuentra que la primera explotación que aparece fue correctamente incluida en el censo, entonces $Z_1 = X_1 = 1$; la segunda explotación debería haberse incluido pero fue omitida, entonces $Z_2 = 0$, $X_2 = 1$; la tercera explotación fue censada pero no correspondía hacerlo, entonces $Z_3 = 1$, $X_3 = 0$. Se define como error bruto de listado a la cantidad $G = \sum_j (Z_j - X_j)^2$ y como error neto de listado a: $D = \sum_j (Z_j - X_j)$. Con M.A.S. de k segmentos tomados de K, un estimador del error total es:

$$D = \frac{K}{k} \sum_{i=1}^{k} D_i$$

Siendo $\hat{s}_D$ el desvío estándar estimado para $\hat{D}$, el estadístico $T = \frac{\hat{D}}{\hat{s}_D}$ se distribuye aproximadamente Normal estándar en la hipótesis nula D = 0, por lo que puede hacerse un test de significación para el error neto.

3.3.2.2. Análisis de los errores de respuesta.

El segundo objetivo de la PES es, mediante la confrontación de los valores registrados para las variables consideradas en la PES y en el censo, determinar errores de respuesta. Estos errores de respuesta tienen orígenes variados: errores de registración, errores de interpretación de los declarado por el productor, errores de apreciación de cantidades por parte del informante, etc.. En este contexto, el error de respuesta medirá la discrepancia entre el valor recogido en la PES (para nuestro propósito, el "valor verdadero") y el valor reportado durante el censo (o encuesta). Valores para medir esta discrepancia sólo se tendrán para aquellas explotaciones que aparecen en ambas instancias: fueron censadas y aparecieron durante la PES.

Llamemos Y_i al valor de la variable Y observado durante el censo para la i-ésima explotación y sea Y'_i el valor de "verdadero" de la variable Y para la misma explotación. Estos "valores verdaderos" son los que se suponen se observarán en la PES para aquellas explotaciones incluidas en esta encuesta de post-enumeración.
Podemos definir el sesgo para la i-ésima unidad como $B_i = Y_i - Y'_i$. El sesgo total se define

como $\sum_i B_i$. Este sesgo y cuanto representa en el total de la variable en cuestión nos dará una idea de la magnitud de este tipo de errores.

Si estuviéramos estimando la media poblacional (μ) o el total ($N\mu$), la situación sería la siguiente: la "verdadera" media a estimar es $\mu_{Y'}$ pero nosotros habremos estimado $\mu_{Y'}$ porque los "valores verdaderos" no se observaron. Sea $\bar{y}$ la media muestral de los valores observados (obtenidos en la encuesta inicial o en el censo por enumeración completa para los elementos comunes con la PES), $\bar{y}$ estima a $\mu_{Y'}$, esta estimación es sesgada, (porque es insesgada para μ_Y y $\mu_Y \neq \mu_{Y'}$) por tanto una medida del error lo dará el ECM:

$$ECM(\bar{y}) = E_s\ (\bar{y} - \mu_{Y'})^2 = E_s\ (\bar{y} - \mu_Y + \mu_Y - \mu_{Y'})^2 = var\ (\bar{y}) + (\mu_Y - \mu_{Y'})^2$$

y nos queda la expresión usual de ECM igual a varianza más el sesgo al cuadrado, con la diferencia que el sesgo aquí no es el sesgo intrínseco al estimador sino que es el sesgo por errores de respuesta. El sesgo $\mu_Y - \mu_{Y'}$ se estimará por la diferencia $\bar{y} - \bar{y}'$ donde $\bar{y}'$ es la media muestral obtenida para las unidades incluidas en la PES. Como para las unidades comunes a la encuesta original (o censo) y la PES se tienen los B_i el sesgo puede estimarse por $\sum_{i=1}^{n} B_i$ siendo n el número de elementos comunes a la encuesta o censo inicial y a la PES.

Un estudio interesante a este respecto es presentado en "Biometrics, 14, 1958" en un artículo de J. Fleisher y otros: "Measurements errors associated with obtaining acreage estimates of cotton fields." Los autores, compararon las superficies sembradas con algodón reportadas por los productores y medidas obtenidas mediante planimetría sobre la foto aérea contra las mediciones objetivas hechas en el campo con cinta métrica. Con nuestra notación, la declaración del productor sobre el área sembrada en el campo "i" será Y_i, la medición planimétrica del área con algodón del campo "i", "otro" Y_i y la medición real con cinta métrica del campo: Y'_i . Para todos los campos (n) se tienen las series de valores Y_i y Y'_i . El valor medio "verdadero" (en este caso, la medición con cinta métrica) fue de $\bar{y}' = 10.825$ acres . Realizados los cálculos sobre las discrepancias, algunos de los resultados obtenidos son:

MEDIDAS DE LA DISCREPANCIA	Estimación de los productores	Planímetro
Sesgo medio ($\bar{b} = \frac{1}{n} \sum_i B_i$)	-0.118	0.806
Sesgo relativo: $\frac{\bar{b}}{\bar{y}'}$	-0.01	0.074
var (B_i)	1.73	0.99
ECM	67.83	79.71
Coeficiente de variación de los B_i	11.15	1.23
Covarianza entre los B_i y los Y'_i	-2.00	10.02

Estos resultados muestran aspectos muy interesantes:

1. Si bien el sesgo de las declaraciones de los productores en el conjunto es muy bajo (subestimaron las áreas en un 1%) contra un sesgo relativo del 7% del planímetro, hay mayor variabilidad en sus errores que en la estimación del planímetro (Coeficiente de variación de 11.15 contra 1.23).

2. La covarianza negativa entre los sesgos de las estimaciones de los productores y los verdaderos valores (-2.00) indica que los productores tendieron a subestimar las áreas sembradas con algodón en los campos grandes y a sobreestimarlas en los chicos. Por el contrario el planímetro tiende a subestimar la superficie en los campos pequeños y a sobreestimarla en los grandes.

3.3.3. Presentación de resultados de la PES.

Los resultados de la PES deberían presentarse de una manera tal que permitieran computar los indicadores anteriores tanto para el análisis de cobertura como de errores de respuesta. En este sentido, realizada la PES, se tendrán para los segmentos sorteados y para las explotaciones comunes con el censo, los vectores de valores Y_i (censo) y Y'_i (PES) para todas las variables incluidas en la PES. Además se tendrá información sobre el número total censado, el número de explotaciones erróneamente incluidas en el censo; el número de explotaciones omitidas, el sesgo por subcobertura o sobrecobertura (la diferencia entre las censadas y las encontradas en la PES) y los elementos para calcular los sesgos por errores

de respuesta para las variables incluidas en la PES. Estos errores y sus magnitudes relativas deberían presentarse clasificados por alguna medida relevante, por ejemplo por tamaño total.

En el Capítulo 16 de la publicación de FAO: " Realización de censos y encuestas agropecuarios" se brindan ejemplos de presentación tabular de los datos de la PES vs. los datos censales. Un estudio de los diferentes valores encontrados dará una perspectiva de la calidad de los datos relevados.

REFERENCIAS BIBLIOGRAFICAS

- **COCHRAN,William**. "Técnicas de Muestreo" 9a. Ed. Cecsa. 1993.
- **Des RAJ**, "'Sampling Theory", Mc. Graw-Hill' 1968.
- **FAO**. "Realización de censos y encuestas agropecuarios". Colección FAO: Desarrollo Estadístico #6. 1996.
- **FAO**. "Encuestas Agrícolas con Múltiples Marcos de Muestreo". Colección FAO: Desarrollo Estadístico #7. 1996.
- **FAO**. "Métodos de Muestreo para Encuestas Agrícolas". Colección FAO: Desarrollo Estadístico #3. 1989.
- **FAO**. "Un sistema integrado de censos y encuestas agropecuarios. Vol 1: Programa Mundial del Censo Agropecuario 2010". Colección FAO: Desarrollo Estadístico #11. 2007.
- **FAO**. 'Métodos Nacionales de Compilación de Estadísticas Agropecuarias" Suplemento No. 4, 1979.
- **FAO**. "Revisión Metodológica del Censo Agropecuario Mundial 2000 (documento interno)". 2011
- **FAO**. "Terminal Statement. Project TCP/DMI /2357: Preparation of the Agricultural Census in Dominica". 1996
- **GALMES, M. & HAN**, Chien-Pai, "Some Aspects of the Crop Survey in Uruguay". Journal of the IASI, Vol.XXXIII #121.-
- **GALMES, M**. "Encuesta Agrícola: Metodología. del diseño de Muestreo", Ministerio de Ganadería y Agricultura, Uruguay, 1994.
- **GALMES, M**. "Métodos de Muestreo" FAO: Programa Subregional de Capacitación en Censos Agrícolas y Encuestas Agrícolas" San Salvador, El Salvador, Diciembre de 1996.
- **LESSLER J. y KALSBEEK, W.** "Nonsampling errors in surveys" Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, 1992.
- **OFICINA DE INFORMACIÓN AGRARIA** "Construcción del Marco de Áreas en el Departamento de Lima", OIA, Ministerio de Agricultura, Perú, 1995
- **SUKHATME P.V. & SUKHATME, B.V.** "Teoría de Encuestas por Muestreo con Aplicaciones". Iowa State University Press, 1970.
- **SHEAFFER, MENDELHALL y OTT.** "Elementos de Muestreo" Editorial
- Iberoamericana, 1992
- **THOMPSON, G.** "Sampling". la. Ed. John Wiley & Sons, 1993.