INSTITUTO INTERAMERICANO DE CIENCIAS AGRICOLAS -OEA
PROGRAMA NACIONAL DE CAPACITACION AGROPECUARIA

CONFERENCIAS SOBRE MUESTREO, OFICINA DE CENSOS DE LOS ESTADOS UNIDOS

Bogotá, 1979

INSTITUTO INTERAMERICANO DE CIENCIAS AGRICOLAS - OEA

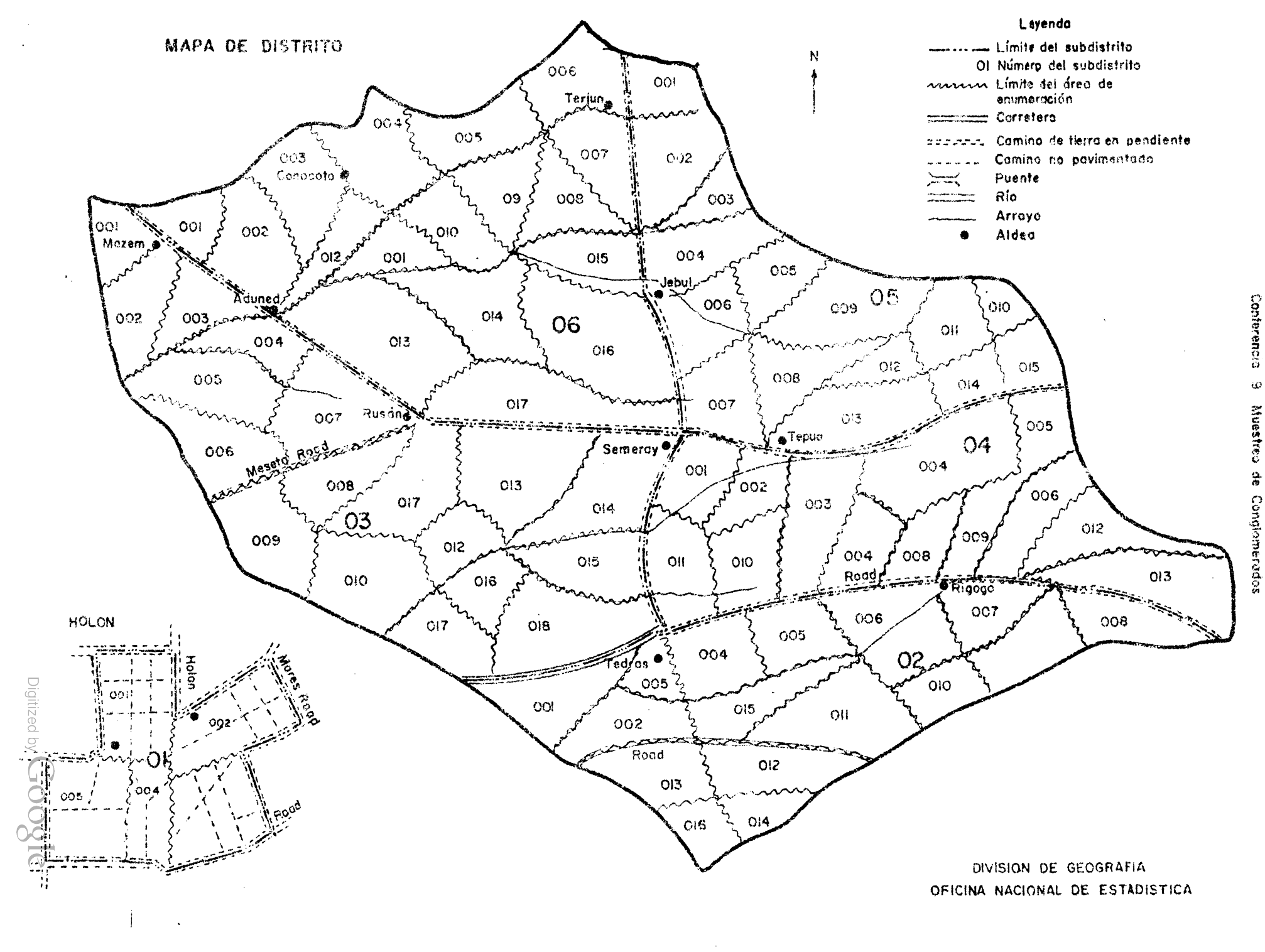
Oficina en Colombia

Programa Nacional de Capacitación Agropecuaria - PNCA

CONFERENCIAS SOBRE MUESTREO

OFICINA DE CENSOS DE LOS ESTADOS UNIDOS

Bogotá, Septiembre de 1979



B U R E A U O F T H E C E N S U S

George H. Brown, Director

Robert F. Drury, Deputy, Director

CHARLES B. LAWRENCE, JR., Assistant Director

For International Statistical Programs

International Statistical Training and Workshop Office

Beulah Washabaugh Chief

Impreso Septiembre 1971

Este informe se designó anteriormente Series ISPO I. No. 1-N y Series ISP 2. No. 1 6

CONSULTESE

U.S. Bureau of the Census: Curso Suplementario para un Estudio de Caso sobre Encuestas y Censos, Conferencias sobre Muestreo. ISP Suplemental Course Serie No. 1 (Versión en Español) Washington, D.C. 1971.

"CONFERENCIAS SOBRE MUESTREO" *

* Estas Conferencias sobre Muestreo, preparadas por la Oficina de Censos de los Estados Unidos, fueron reproducidas para su utilización en los cursos que dicta el Programa Nacional de Capacitación Agropecuaria (PNCA), que administra la Oficina del Instituto Interamericano de Ciencias Agrícolas en Colombia. La revisión de esta reproducción estuvo a cargo del Economista Nizar E. Vergara G., funcionario del PNCA. Con relación a la versión original solo se efectuó el cambio del signo σ (sigma) por S, con el único fin de facilitar la labor de Mecanografía.

Las Conferencias sobre Muestreo son resultado del esfuerzo conjunto de un grupo de personas, mayormente miembros del personal del International Statistical Programs y Statistical Methods División en el U.S. Bureau of the Census. Contribuciones particularmente significativas fueron hechas por Joseph Wakaberg, Jefe, Statistical Methods División; Margaret Gurney, Mathematical Statistician, Statistical Research División; Catherine S. Manno, Mathematical Statistician, International Statistical Programs; y Harold D. Baker, Assistant Professor, Statistical Laboratory, lowa State University. María Alicia de Madariaga, Directora de Estadística, revista del Instituto Interamericano de Estadística, fue principalmente responsable de la presente traducción.

PREFACIO Y RECONOCIMIENTO

El presente manual, No. 1 de la Serie de Cursos Suplementarios ISP. forma parte de la publicación Cursos Suplementarios para los Estudios de Caso en Encuestas y Censos que comprende diversas Conferencias sobre muestreo aplicado básico, principios de demografía, economía agrícola, lenguaje de computadoras y tópicos similares destinadas a sumentar los materiales presentados en los estudios de caso sobre encuestas y Censos.

El desarrollo de los estudios de caso ha respondido a la creciente importancia adquirida por las encuestas por muestreo y los censos en los países en desarrollo y a la necesidad de contar con un cúmulo de materiales interrelacionados sobre los diseños, estimaciones y procedimientos que son necesarios para el establecimiento de un Programa de encuestas y censos. En los documentos se cubren todos los aspectos de la planificación y ejecución de una encuesta por muestreo o un censo- desde la determinación de los temas a investigarse y la recolección de la información hasta la tabulación y análisis de los resultados.

Para proporcionar un marco realístico a la presentación de los materiales se adoptó la técnica de los estudios de caso. Se creó un país mítico, con divisiones administrativas, población urbana y rural, conocimiento general de algunas características sociales y económicas de los habitantes y una oficina nacional de estadística empeñada en el planeamiento de sus programas estadísticos. Para esta empresa se desarrollaron, de acuerdo con las recomendaciones, programas estadísticos similares, normas sobre conceptos y procedimientos. Los estudios de caso están destinados a su presentación en seminarios donde se reunirán los estadísticos responsables, en sus propios países, de implantar y llevar adelante programas de encuestas y censos.

El desarrollo de los estudios de caso estuvo a cargo del personal de la International Statistical Programs Office, U.S. Bureau of the Census, en colaboración con la U.S. Agency for international Development. Hasta la fecha se han preparado los siguientes estudios de caso: FLORENCIA: Un estudio de Caso sobre Elaboración de Datos Censales de Población y Vivienda: PROVIDENCIA: Un estudio de Caso sobre Censos Económicos; ATLANTIDA: Un estudio de Caso en Encuestas de hogares por Muestra; AGROSTAN: Un estudio de Caso para el Censo Agropecuario Mundial de 1970 y NUEVA FLORENCIA: Un estudio de Caso para los Censos de Población y Vivienda de 1979. La preparación de los materiales de los estudios de caso y de los cursos suplementarios estuvo bajo la dirección técnica de Peulah Washabaugh, anteriormente Jefe, Statistical Workshop Branch, International Statistical Programs, U.S. Bureau of the Census.

Un conjunto de personas contribuyó al desarrollo de los conceptos y los materiales, los que sintetizan conocimiento y experiencia dentro de una gama variada aunque relacionada de actividades estadísticas. Se recibieron valiosas aportaciones de miembros del personal del U.S. Bureau of the Census. Instituto Interamericano de Estadística, U.S. Departament of Agriculture y Organización de las Naciones Unidas para la Agricultura y la Alimentación. A todos ellos se expresa aquí el reconocimiento por la colaboración prestada.

CONTENIDO

CONFERENCIA 1. NATURALEZA GENERAL DE LAS ENCUESTAS POR MUESTRA

CONFERENCIA 2. CRITERIOS Y DEFINICIONES

CONFERENCIA 3. MUESTREO SIMPLE AL AZAR

Cuadros del Texto

CONFERENCIA 4. MUESTREO SIMPLE AL AZAR-TEORIA BASICA

CONFERENCIA 5. MUESTREO SIMPLE AL AZAR - TEORIA ADICIONAL

CUADROS DEL TEXTO

CONFERENCIA 6. CONSIDERACIONES PRACTICAS PARA SELECCIONAR UNA MUESTRA

CONFERENCIA 7. MUESTREO ESTRATIFICADO-TEORIA BASICA

CONFERENCIA 8. MUESTREO ESTRATIFICADO- AFIJACION EN LOS ESTRATOS

CUADRO DEL TEXTO

CONFERENCIA 9. MUESTREO DE CONGLOMERADOS

CONFERENCIA 10. MUESTREO DE CONGLOMERADOS - VARIANCIAS

CUADRO DEL TEXTO

CONFERENCIA 11. ESTIMACIONES POR RELATIVOS

CONFERENCIA 12. EL MUESTREO EN LAS ENCUESTAS AGROPECUARIAS DE MEDICIONES OBJETIVAS

FIGURAS

REFERENCIAS

CONFERENCIA 1. NATURALEZA GENERAL DE LAS ENCUESTAS POR MUESTRA

1. PAPEL DEL MUESTREO DENTRO DE LA TEORIA Y LOS METODOS ESTADÍSTICOS.

En un sentido general, la teoría del muestreo puede considerarse como coexistente con los modernos métodos estadísticos. Casi todos los desarrollos modernos en estadística se refieren a inferencias sobre la población teniendo como única información disponible una muestra de los elementos que componen dicha población. Se exponen a continuación algunas de las formas en que ésto se refleja en los programas estadísticos.

1.1 Encuestas

En la mayoría de las encuestas la población es el conjunto de todas las personas (o establecimientos industriales, fincas, etc.) de una ciudad u otra área geográfica. La información se obtiene, o se trata de obtener, de una muestra de la población para inferir características de toda la población.

1.2 Diseño y análisis de experimentos

En el diseño y análisis de experimentos, la población representa todas las posibles aplicaciones de varias técnicas alternativas disponibles. Por ejemplo, consideremos un experimento agrícola para probar varios fertilizantes. La población es infinita ya que representa el uso de los fertilizantes en todas las posibles fincas y a través de todos los tiempos. El problema es diseñar los experimentos de modo tal que den la cantidad máxima de información para extraer inferencias acerca de toda la población, estimadas con una muestra de tamaño limitado.

1.3 Control de la calidad

Cuando se aplican los métodos del control de la calidad en un establecimiento industrial, la población está compuesta, por ejemplo, por todos los productos que fabrica una máquina. se necesitan inferencias acerca de la medida en que los productos cumplen las especificaciones. El término "control de la calidad" se aplica también a una verificación por muestra de la calidad del trabajo de campo cumplido en una encuesta por muestra; la verificación por muestra se efectúa después de terminada la encuesta verdadera.

Así mismo se controla la calidad de labores que se cumplen en la oficina, por ejemplo la crítica,

codificación y perforación de las tarjetas verificando una muestra del trabajo efectuado para determinar si él mismo satisface el estandar de aceptación adoptado.

2. CONTENIDO DE LAS CONFERENCIAS

En este programa de conferencias se tratará sólo un aspecto del muestreo sus aplicaciones en las encuestas. Los principios del muestreo serán expuestos desde un punto de vista más de sentido común que matemático si bien no evitarán del todo los aspectos de orden matemático. Se pondrá especial énfasis en los métodos de muestreo utilizables bajo condiciones diferentes, las fórmulas no se demostrarán matemáticamente pero se indicará su caso. Para ilustrar las fórmulas y los métodos se incluirán dos tipos de ejemplos

a) Ejemplos símples que servirán para aclarar las técnicas y
b) Ejemplos tomados de encuestas verdaderas para mostrar la aplicación real de los métodos discutidos.

Se hará primero una exposición general de todo el tema incluyendo la naturaleza del muestreo probabilístico y la elección de las unidades de muestreo y los marcos de muestreo. Se describirán a continuación los tipos de diseños de muestras más comunes- el muestreo simple al azar, el muestreo estratificado y el muestreo de conglomerados- considerando en cada uno diferentes aspectos del diseño y de los métodos de selección de la muestra. Se tratarán también los distintos métodos de estimación de las caracterísiticas de la población a través de los resultados muestrales así como el procedimiento para determinar el tamaño de muestra requerido para alcanzar un grado particular de confiabilidad y la forma de calcular los errores de muestreo.

Nos ocuparemos del problema de estimar mediante una - muestra- los resultados que se habrían obtenido si se hubiera levantado un censo completo con el mismo cuestionario, procedimientos de enumeración y entrevista, supervisión, etc., que los usados en la muestra. Todos estos aspectos guardan relación con el error de muestreo. Por supuesto que también existen los errores ajenos al muestreo que se producen debido a las respuestas equivocadas en los interrogatorios o a preguntas pobremente redactadas. Esos errores existen tanto en un censo completo como en una encuesta por muestra. En estas conferencias no se tratarán en particular los errores ajenos al muestreo pero cabe decir que los mismos pueden tener gran importancia. En realidad, los errores ajenos al muestreo son, con frecuencia, una limitación más seria para usar las estadísticas que los propios errores de muestreo.

3. RAZONES PARA EL USO DE LAS MUESTRAS

Existen seis razones fundamentales para usar muestras:

1) Una muestra ahorra dinero (si se compara con el costo de un censo completo) cuando no se necesita una precisión absoluta.

2) Una muestra ahorra tiempo cuando se desean tener los datos con mayor rapidez que lo que sería posible con un censo completo.

3) Una muestra puede permitir concentrar la atención en los casos individuales.

4) En la industria, ciertas pruebas son destructivas (por ejemplo, las pruebas de duración de bombillas eléctricas hasta su consumo total) y sólo pueden llevarse a cabo con una muestra de productos.

5) Algunas poblaciones pueden considerarse infinitas y por lo tanto su estudio sólo es factible mediante una muestra. Un ejemplo simple lo constituye la experimentación agrícola para la prueba de fertilizantes. En un cierto sentido un censo puede considerarse como una muestra, en un instante dado de tiempo, de un sistema fundamental de causas el cual tiene características aleatorias.

6) Cuando los errores ajenos al muestreo son necesariamente grandes, una muestra puede dar mejores resultados que un censo completo ya que esos errores se controlan con más facilidad si la operación es de pequeña escala.

4. ILUSTRACIONES DE MUESTREO

Los párrafos siguientes ilustran el uso del muestreo en distintas situaciones.

4.1 Fondos limitados

Es bien conocido el uso de una encuesta por muestra cuando los fondos disponibles para recoger información son limitados. También se puede utilizar el muestreo para ahorrar dinero en la tabulación. Por ejemplo, en el Censo de 1950 de Estados Unidos la mayoría de los datos se recogieron sobre una base del cien por ciento. Sin embargo, muchas tabulaciones se prepararon sobre la base de una muestra (del 20% ó 3 1/3%) para clasificaciones detalladas especiales y ahorrar así el tremendo costo que implicaba la tabulación de 150.000.000 de tarjetas individuales perforadas. En el Censo de 1960 se utilizó en una medida mayor al muestreo tanto en la recolección como en la tabulación de los datos.

4.2 Ahorro de tiempo

Otros ejemplos referentes al Censo de 1950 de Estados Unidos ilustran el uso de las muestras para ahorrar tiempo. La enumeración en ese Censo se efectuó en el mes de abril de 1950. Se consideró que la elaboración de la información requeriría tanto tiempo que la publicación de los resultados recién se podría iniciar en 1951 para continuar a lo largo del año 1952. Se seleccionó una muestra de los resultados la que se elaboró y tabuló rápidamente publicandose los datos finales sobre la base de la misma. Esos resultados se pudieron dar a conocer entre uno a dos años antes que los del Censo completo.

4.3 Concentración en los casos particulares

En algunas encuestas se requiere una entrevista tan intensa y larga que es imposible llevarlas a cabo en otra forma que no sea a través de una muestra. Además el muestreo permite prestar atención particular a un número limitado de casos. Ejemplos en este sentido lo dan los estudios de presupuestos familiares y de condiciones de salud.

4.4 Muestreo para series cronológicas

Puede necesitarse información para una serie cronológica cuando sólo se dispone de datos para ciertos períodos y los resultados se requieren pronto. Puede tratarse de una serie sobre la actividad económica del país con datos disponibles únicamente sobre una base mensual o anual o una serie que origina una curva de conocimiento con sólo pruebas ocasionales posibles.

4.5 Control de los errores ajenos al muestreo

El Censo de 1950 de Estados Unidos proporciona un ejemplo interesante de un caso en el que la relación entre los errores ajenos al muestreo y los errores de muestreo pone en evidencia la mayor conveniencia de los resultados muestrales en comparación con los resultados de un censo completo. Desde 1940 se viene efectuando en Estados Unidos una encuesta mensual por muestra de la fuerza de trabajo. En 1950 se usó como base una muestra de 20.000 hogares; por su parte en el Censo de 1950 se recogió también información sobre la condición de las personas dentro de la fuerza de trabajo. Los resultados Censales mostraron que las cifras tanto de empleados como de desempeados eran bastante diferentes de las cifras correspondientes a esos rubros estimadas, mediante la muestra de la fuerza de trabajo; las diferencias eran mucho mayores que las que podrían esperarse como consecuencia de los errores de muestreo. El problema de declaración en el Censo había introducido un error mucho mayor que el error de muestreo de la encuesta mensual (ese error, tan considerable, había sido causado por el empleo de enumeradores que, en su mayor parte, carecían de experiencia como entrevistadores). Por lo tanto se advirtió a los usuarios de los datos Censales que los resultados muestrales constituían estadísticas nacionales más confiables en relación con la fuerza de trabajo.

5. LIMITACIONES DEL MUESTREO

En ciertas condiciones la utilidad del muestreo es cuestionable. Pueden mencionarse tres condiciones principales:

1) Si se necesitan datos para áreas muy pequeñas tiene que usarse una muestra desproporcionadamente grande ya que la precisión de una muestra depende, en gran parte, del tamaño de la muestra y no de las tasas de muestreo. En casos como ésos una muestra puede resultar tan costosa como un censo completo.

2) Si se necesitan datos a intervalos regulares de tiempo y es importante medir cambios muy pequeños entre un período y el siguiente, pueden requerirse muestras muy grandes.

3) Si los costos generales de una encuesta por muestra son desusadamente elevados debido al trabajo de Selección de la muestra, control, etc., el muestreo puede resultar poco práctico. Por ejemplo, en un país con muchas aldeas pequeñas, es posible que resulte más económico enumerar todos los hogares en las aldeas muestrales que enumerar una muestra de hogares en esas aldeas. Sin embargo, para la elaboración en la oficina, puede usarse una muestra en los hogares enumerados y reducir en esa for

ma el trabajo y los costos de producción de las tabulaciones.

CONFERENCIA 2. CRITERIOS Y DEFINICIONES.

1. CRITERIOS DE ACEPTACIÓN DE UN METODO DE MUESTREO

En las aplicaciones prácticas ha quedado demostrado repetidas veces que los modernos métodos de muestreo pueden proporcionar datos de confiabilidad conocida en forma eficaz y económica. No obstante, si bien es cierto que una muestra es una parte de una población, implicaría tener un concepto equivocado al llamar"muestra"a cualquier conjunto de números simplemente porque se trata de una parte de una población.

Para que una muestra sea aceptable desde el punto de vista del análisis estadístico, es necesario que represente a la población, que tenga una confiabilidad suceptible de medición, y que responda a un plan práctico y eficaz.

1.1 Probabilidad de selección de cada unidad

La muestra debe seleccionarse en forma tal que represente apropiadamente a la población que se está considerando. Esto significa atribuir a cada unidad (finca, hogar, persona, o cualquier otra) una probabilidad conocida de ser elegida la que deberá ser siempre distinta de cero.

1.2 Confiabilidad suceptible de medir

La confiabilidad de las estimaciones derivadas de la muestra debe ser sucentible de medir. Es decir que la muestra, además de dar las estimaciones de las caracterísiticas de la poblacipon (totales, promedios, tanto por ciento, etc.) debe proporcionar medidas de la precisión de tales estimaciones. Como repetimos más adelante, esas medidas de la precisión se podrán usar para determinar el error máximo que razonablemente puede esperarse en esas estimaciones si el procedimiento se cumple en la forma especificada y si la muestra es moderadamente grande. No se puede estimar la precisión al menos que la selección se efectúe de modo tal que no conozca la probabilidad de selección de cada unidad y se use algún tipo de muestra aleatoria.

1.3 Viabilidad

Una tercera característica es que el plan de muestreo sea práctico. Es decir, que el plan sea lo suficientemente simple y directo como para poder llevarlo a efecto en la forma proyectada; en otras palabras, que la teoría y la práctica estén de acuerdo. Por muy atractivo que parezca en el papel un plan de muestreo sólo será útil en la medida en que sea posible ejecutarlo. Cuando los métodos realmente usados son los mismos (o substancialmente los mismos) que los especificados en el plan de muestreo, la teoría de muestreo conocida de las medidas necesarias de la confiabilidad. Esas medidas derivadas de los resultados de la encuesta servirán, además, como una guía poderosa para el mejoramiento futuro de muchos aspectos importantes del diseño muestral.

1.4 Economía y eficiencia

Finalmente, un diseño debe ser eficiente. Entre los distintos métodos de muestreo que satisfacen los tres criterios establecidos más arriba, deberemos elegir, por supuesto, aquél que, en la medida de nuestro conocimiento, sea capaz de producir la mayor cantidad de información al menor costo. Si bien éste no es un aspecto primordial de un plan aceptable de muestreo, constituye una característica altamente deseable. Esto implica que se hará el uso más efectivo posible de todos los recursos y medios disponibles, por ejemplo, mapas, información estadística al alcance, preparación que posee el personal, teoría del muestreo, etc.

Consideraremos aquí solamente los métodos de muestreo que satisfacen los criterios enunciados. Expondremos la teoría básica de distintos diseños alternativos posibles de ejecutar y los métodos para medir su precisión. Destacaremos así mismo métodos prácticos de aplicación y ciertas consideraciones relativas a la eficiencia.

2. DEFINICION DE TERMINOS

2.1 Unidad de análisis

La unidad de análisis es la unidad para la que deseamos obtener información estadística. En las encuestas de tipo usual, pueden ser personas, hogares, fincas o firmas comerciales. Podrían ser también tarjetas perforadas o productos surgidos de algún proceso mecánico para algunos otros tipos de análisis. La unidad de análisis se denomina frecuentemente como un elemento de población. En una misma encuesta puede existir más de un elemento, por ejemplo, familias y personas, o número de fincas y hectáreas (o áreas) cultivadas.

2.2 Población o Universo

La población o universo es el conjunto completo de todas las unidades de análisis cuyas características se van a estimar. Las conferencias que forman este programa suplementario de muestreo tratarán fundamentalmente de poblaciones finitas de N unidades.

2.3 Unidades de muestreo

La unidad de muestreo es una unidad seleccionada del marco de muestreo Puede ser la unidad de análisis, aun cuando no es necesario. Por ejemplo, para obtener información acerca de personas podríamos usar una lista completa de un censo, o un registro de personas y seleccionar directamente una muestra de personas. Sin embargo, también podríamos seleccionar una muestra de familias e incluir en la encuesta todas las personas de las familias seleccionadas. En forma similar, podríamos seleccionar edificios completos e incluir todas las personas que viven en las estructuras seleccionadas.

2.4 Marco de muestreo

La totalidad de las unidades de muestreo de donde se extraerá la mues-

tra constituye el marco de muestreo. El marco puede ser una lista de personas o de unidades de vivienda, un archivo de registros, o un juego de tarjetas perforadas. Puede ser también un mapa subdividido o una guía de nombres y direcciones impresa en la cinta de una computadora.

2.5 Probabilidad de selección

La probabilidad de selección es la que tiene cada unidad en la población de ser incluída en la muestra. La probabilidad es un valor que oscila entre cero y uno.

2.6 Estadística

Una estadística es una cantidad que se calcula con las observaciones muestrales correspondientes a una característica, generalmente para hacer alguna inferencia acerca de la población. La característica puede ser cualquier variable asociada con un miembro de la población, por ejemplo, la edad, ingreso, condición de empleo, etc. La cantidad puede ser un total, un promedio, una mediana o cualquier otro percentil. Puede ser también una tasa de cambio, un tanto por ciento, una desviación estándar o cualquier otra cantidad cuyo valor en la población deseamos estimar.

2.7 Información independiente

Información independiente consiste de los datos conocidos antes de tomarse la encuesta o simultáneamente con ésta que no estan basados en la encuesta pero que se usan para mejorar el diseño de la investigación. Tales datos pueden utilizarse para la estratificación, para establecer las probabilidades de selección o en la estimación de los resultados finales a través de los datos de la muestra.

2.8 Fórmula de estimación

La fórmula de estimación es una fórmula en la que se usan los resultados muestrales para producir una estimación referente a toda la población. Por ejemplo,

Se usa para estimar

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i$$

$$\bar{X} = \frac{1}{N}\sum_{i}^{N} X_i$$

La fórmula se denomina frecuentemente estimador.

2.9 Intervalo de confianza

Un intervalo alrededor del valor verdadero dentro del cual se pueden establecer probabilidades fijas de que la estimación caerá en el mismo.

2.10 Muestra simple al azar (llamada también muestra al azar sin restricciones).

Se da el nombre de muestreo simple al azar al tipo más sencillo de muestreo. Si n es el tamaño de la muestra, cada una de las posibles combinaciones de n unidades elementales que se pueden formar con N unidades elementales que constituyen la población tiene la misma probabilidad de ser incluida en la muestra que cualquier otra combinación de n unidades. Puede demostrarse que en el muestreo simple al azar cada elemento de la población tiene la misma probabilidad de ser seleccionado que cualquier otro elemento. Esto es cierto ya sea que la muestra se seleccione con o sin reposición (véase conferencia 3).

TAREA DE ESTUDIO

Ejercicio 1. Con el propósito de seleccionar una muestra de la población total de una ciudad, se extrae una muestra de la guía de teléfonos de esa ciudad y se entrevista a las familias de las personas que resulten seleccionadas. ¿Satisface esta muestra los criterios de aceptabilidad? Dé una explicación.

Ejercicio 2. Para determinar la población de una ciudad en la que todos los niños de edad escolar asisten a la escuela, se selecciona una muestra de niños en las escuelas y se entrevista a sus familiares. Exponga dos razones por las que esta muestra no satisface los criterios de aceptabilidad.

CONFERENCIA 3. MUESTREO SIMPLE AL AZAR

1. INTRODUCCION

El objetivo principal de esta conferencia es presentar algunos ejemplos para aclarar los términos y conceptos expuestos antes. Usaremos ejemplos simples a fin de que las relaciones necesarias se puedan descubrir fácilmente. Serán por lo demás, ejemplos, algo artificiales, ya que nunca muestrearíamos poblaciones tan pequeñas como las de estas ilustraciones. Sin embargo, la extensión de los métodos ilustrados con poblaciones pequeñas, a situaciones más prácticas y de mayor tamaño, resultará clara.

2. METODOS PARA SELECCIONAR LA MUESTRA

Supongamos que tenemos una población hipotética de 12 individuos y que deseamos estimar el ingreso promedio de esas personas, a través de una muestra. La población completa aparece en el cuadro siguiente.

CUADRO 34. INGRESOS EN UNA POBLACION HIPOTETICA DE 12 PERSONAS

Individuos	Ingreso
A.....	$ 1,300
B	6,300
C	3,100
D	2,000
E	3,600
F	2,200
G	1,800
H	2,700
I	1,500
J	900
K	4,800
L	1,900
Ingreso total.....	32,100
Ingreso promedio.....	2,675

Supongamos que deseamos calcular las estimaciones mediante una muestra de dos individuos. La muestra se puede seleccionar en varias formas. Por ejemplo, podríamos usar 12 fichas de igual tamaño, bien pulidas, cada una de las cuales tendría escrita las letras A, B, C,..., L, no existiendo dos de ellas marcadas con la misma letra. Colocaríamos luego las fichas en un recipiente, las mezclaríamos muy bien y extraeríamos dos al azar considerando que las fichas representan los individuos seleccionados.

Este tipo de selección puede llevarse a cabo en dos formas diferentes. Puede extraerse una ficha, reemplazarla en el recipiente y extraer la segunda. En este caso la segunda ficha podría ser igual a la primera. Este procedimiento se denomina muestreo con reposición.

Por otra parte se podría extraer la segunda ficha al mismo tiempo que la primera o sea la podría seleccionar sin reponer la primera; en uno u otro caso las fichas serían diferentes. Este es el muestreo sin reposición. Cuando se extraen muestras de una población finita, la, práctica usual es aplicar el muestreo sin reposición. En su mayor parte la teoría que expondremos aqui se referirá a este método.

Existen otras formas de seleccionar dos individuos al azar. En el muestreo sin reposición, se consideran todos los pares posibles de individuos -

AB, AC, AD,..., BC, BD,...,etc. Podríamos escribir un par de letras, por cada uno de los 66 pares, en cada ficha y seleccionar una ficha única. Las muestras posibles y las probabilidades de selección son iguales que las del caso anterior.

En la práctica no se usan fichas para seleccionar unidades individualmente o en pares. El método común es usar una tabla de números al azar y elegir en la misma dos números comprendidos entre 1 y 12. Los dos números representan a dos individuos. El uso de las tablas de números al azar tiene el mismo efecto que el uso de las fichas. En la Conferencia 6 se explica el uso de las tablas de números al azar.

Cualquiera sea el método que se aplique se satisfacen los criterios para una muestra aceptable. En cada uno de los métodos, cada individuo tiene una probabilidad de selección; las probabilidades son conocidas y pueden calcularse. Tratándose del pequeño universo que estamos considerando, los tres métodos son prácticos. (En situaciones más ajustadas a la realidad, sólo sería práctico el método de los números aleatorios). Por último, los tres satisfacen las condiciones de una muestra simple al azar ya que todas las combinaciones posibles de dos individuos son igualmente probables.

3. DISTRIBUCION DE LAS ESTIMACIONES MUESTRALES

Si calculamos el ingreso promedio de los dos individuos que resultaron seleccionados en la muestra y usamos ese valor como una estimación del ingreso promedio de la población, debemos naturalmente esperar que la estimación muestral varíe en función de los individuos seleccionados. Una de las condiciones más importantes de un diseño de muestra aceptable es que la magnitud de esta variabilidad pueda predecirse siempre que el procedimiento se cumpla en la práctica en la forma especificada y siempre que exista un número moderadamente grande de elementos en la muestra. Esta condición requiere conocer la probabilidad de selección de cada muestra posible. Es decir, es posible predecir exactamente la probabilidad de que una estimación muestral caiga dentro de un intervalo especificado con respecto al valor verdadero en la población .

Consideremos una vez más la población de 12 individuos. El ingreso promedio en cada combinación puede calcularse y obtenerse una lista como la siguiente:

Individuos en la muestra	Ingreso promedio
AB	$ 3,800
AC	2,200
AD	1,650
AE	2,450

Supongamos que tenemos listados todos los 66 pares posibles de individuos y el ingreso promedio de cada par. Si calculamos luego un promedio de esos 66 ingresos promedio podremos ver que tal promedio es 32,675, exactamente el mismo que el promedio verdadero (véase Ejercicio 3). Es decir, la media aritmética de las estimaciones basadas en todas las muestras posibles del tamaño dado es igual al valor verdadero que se trata de estimar. El mismo resultado se obtendría con muestras de cualquier tamaño. La media de todos los posibles valores de una estimación se llama al valor esperado de la estimación y se representa con el símbolo y las estimaciones que tienen la propiedad de que sus valores esperados son iguales a los valores verdaderos se denominan estimaciones insesgadas. En el muestreo simple al azar la media muestral es una estimación insesgada del promedio verdadero en la población. No todos los tipos de estimaciones tienen esta propiedad (por ejemplo, no es cierta para la estimación de la desviación estándar; véase la Conferencia 4).

En el cuadro 38 se muestra una distribución de frecuencias de las medias muestrales para muestras de 1,2,3,4,5,6 y 7 individuos. Para cada tamaño de muestra se incluye el promedio de las medias.

Una comparación de las distribuciones de las estimaciones muestrales obliga a hacer dos comentarios. Primero, a medida que aumenta el número de personas en la muestra, las medias de las muestras tienden a concentrarse más y más en, o alrededor, del intervalo de clase 32,600 - $2,799 que incluye el valor promedio verdadero que es $2,675. En otras palabras, las muestras tienden a dar estimaciones relativamente más confiables que se aproximan más al valor verdadero o sea, las estimaciones tienden a hacerse más confiables a medida que aumenta el tamaño de la muestra. Segundo, las distribuciones en tanto por ciento de las estimaciones muestrales pueden usarse para predecir la probabilidad de obtener una estimación muestral dentro de intervalos especificados del valor verdadero. Por ejemplo, la proporción de resultados muestrales que caen dentro del intervalo $2,000 - $3,400 es 47 por ciento en el caso de muestras de tamaño 2; 58 por ciento en el caso de muestras de tamaño 3;69 por ciento en el caso de muestras de tamaño 4; y 78, 87 y 94 por ciento en el caso de muestras de tamaño 5,6 y 7, respectivamente (véase cuadro 38). Extrayendo muestras suficientemente grandes. La proporción de estimaciones muestrales que cae dentro de un intervalo designado alrededor del valor esperado puede hacerse tan próxima a 100 por ciento como se desee. Es decir, podemos predecir la precisión de una muestra si tenemos la distribución de todas las estimaciones muestrales para un tamano dado de muestra en la población. Más adelante veremos cómo podemos hacer esas predicciones sin tener todos los valores muestrales posibles en realidad. Sin tener ninguna información acerca de toda la población.

La concentración cada vez mayor de las estimaciones muestrales alrededor del valor verdadero ilustra la consistencia, una cualidad que poseen importantes tipos de estimaciones muestrales. Una estimación es consistente si la proporción de estimaciones muestrales que difieren del valor esperado en menos de una cantidad especificada se acerca al 100 por ciento a medida

que el tamaño de la muestra aumenta. Esto significa que si la muestra es suficientemente grande, existe un riesgo muy pequeño en el uso de las estimaciones muestrales. (Según la ilustración presentada, podría parecer que el aumento en concentración se debe al hecho de que, cuando aumenta el tamaño de la muestra, aumenta también la proporción de la población en la muestra. En realidad, se observarían resultados similares si la muestra aumentara de tamaño aun cuando solamente se incluyera una proporción pequeña del universo.)

CUADRO 38. TOTAL DE ESTIMACIONES POSIBLES DEL INGRESO PROMEDIO DERIVADAS DE MUESTRAS EXTRAIDAS SIN REPOSICION DE LA POBLACION QUE APARECE EN EL CUADRO 3A

Ingreso promedio estimado con la muestra	Número de muestras que tienen la estimación indicada del ingreso promedio con una muestra de tamaño n						
	n=1	n=2	n=3	n=4	n=5	n=6	n=7
$800 a $1,159	1	1	-	-	-	-	-
$1,200 a $1,399	1	1	3	1	-	-	-
$1,400 a $1,599	1	5	10	11	7	1	-
$1,600 a $1,799	-	6	15	25	25	16	6
$1,800 a $1,999	2	5	20	2	55	50	27
$2,000 a $2,199	1	6	22	50	78	34	61
$2,200 a $2,399	1	6	22	52	90	109	98
$2,400 a $2,599	-	6	19	52	101	139	136
$2,600 a $2,799	1	3	17	49	108	151	150
$2,800 a $2,999	-	4	16	57	101	133	130
$3,000 a $3,199	1	3	16	46	81	107	108
$3,200 a $3,399	-	3	16	38	61	79	62
$3,400 a $3,599	-	2	13	26	46	43	14
$3,600 a $3,799	1	2	10	21	27	12	-
$3,800 a $3,999	-	3	7	11	10	-	-
$4,000 a $4,199	-	3	4	10	2	-	-
$4,200 a $4,399	-	2	6	3	-	-	-
$4,400 a $4,599	-	1	1	1	-	-	-
$4,600 a $4,799	-	1	2	-	-	-	-
$4,800 a $6,399	2	2	1	-	-	-	-
Número de muestras	12	66	220	495	792	924	792
Promedio de todas las muestras posibles *	$2,675	2,675	2,675	$2,675	$2,675	2,675	2,675

* Valor esperado.

4. PREDICCION DE LA CONFIABILIDAD DE LAS ESTIMACIONES MUESTRALES (INTERVALO DE CONFIANZA.

En una situación real, no podemos seleccionar todas las muestras posibles y examinar las estimaciones derivadas de las mismas. Dependemos de una sola muestra. Por lo tanto es necesario encontrar alguna medida del alcance con que las estimaciones derivadas de varias muestras difieren del valor verdadero; esta medida, para que sea útil, debe poder estimarse a través de la muestra misma. Antes de mostrar cómo y por qué podemos hacer esto, introduciremos ciertas definiciones y relaciones que se derivan de la teoría del muestreo.

4.1 Desviación estándar

Mostraremos que existe una medida de la variabilidad en la población original, que puede estimarse mediante las observaciones de una muestra única, y con la cual es posible estimar el error esperado en la media muestral. La medida de la variabilidad en la población se denomina la desviación estándar; su cuadrado es la varianza que se simboliza con S^2 ó VAP. La varianza se define como el promedio de los cuadrados de los desvíos de todas las observaciones individuales con respecto al valor medio. Así, si se pudieran observar todos los valores en el universo, la variancia se calcularía en la forma siguiente:

$$S^2 = \frac{(X_1-\bar{X})^2 + (X_2-\bar{X})^2 + \ldots + (X_N-\bar{X})^2}{N} = \frac{1}{N}\sum (X_i - \bar{X})^2$$

donde, las X con subíndices son las observaciones individuales, y $\bar{X}$ es la media de las N observaciones en los N elementos que componen el universo.

4.2 Error estándar

En forma similar, calculemos las medias de todas las muestras posibles de tamaño n. Si elevamos al cuadrado sus desvios con respecto a la media verdadera y sacamos el promedio de esos cuadrados, tenemos la variancia de las medias muestrales. La raíz cuadrada de éste número es la desviación estándar de las medias, o como se le llama comúnmente, el error estándar de las medias de muestras de tamaño n. Como debíamos esperar, el error estándar para todos los tamaños posibles de muestra en el cuadro 38, podemos construir un cuadro como el que aparece a continuación:

---

1 Véase en la conferencia4 una explicación [illegible] de los símbolos.

CUADRO 3C. ERROR ESTANDAR DE LAS ESTIMACIONES DEL INGRESO PROMEDIO PARA DISTINTOS TAMAÑOS DE MUESTRA

Tamaño de muestra	Error estándar de la media estimada ($S_{\bar{x}}$)
1.....	$ 1.505
2.....	1.015
3.....	7.6
4.....	642
5.....	537
6.....	454
7.....	383

El error estándar de muestras de tamaño 1 es igual a la desviación estándar de la población. Veremos en la Conferencia 4 que cuando se conoce la desviación estándar es fácil obtener el error estándar para muestras de cualquier tamaño (en el caso de muestreo simple al azar) sin necesidad de tener que calcular las diferentes estimaciones muestrales posibles. Tenemos también que la desviación estándar y el error estándar pueden estimarse con una sola muestra.

4.3 Enfoque de la distribución normal

Comparando los cuadros 3B y 3C puede verse que cuando el tamaño de la muestra aumenta, las estimaciones muestrales difieren cada vez menos del valor esperado y, al mismo tiempo, el error estándar se hace más pequeño.

En los problemas prácticos de muestreo, cuando se usa una muestra razonablemente grande (por lo general 100 ó más casos), la distribución de los resultados muestrales a través de todas las muestras posibles se aproxima bastante fielmente a la distribución normal la conocida curva de forma acampanada. Para esta distribución se conocen y han sido publicadas las probabilidades de estar dentro de un intervalo especificado del valor promedio.

Esas probabilidades dependen exclusivamente del valor del error estándar.

Por ejemplo, la probabilidad de estar dentro de un error estándar es 68 por ciento, de dos veces el error estándar, 95 por ciento; y de tres veces el error estándar, 99,7 por ciento.

Las deducciones que se derivan de ésto son de fundamental importancia para la teoría del muestreo. Supongamos que hemos extraído de una población una muestra simple al azar, que hemos calculado la media de la muestra ($\bar{x}$) podemos, con bastante confianza, aseverar que $\bar{x} \pm S_{\bar{x}}$ nos dará un intervalo tal que casi en dos de tres veces estaremos en lo cierto cuando supongamos que la media verdadera cae en ese intervalo en forma similar, $\bar{x} \pm 2\, S_{\bar{x}}$ nos dará un intervalo de confianza con respecto al cual la hipótesis será correcta 95 por-

ciento de las veces; y $\bar{x} \pm 3\ S_{\bar{x}}$ un intervalo con respecto al cual la hipótesis será cierta 99.7 por ciento de las veces.

CUADRO 3D. CONCENTRACION DE LOS RESULTADOS MUESTRALES ALREDEDOR DE LA MEDIA POBLACIONAL

Muestra de tamaño n	$S_{\bar{x}}$	Tanto por ciento de medias muestrales en el cuadro 3B que difieren de la media poblacional en		
		Menos de $S_{\bar{x}}$	Menos de 2 $S_{\bar{x}}$	Menos de 3 $S_{\bar{x}}$
1.....	$1.505	75	92	100
2.....	1.015	64	97	100
3.....	786	65	96	100
4.....	642	64	97	100
5.....	537	65	97	100
6.....	454	64	97	100
7.....	383	65	97	100
Distribución normal.	...	68	95	99.7

Usando el ejemplo de los mismos 12 individuos, el cuadro 3D ilustra que lo hechos siguen bastante fielmente la teoría aun cuando se usen muestras pequeñas. En el caso de muestras grandes (que se pueden obtener de esta población pequeña si la muestra se extrae con reposición) los resultados serían mucho más similares a la distribución normal.

TAREA DE ESTUDIO

Problema: Usted desea calcular algunos promedios (medias) y desviaciones estándar del número de vacas por finca. Suponga que conoce, para cada una de 8 fincas, el número de vacas por finca como se expresa en el Cuadro siguiente:

Finca	Numero de vacas
1	4
2	5
3	0
4	3
5	2
6	1
7	1
8	0

Ejercicio 1. Calcular el número promedio de vacas por finca .

Ejercicio 2. Calcular la desviación estándar del número de vacas por finca

Ejercicio 3. Formar todas las muestras posibles de dos fincas y calcular el número promedio de vacas por finca en cada muestra.

Ejercicio 4. Preparar una distribución de frecuencias que muestre el número de muestras (de dos fincas cada una) para las que el promedio muestral cae en cada uno de los grupos siguientes:

Menos de 1,00
De 1,00 a 1,49
De 1,50 a 1,99
De 2,00 a 2,49
De 2,50 a 2,99
De 3,00 a 3,49
De 3,50 a 3,99
4,00 o más

Ejercicio 5. Calcular el promedio de las 28 medias obtenidas en el ejercicio 3 y compararlo con la media verdadera.

Ejercicio 6. Convertir los datos del cuadro 3B en distribuciones en tanto por ciento para n = 1,3,5 y 7 (mediante la división de las frecuencias por el total mostrado en la penúltima línea del cuadro). Dibujar el histograma para n = 1, n = 3, n=5, y n=7 en un mismo gráfico (sobrepuestos o paralelos) usando lápices de distintos colores si fuera necesario. Hacer los ajustamientos que corresponda por el hecho de que el primer intervalo es dos veces más largo que el intervalo estándar y el último intervalo es ocho veces más largo que el intervalo estándar. Denominar el gráfico y los histogramas individuales.

Observe como a medida que n se vuelve mayor, las distribuciones se concentran más alrededor de la media. Observe además cómo las distribuciones se aproximan más a la distribución normal a medida que n se vuelve mayor.

3.2 Error relativo

Con frecuencia deseamos considerar, no el valor absoluto del error estándar, sino un valor con relación a la magnitud de la estadística (promedio, total, etc.) que se trata de estimar. Con éste fín podemos expresar el error estándar como una proporción (o un por ciento) del valor que se estima. Esta fórmula se denomina el error estándar relativo, o coeficiente de variación, y se indica con el símbolo V, agregando un subíndice para mostrar la estadísitica a la que se refiere el error. Cuando no lleva subíndice V se refiere al error relativo de las medias de muestras de tamaño 1 (es decir, la desviación estándar de la población expresada como una proporción de la media poblacional) Así, para la estimación de la media muestral, el coeficiente verdadero de variación es:

(4.10) $$V_{\bar{x}'} = \frac{s_{\bar{x}'}}{\bar{X}} = \frac{1}{\bar{X}} \sqrt{\left(\frac{N-n}{N-1}\right) \frac{s^2}{n}}$$

que es aproximadamente igual a

(4.11) $$\sqrt{\left(\frac{N-n}{N}\right) \frac{1}{n} \frac{s^2}{\bar{x}}} \quad o \quad \sqrt{\left(\frac{N-n}{N}\right) \frac{v^2}{n}}$$

En forma similar para la estimación total

(4.12) $$V_{x'} = \frac{s_{x'}}{X} = \frac{N \sqrt{\left(\frac{N-n}{N}\right) \frac{s^2}{n}}}{N\bar{X}} \quad o \quad \sqrt{\left(\frac{N-n}{N}\right) \frac{v^2}{n}}$$

Téngase en cuenta que $V_{\bar{x}'} = V_{x'}$. (El símbolo se representa " aproximadamente igual a").

El error estándar del total estimados es N veces el error estándar de la media, mientras que el coeficiente de variación de las dos estimaciones es el mismo, este resultado, si se analiza, no es inesperado. Un total estimado se obtiene multiplicando la media muestral (una estimación) por el número de elementos en la población (que es un número conocido); la única fuente de error es la media muestral. Por lo tanto, debemos esperar que, cuando expresemos el arror total como una proporción o un porcentaje, dicho error sea el mismo de la media; sin embargo, cuando el error del total se expresa en términos absolutos, tiene que ser N veces mayor que el de la media dado que N es el factor de multiplicación.

4. FORMULAS PARA DETERMINAR EL TAMAÑO DE LA MUESTRA

Apliquemos el razonamiento de la exposición anterior al problema de determinar el tamaño de la muestra necesario para alcanzar un nivel deseado de confiabilidad, con un grado de confianza.

Definimos:

E= Límite deseado de error para la media estimada.

K= Múltiplo del error estándar, seleccionado para alcanzar un grado dado de confianza $E = K \, s_{\bar{x}'}$

Por ejemplo, K= 1 dará una probabilidad de 2 en 3 casos (probabilidad de 0.68) de que la estimación esté dentro del límite de error deseado. Para K=2, el nivel de confianza es 95 en 100 casos.

Haciendo $E = K \, s_{\bar{x}'}$ y usando la ecuación (4.6)

obtenemos:

(4.13) $$n_{\bar{x}'} = \frac{K^2 \, N \, S^2}{K^2 \, S^2 + NE^2}$$

4.1 Ejemplo:

Consideremos una población compuesta de 1.000 fincas en la que la variancia del número de vacunos por finca es 250 ($N = 1.000$ y $S^2 = 250$). Estimemos el número promedio de vacunos por finca mediante una muestra; deseamos tener una confianza razonable de que la estimación resultará próxima al valor verdadero. Supongamos que la estimación muestral no tendrá un error superior a 1 (un vacuno) con respecto al promedio verdadero y establezcamos una seguridad de que en 95 de cada cien casos el error no será mayor de 1. En ese caso,

$E = 1$

$K = 2$ (dado que 2S nos dá un nivel de confianza del 95%)

$N = 1.000$

$S^2 = 250$

$E^2 = 1$

$K^2 = 4$

Aplicando la Ecuación (4.13) vemos que n debe ser igual o mayor que

$$\frac{4(1.000)(250)}{4(250) + (1.000)1} = \frac{1.000.000}{2.000} = 500$$

CONFERENCIA 4. MUESTREO SIMPLE AL AZAR-TEORIA BASICA

Si en un caso similar aceptáramos un error no mayor de 3, con un nivel de confianza del 95 por ciento el único cambio en la fórmula sería usar para los valores de E y E^2 los siguientes:

$E = 3$

$E^2 = 9$

Luego tendríamos:

$$n = \frac{4(1.000)(250)}{4(250)+(1.000)9} = \frac{1.000.000}{10.000} = 100$$

o sea que sería suficiente una muestra de 100 casos.

4.2 Comentarios

Podemos derivar fórmulas similares para las estimaciones de totales en lugar de promedios y para estimaciones cuando el error deseado se expresa en términos relativos en lugar de términos absolutos.

Para estimar un total con un error absoluto E y un nivel de confianza K, usamos en la ecuación (4.7), $E = KS_{\bar{y}'}$ para obtener

(4.14) $$n_{\bar{y}'} = \frac{SK^2N^3S^2}{K^2N^2S^2+NE^2} = \frac{K^2N^2S^2}{K^2NS^2+E^2}$$

Para los errores relativos, si E es una proporción de la estimación, sustituímos en la Ecuación (4.11) o (4.12), $V_{\bar{y}'}$ o $V_{\bar{x}'}$, por $\frac{E}{K}$:

(4.15) $$n_{rel} = \frac{K^2NV^2}{K^2V^2 + NE^2}$$

Esta se aplica tanto a medias como a totales.

En la práctiva, por lo general no conocemos S^2 o V^2. En realidad, antes de la encuensta, ni siquiera conocemos S^2. En su lugar usamos una estimación aproximada de S^2 o V^2, etc., obtenida por alguno de los métodos que se explican en la sección 8 de la Conferencia 6.

TAREA DE ESTUDIO

Problema A: Refiérase a los datos de las 8 fincas presentados en el problema de la Conferencia 3.

Ejercicio 1: Supongamos que usted tiene información sobre el número de vacunos en una muestra de tres fincas - fincas No. 1,2 y 8. Conoce también que existen 8 fincas en el grupo (población) del que se extrajo la muestra. Estimar el número de vacunos en las ocho fincas.

Ejercicio 2: Usando la fórmula del error estándar de la media, calcular el error estándar del número promedio de vacunos por finca para una muestra de dos fincas. Véase la Ecuación (4.4) dado que N no es un número grande).

Ejercicio 3: Usando la distribución de frecuencias que se preparó en el ejercicio 4 de la Conferencia 3, determinar:

a) ¿Qué proporción de muestras (de dos fincas cada una) está dentro de más o menos una vez el error estándar del promedio?

b) ¿Qué proporción está dentro de más o menos 2 veces el error estándar.?

c) ¿Qué proporciones debería usted esperar sobre la base de la teoría?.

Problema B: Usted quiere preparar una encuesta de hogares para estimar el ingreso promedio anual por hogar. El número de hogares es 2.000.000. Sobre la base de los datos de un censo anterior, se estima que la varianza poblacional del ingreso anual por familia es 1.000.000 (es decir, S=1000).

Ejercicio 4: ¿Qué tamaño de muestra se necesita para estimar el ingreso anual promedio con una confianza del 95 por ciento de que el resultado será exacto en más o menos $100?.

Ejercicio 5: ¿Qué tamaño de muestra se necesita para estimar el ingreso anual promedio dentro de más o menos 150; igualmente con un nivel de confianza del 95 por ciento?.

Problema C: Refiérase a la ecuación (4.13) y a la Ecuación (4.6) en la Conferencia 4.

Ejercicio 6: Derivar la fórmula para $n_{\bar{x}}$, en la Ecuación (4.13) partiendo de la fórmula de $S_{\bar{x}}$ en la Ecuación (4.6) usando las definiciones de E y K dadas con la Ecuación (4.13).

CONFERENCIA 5. MUESTREO SIMPLE AL AZAR - TEORIA ADICIONAL

1. MUESTREO PARA PROPORCIONES

La estimación de la proporción de unidades que tienen una característica constituye una clase importante de estadística para la que resultan particularmente sencillas las fórmulas de la variancia y de la determinación del tamaño de la muestra.

1.1 Tipos de estadísticas para las que se usan proporciones

En el análisis estadístico las proporciones se pueden presentar en dos formas. En el primer lugar, con frecuencia, solemos tener interés en conocer más una proporción que un total o un promedio. Por ejemplo, la proporción de desempleados en la población, el porcentaje de familias con ingresos mayores que una suma dada, o la proporción de firmas interesadas en adquirir un determinado producto. En segundo lugar, podemos desear clasificar a una población en ciertos grupos diferentes y saber qué porcentaje de la población constituye cada grupo. Tales grupos podrían responder a un ordenamiento natural como serían, por ejemplo las distribuciones según edad (o a 4 años, 5 a 9, 10 a 14, etc. o según clases de ingresos, o no responder a un ordenamiento natural como ocurre en una clasificación de firmas comerciales de acuerdo con las ramas de actividad económica donde los ordenamientos de los grupos pueden variar en formas diferentes. El análisis es el mismo en todos los casos en que la estadística que se mide es la proporción del total en cada grupo.

1.2 Relación con la teoría anterior

Supongamos que adoptamos el siguiente razonamiento para la población y la

muestra. Consideremos una clase particular de unidades en la cual estamos interesados y usamos los símbolos siguientes:

A..... Es el número total de unidades, en la población, en esta clase.

a..... Es el número de unidades, en la muestra, en esa clase.

P..... Es la proporción verdadera, en la población de unidades en esa clase.

p..... Es la proporción, en la muestra, en esa clase.

Q..... Es la proporción, en la población, de unidades que no pertenecen a esa clase
(Q = 1 - P)

q..... Es la proporción, en la muestra, de unidades que no pertenecen a esa clase (q = 1 - p)

Nótese que $P = \frac{A}{N}$ $p = \frac{a}{n}$

Todas las fórmulas presentadas en las conferencias anteriores pueden aplicarse a este caso particular si es que consideramos que cada elemento de la población posee una característica que sólo puede tomar uno de dos valores; cero o uno. Si un elemento pertenece a una cierta clase en la cual estamos interesados, le asignamos el valor 1: si no pertenece, le asignamos el valor 0. Examinando la población completa podemos ver que los miembros de la clase no tienen, cada uno, el valor 1 y que los que no pertenecen a la clase, el valor 0. Así, si sumamos, para todos lo elementos en la población, el valor asignado, obtendremos A. En otras palabras, A puede considerarse equivalente a $X = \sum_{1}^{N} X_i$, expresión que ya hemos estudiado anteriormente. En términos similares, $P = \frac{A}{N}$ puede x interpretarse en la misma forma que $\bar{X} = \frac{X}{N}$. Cabe así usar las fórmulas ya conocidas. Veremos que en este caso su aplicación es particularmente sencilla.

1.3 Fórmulas aplicables

En el muestreo para proporciones se aplican las fórmulas siguientes (con muestreo simple al azar):

(5.1) $p' = p = \frac{a}{n}$ y $a' = p' N$.

Es decir que puede obtenerse una estimación de la proporción en la población usando la proporción en la muestra y una estimación del número total de unidades que tienen la característica, multiplicando la proporción en la muestra por el número total de unidades en la población. Además

(5.2)
$$S^2 = PQ; \qquad s^2 = pq$$

PQ es la variancia de la población. Nótese que se trata de la variancia de la distribución de la población cuando se ha dado a cada elemento el valor 1 ó 0 según que dicho elemento pertenezca o no a una clase dada (es decir, posea o no el atributo en cuestión). Esa variancia se puede estimar mediante pq, salvo en el caso en que n sea un número muy pequeño (por ejemplo $n = 30$), de ser así la formula es:

$$s^2 = pq\left(\frac{n}{n-1}\right)$$

La variancia de las estimaciones de la proporción calculadas son todas las muestras de tamaño n, es

(5.3)
$$S_p^2 = \frac{PQ}{n}\left(\frac{N-n}{N-1}\right)$$

La estimación de esta variancia, calculada con una sola muestra de n observaciones, es

$$s_{p'}^2 = \frac{PQ}{n}\left(\frac{N-n}{N-1}\right)$$

Estas fórmulas son iguales a las presentadas anteriormente para $S_{\bar{x}}^2$, sustituyendo S^2 por PQ, y para $s_{\bar{x}}^2$, sustituyendo s^2 por pq. Véase las Ecuaciones (4.4) y (4.8).

En forma similar, las fórmulas dadas en la Conferencia 4 para el error estándar relativo de una media y el error estándar de un total estimado, toman ahora la forma que figura a continuación a la izquierda; las fórmulas para estimar esas cantidades, son las expresiones que aparecen a la derecha:

(5.4)
$$V_{p'} = \sqrt{\frac{Q}{Pn}\left(\frac{N-n}{N-1}\right)} \qquad V_{p'} = \sqrt{\frac{q}{pn}\left(\frac{N-n}{N-1}\right)}$$

$$S_{a'} = N\sqrt{\frac{PQ}{Pn}\frac{N-n}{N-1}} \qquad s_{a'} = N\sqrt{\frac{pq}{n}\left(\frac{N-n}{N}\right)}$$

Una vez más, el error estándar relativo del total es igual al error estándar relativo de la media.

Las fórmulas para calcular el tamaño de muestra necesaria para estimar, con una exactitud dada, una proporción puede expresarse de dos maneras. Si

Si el error E está dado en términos absolutos , el tamaño de muestra n se obtiene mediante la Ecuación (4.13); en esa ecuación $S^2 = PQ$:

(5.6) $$n_{p'} = \frac{K^2 NPQ}{K^2 PQ + NE^2}$$

Si el error E está expresado como un error relativo, se usa la Ecuación (4.15) con $V = \frac{Q}{P}$:

(5.7) $$n_{Rel} = \frac{K^2 N Q}{K^2 Q + NE^2}$$

Para obtener el tamaño de muestra requerido para estimar el número total de unidades en una clase, con un error absoluto E, se usa la Ecuación (4.14) con $S^2 = PQ$:

(5.8) $$n_{a'} = \frac{K^2 N^3 P Q}{K^2 N^2 PQ + N E^2} = \frac{K^2 N^2 PQ}{P^2 NPQ + E^2} .$$

Para obtener el tamaño de muestra requerido para estimar el número total de unidades en una clase, con un error relativo E, se usa la ecuación (5.7) presentada anteriormente.

1.4 Ejemplos

Los ejemplos siguientes ilustran el cálculo del error de muestreo y la determinación del tamaño de muestra necesario para estimar una proporción.

1.41 Estimación del error de muestreo

Supongamos que la proporción de fincas que cultivan maíz en una región es .40. ¿Cuál sería el error de muestreo si se estimara dicha proporción con una muestra aleatoria de 500 fincas, sabiendo que el número total de fincas en la región es 10,000?. En este caso;

N= 10.000
n= 500
p= .40
Q= .60

Tenemos:

$$S^2_p = \frac{pQ}{n}\left(\frac{N-n}{N-1}\right) = \frac{PQ}{n}\left(\frac{N-n}{N}\right)$$

$$= \left(\frac{(.4)(.6)}{500}\right)\left(\frac{10,000-500}{10,000}\right)$$

$$= \left(\frac{.24}{500}\right)\left(\frac{9,500}{10,000}\right) = .000456$$

$$S_{p'} = \sqrt{.000456} = .021$$

¿Qué interpretación debe darse a la cifra .021?.
Este número nos dice que si fijamos un intervalo alrededor de la proporción verdadera de amplitud .40 $\pm$ .021 (o sea, de .379 a .421), existe una probabilidad razonablemente buena (68 por ciento) de dque con una muestra de 500 fincas obtengamos una proporción cuyo valor esté comprendido entre .379 y .421. Si duplicamos el intervalo para tener una amplitud de .358 a .442, la probabilidad de que la estimación muestral esté dentro de la misma será de alrededor del 95 por ciento. Y si tomamos tres veces.021, o sea .063, la probabilidad de que la estimación esté dentro del intervalo será .997 (casi la certeza). En la práctica se acostumbra usar una amplitud de dos veces sigma (2 desviaciones estándar) para asegurar una suficiente confianza en la exactitud a las estimaciones. Cuando los resultados de la muestra se van a usar para tomar decisiones de mucha importancia y deseamos estar casi absolutamente seguros del intervalo dentro del cual estará la estimación muestral, podemos usar un nivel de tres sigma.

En este ejemplo se conoce tanto la proporción (.40) como la probabilidad de que la estimación muestral esté dentro de un cierto intervalo alrededor de la proporción. En la práctica nos interesa por lo común una situación inversa a ésta ya que no conocemos la proporción verdadera si bien tenemos una estimación muestral de .40 calculada mediante una muestra de 500 fincas extraída de una población de 10,000. Deseamos establecer los siguientes intervalos con respecto a la cifra muestral dentro de los cuales podemos esperar que caiga el valor verdadero de la media. Prácticamente, en todos los caso podemos, reemplazando la expresión "valor verdadero" por "estimación muestral", hacer las mismas afirmaciones. Es decir, si según la muestra, existe un .40 de fincas que cultivan maíz y establecemos un intervalo de .40 $\pm$ .021, tendremos una probabilidad de alrededor del 68 por ciento de que la cifra verdadera esté dentro de ese intervalo; una probabilidad del 95 por ciento, si el intervalo es .358 a .442; etc.

1.42 Tamaño de muestra necesario para una confiabilidad dada

¿Qué tamaño debe tener la muestra para estimar, dentro del .05, la proporción en las 10,000 fincas que cultiva maíz.

1) Si la proporción verdadera es P= .40?
2) si la proporción verdadera es P= .90 ?

Debemos usar la fórmula del tamaño de la muestra. Como el error (.05) está expresado en términos absolutos, corresponde aplicar la Ecuación (5.6):

$$n_p, = \frac{k^2 N P Q}{k^2 P Q + N E^2}$$

Como antes tenemos, N = 10,000 y

para (a) : P = .40 y Q = .60:
para (b) : P = .90 y Q = .10.

En el caso de una seguridad del 95%, tomamos:

$k = 2$; $k^2 = 4$

El error será

$E = .05$; $E^2 = .0025$.

Sustituyendo en la Ecuación (5.6) se llega,

Para (a): $$n_p, = \frac{4(10.000)(.4)(.6)}{4(.4)(.6) + 10.000(.0025)}$$

$$n_p, = \frac{9,600}{.96 + 25} = 370.$$

para (b): $$n_p, = \frac{4(10.000)(.9)(.1)}{4(.9)(.1) + 10.000(.0025)}$$

$$= \frac{3.600}{.36 + 25} = 142.$$

Se habrían obtenido los mismos resultados usando la fórmula que expresa n en función del error relativo (Ecuación 5.7). Para (a) el error del .05 expresado como porcentaje de la estimación es

$\frac{.05}{.40} = .125$; para (b), es $\frac{.05}{.90} = .056$. Usando esos resultados en la Ecuación (5.7)

$$n_{Rel} = \frac{K^2 N Q}{K^2 Q + PNE^2},$$

Se derivan los mismos valores de n, es decir, 370 para (a) y 142 para (b).

1.5 Procedimiento cuando P se refiere a un subconjunto de una clase

Con frecuencia la proporción que se estima es un porcentaje, node la población total, sino de una clase particular. Por ejemplo, podemos tener in-

terés en conocer, no ya el número de desempleados expresado como un porcetaje de toda la población sino el número de desempleados expresado como un porcentaje con respecto a las personas en la fuerza de trabajo. O podemos necesitar conocer la proporción de firmas con más de cinco empleados dentro de una cierta actividad económica. En tales casos puede prepararse una aproximación bastante cercana a un análisis exacto usando las fórmulas dadas antes pero interpretando los números N y n como si se refirieran a la clase que nos interesa. Es decir, N sería, no la población total, sino el número de personas en tal clase (por ejemplo, el número de personas que componen la fuerza de trabajo) según la estimación muestral; n sería el número de casos en la muestra en la clase mencionada; a sería el número de casos en la muestra en el subgrupo (por ejemplo, el número de desempleados).

1.6 Cuadro de valores de $\sqrt{\frac{PQ}{n}}$

El cuadro 5A muestra los valores de $\sqrt{\frac{PQ}{n}}$ para valores dados de P y n Como se expone en las secciones 2.2 y 2.3 más adelante, podemos usar la fórmula simplificada

(5.9) $$p = \sqrt{\frac{PQ}{n}}$$

Para calcular el error estándar de la proporción de unidades que poseen cierto atributo si la muestra es una muestra al azar (simple) sin restricciones y si N es tan grande en relación con n como para que el factor $\frac{N-n}{N}$ sea un número muy próximo a 1.

Como no se conoce la proporción verdadera en la población (p) se puede usar en la Ecuación (5.9) la estimación derivada de la muestra (p) para obtener una estimación del error estándar de P:

(5.10) $$s_{p'} = \sqrt{\frac{pq}{n}}$$

La mayoría de las muestras no son muestras aleatorias simples sino estratificadas. En las Conferencias 7 y 8 veremos que el efecto de este hecho es la disminución del error de muestreo si se compara con el de una muestra simple al azar de igual tamaño. Sin embargo, la mayoría de las muestras que se usan en las encuestas son además conglomeradas lo que (como veremos en las Conferencias 9 y 10) tiene el efecto opuesto de hacer mayor el error de muestreo si se lo compara con el que se obtendría con una muestra simple al azar de igual tamaño. Cuando la muestra es, al mismo tiempo, estraficada y conglomerada, las fórmulas del error estándar resultan más complejas (véase las Conferencias siguientes).

Algunas veces no es posible desarrollar fórmulas exactas pero puede obtenerse una estimación aproximada del error estándar usando la fórmula simple de la Ecuación (5.9) con una cierta tolerancia por el efecto neto esperado del apartamiento de la aleatoriedad en el diseño muestral.

CUADRO 5A. ERROR ESTANDAR DE UNA ESTIMACION DE UNA PROPORCION EN EL MUESTREO SIMPLE AL AZAR

$$S_{p'} = \sqrt{\frac{PQ}{n}}$$ para valores especificados de P y n)

n=número de casos M.	P=proporción de unidades q'poseen 1 característica(Q=1-P tieneel mismo E. EST												
	.001 o .999	.002 o .998	.01 o .99	.02 o .98	.03 o .97	.04 o .96	.05 o .95	.10 o .90	.15 o .85	.20 o .80	.25 o .75	.30 o .70	.40 o .60	.50
50	.0045	.0063	.0141	.0198	.024	.028	.031	.042	.051	.057	.061	065	069	071
100	.0032	.0045	.0099	.0140	.017	.020	.022	.030	.036	.040	.043	046	049	050
200	.0022	.0032	.0071	.0099	.012	.014	.016	.021	.025	.028	.031	033	035	035
300	.0018	.0026	.0058	.0081	.0099	.012	.013	.017	.021	.023	.025	027	028	029
400	.0016	.0023	.0050	.0070	.0086	.010	.011	.015	.018	.020	.022	023	024	025
500	.0014	.0320	.0045	.0063	.0076	.0089	.0098	.013	.016	.018	.019	021	022	022
600	.0013	.0018	.0041	.0057	.0070	.0082	.0090	.012	.015	.016	.018	019	020	020
700	.0012	.0017	.0038	.0053	.0065	.0076	.0083	.011	.014	.015	.016	017	019	019
800	.0011	.0016	.0035	.0050	.0061	.0071	.0078	.011	.013	.014	.015	016	017	018
1,000	.0010	.0014	.0032	.0044	.0054	.0063	.0070	.0095	.011	.013	.014	015	015	016
1,200	.0009	.0013	.0029	.0040	.0049	.0058	.0064	.0087	.010	.012	.013	013	014	014
1,500	.0008	.0012	.0026	.0036	.0044	.0052	.0057	.0077	.0093	.010	.011	012	013	013
1,700	.0008	.0011	.0024	.0034	.0042	.0049	.0053	.0073	.0087	.0097	.011	01	012	012
2,000	.0007	.0010	.0022	.0031	.0038	.0045	.0049	.0067	.0081	.0090	.0097	010	011	011
2,500	.0006	.0009	.0020	.0028	.0034	.0040	.0044	.0060	.0072	.0080	.0087	0092	0098	0100
3,000	.0006	.0008	.0018	.0026	.0031	.0039	.0040	.0055	.0066	.0073	.0079	0084	0090	0092
3,500	.0005	.0008	.0017	.0024	.0029	.0034	.0037	.0051	.0061	.0068	.0073	0078	0083	0084
4,000	.0005	.0007	.0016	.0022	.0027	.0032	.0035	.0047	.0057	.0063	.0068	0073	0077	0079
4,5'	.0005	.0006	.0015	.0021	.0025	.0030	.0033	.0045	.0054	.0060	.0065	0069	0073	0074
5,000	.000	.0006	.0014	.0020	.0024	.0028	.0031	.0042	.0051	.0057	.0061	0065	0069	0071

1 Valores de n mayores de 5.000: cuando n se multiplica por 100, el error estándar se divide por 10.

2 En la práctica se usaría el valor muestral puesto que se desconoce el valor poblacional P.

Si las unidades de análisis están conglomeradas formando grupos bastante pequeños por ejemplo, 5 unidades de vivienda ó 25 personas en cada conglomerado, y dentro de un conglomerado personas de características bastante similares como ocurre en las zonas rurales podría multiplicarse el error estándar de la proporción que aparece en el Cuadro 5A por un factor tal como 1.25. En un conglomerado grande, como ser una manzana en una ciudad que contuviera 40 ó 50 unidades de vivienda, el factor que se podría aplicar a los valores del Cuadro 5A sería 1.75 a pesar de que las personas que componen el conglomerado fueran menos similares en una zona urbana que en una zona rural.

El tamaño del factor de corrección que se use depende del diseño muestral y de la naturaleza de la población; algunas veces ese factor puede ser estimado, aproximadamente, por un estadístico experimentado en muestreo sobre la base de la experiencia anterior y de las fórmulas matemáticas que involucren la "correlación intraclass" (véase conferencia 10).

2. RELACION ENTRE EL TAMAÑO DE LA MUESTRA Y EL TAMALO DE LA POBLACION

Volvamos nuevamente a la fórmula básica de la que se han derivado las fórmulas anteriores. La fórmula básica es la presentada como Ecuación (4.4) en la Conferencia 4:

$$(5.11) \qquad S_{\bar{x}'} = \left(\frac{N-n}{N-1}\right)\frac{S^2}{n}$$

Nótese que la variancia de muestreo de la media es igual a la variancia de las observaciones individuales en la población multiplicada por el factor $\frac{1}{n}\left(\frac{N-n}{N-1}\right)$. ¿Qué ocurre cuando la muestra aumenta desde su menor tamaño posible (n=1) a su mayor tamaño posible (n=N)? Cuando n=1,

$$S^2_{\bar{x}'} = \left(\frac{N-1}{N-1}\right)\frac{S^2}{1} = S^2.$$

Esta expresión establece el hecho familiar de que el error estándar (al cuadrado) de las medias de muestras de tamaño 1 es igual a la desviación estándar (al cuadrado) de las observaciones individuales en la población. En el extremo opuesto, es decir cuando n=N,

$$S^2_{\bar{x}'} = \left(\frac{N-N}{N-1}\right)\frac{S^2}{N} = 0.$$

En otras palabras, si la muestra incluye toda la población, la estimación de la media está exenta de error.

Para los tamaños de muestra comprendidos entre ambos extremos, ¿qué efecto tiene la fracción de muestreo $\frac{n}{N}$ (tasa de muestreo) sobre el error estándar? La respuesta a este interrogante, algunas veces inesperada para los estudiantes, es que para poblaciones de tamaño grande en comparación con el ta-

maño de la muestra, la precisión de la media estimada está determinada por el tamaño absoluto de la muestra (n) y no por la fracción de muestreo ($\frac{n}{N}$).

Este resultado es una consecuencia del hecho de que cuando N es un número grande en comparación con n, el factor $\frac{N-n}{N-1} \doteq 1$ (El símbolo $\doteq$ se lee "aproximadamente igual a.") luego, $s^2_{\bar{x}} \doteq \frac{s^2}{n}$ Por lo tanto es evidente que el error depende de s^2 y de n, no de $\frac{n}{N}$

Por otra parte, en el caso de poblaciones pequeñas, la fracción de muestreo afecta los resultados. Por ejemplo, supongamos que tenemos dos poblaciones de igual media y varianza: $\bar{x} = 50$ y $s^2 = 100$, pero de tamaños diferentes, $N_1 = 40$ y $N_2 = 400$. Si de cada población tomamos una muestra de igual tamaño, digamos $n_1 = n_2 = 20$, los errores estándar estarán relacionados (en forma inversa) con las fracciones de muestreo. La Ecuación (5.11) nos da:

	N	n	$\frac{n}{N}$	$C_{\bar{x}}$,
1a. Población	40	20	.50	1.6
2a. Población	400	20	.05	2.2

El número de unidades muestrales necesario para alcanzar la misma precisión será mayor para la segunda población (que es la de mayor tamaño). Sinembargo, el número de unidades muestrales necesario para alcanzar una confiabilidad dada no aumenta indefinidamente cuando crece el número de elementos en la población.

2.1 Ejemplo

El Cuadro 5B indica el tamaño de muestra necesario para tener una estimación de la media de la población dentro de un error del 5 por ciento con una probabilidad correspondiente a los límites 2 $V_{\bar{x}}$, (es decir, E= .05 y k= 2), en el caso de poblaciones que oscilan entre 50 y 10,000,000 de elementos y con $V^2 = .10$ en cada caso. Esos resultados se obtuvieron usando la Ecuación (4.10) de la Conferencia 4.

CUADRO 5B. NUMERO DE ELEMENTOS NECESARIOS PARA UNA PRECISION DADA

(E = .05 y k = 2 cuando v^2 = .10)

Número de elementos en la población(N)	Número de elementos necesarios en la muestra (n)	$\frac{n}{N}$
50*	38	.76
100	62	.62
1,000.....	138	.14
10,000.....	158	.016
100,000.....	160	.0016
1,000,000.....	160	.00016
10,000,000.....	160	.000016

* Usese la Ecuación (4.10) si N es menos de 50.

Este cuadro pone en evidencia que para poblaciones pequeñas el tamaño de muestra necesario para una precisión dada aumenta a medida que la población crece; no obstante cuando la población se hace muy grande el tamaño de la muestra se acerca a un número fijo. El tamaño de muestra mayor que necesitaríamos para esta exactitud (con v^2 = .10) sería 160 elementos, siendo éste aproximadamente el número que necesitaríamos ya fuera que la población tuviera 10,000 ó 10,000,000 de elementos. Además, si para una población tan pequeña como 1,000 elementos usaramos una muestra de 160, dicha muestra sería algo mayor que lo necesario si bien el exceso carecería de mayor importancia.

2.2 Multiplicador finito (o factor de corrección por la población finita)

La fórmula de la variancia relativa de una media, en el caso de una muestra simple al azar es

$$v^2_{\bar{x}'} = \left(\frac{N-n}{N-1}\right)\frac{v^2}{n}$$

ó

$$v^2_{p'} = \left(\frac{N-n}{N-1}\right)\frac{Q}{Pn},$$

Esta fórmula puede descomponerse en dos partes:

$$\frac{N-n}{N-1} \quad y \quad \frac{v^2}{n} \quad o \quad \frac{Q}{Pn}.$$

El tamaño de la población entra a formar parte de la fórmula únicamente a través de la expresión

$$\frac{N-n}{N-1}$$

Por lo general este factor se le denomina como multiplicador finito. Si la población fuera infinita, el factor mencionado sería igual a 1 y las fórmulas se expresarían con mayor sencillez; así:

$$V^2_{\bar{x}'} = \frac{V^2}{n} \quad o \quad V^2_{p'} = \frac{Q}{Pn}$$

El valor $\frac{N-n}{N-1}$ es aproximadamente igual a $1 - \frac{n}{N} = 1 - f$, siendo $\frac{n}{N} = f$, la tasa de muestreo. Si la tasa de muestreo es pequeña, digamos menos de .05, el efecto del multiplicador finito es de poca importancia y, para los fines prácticos, la situación es igual al muestreo de una población infinita por ejemplo, en el Cuadro 5B, para una población de 10.000 la tasa de muestreo es .0016 y el correspondiente multiplicador finito .9984. Esos multiplicadores son casi iguales y la diferencia con la unidad puede ignorarse sin correr riesgo ninguno.

2.3 Simplificación para poblaciones grandes

En el censo de poblaciones grandes y tasas de muestreo pequeñas, el multiplicador finito puede ignorarse obteniendo así fórmulas más sencillas.

FORMULA SIMPLIFICADA

	Para el valor verdadero	Para la estimación derivada de la muestra
Variancia de la media	$S^2_{\bar{x}'} = \frac{S^2}{n}$	$s^2_{\bar{x}'} = \frac{s^2}{n}$
Variancia de la proporción	$S^2_{p'} = \frac{PQ}{n}$	$s^2_{p'} = \frac{pQ}{n}$
Variancia relativa de la media	$V^2_{\bar{x}'} = \frac{V^2}{n}$	$V^2_{\bar{x}'} = \frac{V^2}{n}$
Variancia relativa de la proporción	$V^2_{p'} = \frac{Q}{Pn}$	$V^2_{p'} = \frac{q}{pn}$
Variancia total	$S^2_{x'} = N^2 \frac{S^2}{n}$	$s^2_{x'} = N^2 \frac{s^2}{n}$

Variancia del número total de unidades que poseen un atributo	$S^2_{a'} = N^2 \frac{PQ}{n}$	$s^2_{a'} = N^2 \frac{pq}{n}$
Variancia relativa del total	$v^2_{x'} = v^2_{x'} = \frac{v^2}{n}$	$V^2_{x'} = v^2_{x'} = \frac{v^2}{n}$
Variancia relativa del número total de unidades que poseen un cierto atributo	$v^2_{a'} = v^2_{p'} = \frac{Q}{Pn}$	$v^2_{a'} = v^2_{p'} = \frac{q}{Pn}$

Se dan a continuación fórmulas simplificadas para el tamaño de muestra necesario para un precisión especificada:

1) Cuando los errores están expresados en términos absolutos, los tamaños de muestra necesarios para estimar las medias son:

$$n_{x'} = \frac{K^2 S^2}{E^2}$$ Comparar con (4.13)

y

$$n_{p'} = \frac{K^2 PQ}{E^2}$$ Comparar con (5.6)

Para estimar totales, los tamaños de muestra necesarios son:

$$n_{x'} = N^2 \frac{K^2 S^2}{E^2}$$ Comparar con (4.14)

y

$$n_{a'} = N^2 K^2 \frac{PQ}{E^2}.$$ Comparar con (5.8)

2) Cuando los errores están expresados en términos relativos

$$n_{Rel} = \frac{K^2 V^2}{E^2}$$ Comparar con (4.15)

y

$$n_{Rel} = K^2 \frac{Q}{PE^2}.$$ Comparar con (5.7)

Nótese que ambas fórmulas en (2) se aplican tanto a la media como al total.

TAREA DE ESTUDIO

Problema A: Se selecciona una muestra al azar de personas empleadas a una tasa de uno por ciento. La muestra comprende 30,000 personas empleadas de las cuales 12,000 están dedicadas a actividades agropecuarias.

Ejercicio 1. Del total de personas empleadas, ¿cuál es la proporción que se dedica a actividades agropecuarias y cuál es el error de muestreo de esa proporción?.

Problema B: Un país tiene una población total de 10,000,000 habitantes. Se selecciona una muestra simple al azar de 1/10 de uno por ciento, es decir, 10,000 individuos. En la muestra existen 4,000 personas que forma parte de la fuerza de trabajo de las cuales 200 se clasifican como desempleadas.

Ejercico 2. ¿Cuál es la proporción de desempleados dentro de la fuerza de trabajo del país y cuál es el error estándar de esaproporción

Problema C: Una ciudad tiene una población total de 50,000 personas. Se Selecciona una muestra simple al azar del 20 por ciento, es decir, 10,000 individuos. En la muestra 4,000 personas pertenecen a la fuerza de trabajo de las cuales 200 son desempleadas.

Ejercicio 3. ¿Cuál es la proporción de desempleados dentro de la fuerza de trabajo de esa ciudad y cuál es el error estándar de esa proporción?.

Ejercicio 4. Explique por qué razones existen diferencias en los errores estándar de los ejercicios 2 y 3.

CONFERENCIA 6. CONSIDERACIONES PRACTICAS PARA SELECCIONAR UNA MUESTRA

1. USO DE LAS TABLAS DE NUMEROS AL AZAR

Las tablas de números al azar se usan en el muestreo para evitar el tener que realizar ciertas operaciones, tales como la selección de fichas numeradas de una urna, para determinar las unidades que se deben incluir en la muestra. La experiencia, además, ha hecho ver que es prácticamente imposible mezclar en forma completa las fichas entre cada selección; que ciertos recursos, como usar cartas o dados, no son convenientes debido a las imperfecciones en la construcción de las cartas o los dados; que al pensar en números al azar las personas tienden a favorecer ciertos dígitos, etc. En consecuencia, todos esos procedimientos no dan, en realidad, a cada miembro de la población una misma probabilidad de selección. En cambio, el uso de una

tabla de números al azar reduce la cantidad de trabajo y asegura, en un grado mayor, a todos los elementos una misma probabilidad de selección.

Existen muchas tablas de números al azar. Dentro de la serie Practs Computers, existen varios, en particular las compiladas por Tippett y por Kendall y Smith. La corporación RAND ha publicado a MILLION RANDOM DIGITS. Así mismo en Statistical Tables por Fischer y Yates se incluye una tabla, al igual que en otras varias fuentes. En muchas de esas publicaciones se da una descripción de los métodos de compilación y uso de las tablas. A su vez se pueden emplear computadoras para generar números seudo-aleatorios.

En general esas tablas muestran conjuntos de dígitos aleatorios ordenados en grupos tanto en sentido horizontal como vertical. Para seleccionar un conjunto de números aleatorios podemos comenzar en cualquier lugar de una página. Además, una vez seleccionado el primer número, se puede continuar una columna hacia abajo, o hacia arriba, una fila hacia un lado o el otro, o de acuerdo con cualquier pauta deseada.

1.1 Ejemplos

Para obtener un número al azar entre 1 y un cierto número dado, por ejemplo entre 1 y 273, siga los siguientes pasos: Observe el número de dígitos que componen el número límite superior (en 273 hay tres dígitos); use ese mismo número de columnas contado a partir de la primera (o cualquier otra predeterminada) columna: y comience desde arriba (o desde alguna línea predeterminado). Cada línea en un grupo de tres columnas contiene un número de tres dígitos. Elija el primero de esos números comprendido entre 001 y el límite superior dado entre 001 y 273en nuestro ejemplo. Rechace los números que sean mayores de 273 al igual que 000. Si se desea más de un número aleatorio, continúa hacia abajo a lo largo de las tres columnas, eligiendo cada número de tres dígitos comprendido entre 001 y 273 hasta tener la cantidad deseada de números al azar de tres dígitos. Si un mismo número aparece dos o más veces, alíjalo sólo una vez.

Supongamos que partimos de una tabla de números al azar como la siguiente:

1 0 8 9	8 7 1 9
9 3 8 5	7 9 0 2
6 9 3 4	8 6 6 0
0 0 5 2	1 0 0 7
5 7 3 6	9 2 4 9
1 9 0 1	5 9 8 8
5 3 7 2	6 2 1 2

Dentro de los límites de los números que figuran en los ejemplos siguiente, seleccionaremos en la tabla anterior números al azar usando cada número seleccio-

nado una sola vez.

Ejemplo A:

Seleccionar tres números al azar entre 1 y 10. Elegimos primero una columna arbitrariamente decidiendo que 0 representa a 10. Supongamos que hemos elegido la quinta columna. El primer número en esa columna es 8; el segundo, 7; el tercero, 8 nuevamente. Como este número ya ha sido seleccionado, lo pasamos por alto y tomamos el número siguiente que es 1. Los tres números seleccionados son, por lo tanto, 8, 7 y 1.

Ejemplo B:

Seleccionar cinco números al azar entre 1 y 30. Supongamos que tomamos las dos primeras columnas como punto de partida. Primero elegimos 10; rechazamos 93 ya que no está comprendido entre 01 y 80; elegimos 69; rechazamos 00(que representa a 100); y tomamos luego 57, 19 y 53.

1.2 Precauciones en el uso de las tablas de números al azar

Si una tabla de números al azar se usa con frecuencia se debe tener cuidado de no consultar siempre la misma parte. Por ejemplo, si todas las veces se toma el primer número de la misma columna en la misma página, se está utilizando en forma repetida un mismo conjunto de números y no se alcanzará por lo tanto una verdadera aleatorización. Cuando se usa con frecuencia una tabla de números al azar conviene elegir cada vez un nuevo punto de arranque.

2. MARCO DE MUESTREO

Para seleccionar una muestra correctamente es necesario tener un marco de muestreo, es decir, una lista de todos los elementos (o un equivalente, como podría ser una lista de manzanas, unidades de vivienda etc.) a fín de poder conocer la probabilidad de selección de cada elemento. El marco no tiene que ser literalmente una lista. Si se muestrean tarjetas, cuestionarios, etc., los mismos documentos pueden considerarse como el marco si bien, en este caso, es necesario tener la seguridad de que el archivo está completo. Por ejemplo, al extraer una muestra de un archivo de registros, debemos asegurarnos de que no falten registros del mismo - retirados ya sea para usarlos o que están a la espera para ser archivados ya que entonces esos registros no tendrían ninguna probabilidad de ser seleccionados. Asimismo, si se utiliza un registro de población llevado por alguna autoridad local debemos asegurarnos de que la lista está actualizada. Por ejemplo, podría ser que no se hubieran incorporado a la lista las nuevas familias creadas por matrimonio. Como las nuevas familias, al igual que las que recientemente acaban de mudarse, difieren en sus características de las más antiguas y asentadas, la muestra resultaría sesgada.

Cuando se utilizan listas o registros locales una medida aconsejable en efectuar un verdadero recuento de su completabilidad sobre una base más o menos

aproximada. Para ésto puede visitarse el área que se va a muestrear y seleccionar unas pocas familias (o fincas, o firmas comerciales) dispersas en la zona verificando posteriormente si figuran en la lista. De ser posible, es mejor elegir familias del tipo de las que podrían faltar de la lista ya que esto proporciona una mejor prueba. Se puede tener así una idea aproximada de la exactitud de la lista.

3. PROBABILIDAD DE SELECCION DE LAS UNIDADES

Si algunas unidades tienen más de una probabilidad de ser seleccionadas surgen algunos problemas por ejemplo, cuando se extrae una muestra de un archivo en el que algunas personas figuran más de una vez, cuando se lecciona una muestra de familiares a través de una muestra de personas, etc. Para ilustrar este punto, pensamos en una muestra de niños que concurren a las escuelas usada para seleccionar familias. Es evidente que si se extrae una muestra de familias seleccionando primero una muestra de personas e incluyendo luego las familias a las que pertenecen esas personas. Las familias tendrían probabilidades desiguales, ya que cuanto más grande sea la familia mayor será su probabilidad de resultar seleccionada.

De forma similar, cuando se selecciona una muestra de clientes de una firma comercial usando un archivo de registros con una boleta (o tarjeta) es usada por cada compra. Los clientes que hayan efectuado más de una compra tendrían una probabilidad mayor de selección.

Para evitar los sesgos que resultan de dar a algunas de las unidades una probabilidad de selección mayor que a otras, es conveniente limitar el procedimiento de muestreo de modo que cada unidad tenga sólo una probabilidad de selección. Por ejemplo, si se selecciona una muestra de familias, se podría establecer una regla en el sentido de incluir la familia solamente si la persona que resultó seleccionada es el jefe de familia. Dado que cada familia tiene únicamente un jefe, todas las familias tendrían una probabilidad igual de ser seleccionadas.

La persona especificada de la que depende que se seleccione la familia no necesita ser el jefe; puede ser también el miembro más anciano, o el de menos edad, etc. El único requisito es que en cada familia exista una y sólo una persona como ésa. En forma similar, si se extrae una muestra de clientes, podría limitarse la muestra usando sólo las tarjetas con la fecha más lejana para cada cliente, etc.

Si bien la técnica explicada en los párrafos anteriores es generalmente recomendable ya sea que la muestra se extraiga de un archivo, un conjunto de cuestionarios, o se seleccione en el terreno, existen otras técnicas que también podrían usarse. Esas técnicas proporcionan estimaciones insesgadas del universo si bien no satisfacen estrictamente las condiciones del muestreo simple al azar. Presentamos algunas de ellas :

1) Una vez seleccionada la muestra inicial incluyendo todas las familias para las que se había seleccionado una (o más) personas, se agrupa la muestra según tamaño de las familias. Es evidente que las familias

de dos miembros tendrán una probabilidad doble de ser seleccionadas que las familias de un solo miembro; que las de tres miembros tendrán una probabilidad triple, etc. Por lo tanto, en lugar de entrevistar todas las familias en la muestra, se entrevista solamente la mitad de las familias de dos miembros, un tercio de las familias de tres miembros, etc.

2) Se procede como en el caso anterior pero se entrevistan todas las familias en lugar de la mitad, la tercera parte, etc. Sin embargo, al tabular los resultados, se tabula cada tamaño de clase separadamente y se multiplican los resultados de las familias de dos miembros por por 1/2, los de las familias de tres miembros por 1/3, etc., antes de sumar los resultados.

4. MARCOS QUE INCLUYEN UNIDADES FUERA DE ALCANCE

Algunas veces el único marco disponible es una lista que incluye ciertas unidades que están fuera del alcance del universo definido para la encuesta. Por ejemplo, supongamos que se desea efectuar un análisis especial de las características censales de la población masculina. La única fuente para extraer la muestra es un archivo de tarjetas que contiene tarjetas para todas las personas, tanto hombres como mujeres, y que no permite separar las tarjetas de mujeres. Aun así el archivo podría usarse como marco aunque mediante el procedimiento de selección al azar resultarán designadas tarjetas de hombres y de mujeres. En un caso como éste el procedimiento aproximado es tomar sólo las tarjetas seleccionadas correspondientes a los hombres y rechazar las correspondientes a las mujeres.

No sustituya

Un procedimiento que se utiliza algunas veces equivocadamente (y que puede crear serios sesgos) es sustituir cada tarjeta "mujer" seleccionada por la próxima tarjeta "hombre" del archivo. Se cometen así dos errores:

1) La tasa de muestreo resulta más alta que la especificada. Además no puede calcularse la tasa de muestreo realmente obtenida al menos que se conozca el número total de hombres. Asimismo no permite usar la inversa de la tasa de muestreo, es decir N -- n, como multiplicador para producir estimaciones de totales a partir de los resultados muestrales.

2) Una objección más seria a esta sustitución es que se pueden introducir sesgos en el proceso de selección. Supongamos que tenemos una lista de todas las unidades de vivienda y que deseamos seleccionar una muestra de viviendas ocupadas únicamente. Si usamos un procedimiento por el cual se sustituye cada unidad desocupada que cae en la muestra por la unidad ocupada próxima, las unidades ocupadas en la vecindad de las desocupadas tendrán una probabilidad de selección doble la que les corresponde por figurar en la lista y la que les corresponde por la unidad vacante vecina. Si las unidades desocupadas son más -

fáciles de encontrar en los barrios pobres o de baja condición, las unidades ocupadas en tales barrios resultarían sobrerepresentadas en la muestra.

5. MUESTREO SISTEMATICO

La cantidad de trabajo necesario para extraer una muestra simple al azar puede ser bastante considerable cuando el número de unidades que hay que seleccionar es grande. Por ejemplo, para obtener una muestra del 5% de 20,000 elementos, sería necesario seleccionar 1.000 números al azar de una tabla de números al azar y luego seleccionar las unidades designadas en la población. En la práctica muchos estadísticos prefieren seguir un método diferente. Por lo general se extrae una muestra de ese tamaño tomando un número al azar comprendido entre 1 y 20 seleccionando a continuación cada 20 ésimo elemento. Así, si el número al azar es 3 se toman luego los elementos 3,23,43,63, etc., hasta llegar a 19,983. La inversa de la tasa de muestreo (20, en este caso) se denomina el intervalo de muestreo . Los procedimientos para estimar la media, el total o una proporción son los mismos del muestreo simple al azar.

Este tipo de muestreo se denomina muestreo sistemático. No es exactamente igual al muestreo simple al azar pero es un método aceptable de muestreo ya que se conoce la probabilidad de selección de cualquier elemento y se pueden calcular los errores de muestreo.

Si los elementos en la población están ordenados en una forma casi aleatoria (es decir, con muy pequeña correlación entre los elementos sucesivos), los resultados del muestreo sistemático guardarán íntimo acuerdo con los del muestreo simple al zar. La experiencia enseña que, en general, los dos métodos dan resultados de aproximadamente la misma exactitud. Con frecuencia la muestra sistemática tendrá un error de muestreo algo más pequeño dado que asegurará el que la muestra esté dispersa a través de toda la población. Para evaluar la confiabilidad de las estimaciones de una muestra sistemática podemos usar las fórmulas del muestreo simple al azar; el resultado usualmente sobrestimará algo el error estándar del muestreo sistemático. En otras palabras, subestimaremos ligeramente la confiabilidd de las estimaciones. Existen ciertas fórmulas para calcular exactamente los errores estándar de muestras sistemáticas; sin embargo no las consideraremos en estas conferencias.

5.1 Precausiones en el uso del muestreo sistemático

Existe una situación en la que el muestreo sistemático dará una confiabilidad muy pobre. Se trata del caso en que el ordenamiento de los elementos en la población sigue una pauta muy regular y el intervalo de muestreo de la muestra sistemática cae dentro de dicha pauta. Por ejemplo, supongamos que todas las familias de una cierta población están compuestas exactamente por cuatro personas- el jefe, la esposa y dos hijos. La población ha sido listada en ese orden dado y deseamos extraer una muestra sistemática del 25% a partir de dicha lista obtener cierta información. Como el proce-

dimiento de muestreo consiste en tomar cada cuarta persona a partir de un arranque al azar, se podrían obtener cuatro muestras posibles:

1) Si el comienzo aleatorio es 1 - la muestra estaría compuesta enteramente por jefes de familia.

2) Si el comienzo aleatorio es 2- la muestra estaría compuesta enteramente por esposas de jefes de familia.

3) Si el comienzo aleatorio es 3 ó 4 - la muestra estaría compuesta enteramente por hijos.

En un caso así los resultados de una muestra a otra tendrían casi la máxima variación posible y sería muy posible que las estimaciones basadas en una de las muestras difieran marcadamente de los valores verdaderos en la población. Sin embargo, aun en este caso extremo, las estimaciones serían insesgadas, es decir, las medias de las estimaciones en todas las muestras posibles serían iguales a los promedios poblacionales.

Si bien es muy difícil que en la práctica se presenten casos como el dado, pueden existir situaciones cercanas al mismo. Si se tiene alguna sospecha de cualquier irregularidad en la secuencia del listado, que pudiera coincidir con el intervalo de muestro, se debe evitar o modificar el muestreo sistemático.

5.2 Muestreo sistemático modificado

Una variación del muestreo sistemático que podría utilizarse cuando existe algún ordenamiento sistemático en la población consiste en aplicar, dentro de cada intervalo de muestreo, un número al azar distinto. Como ejemplo, usemos el caso anterior de una muestra del 25 por ciento donde los miembros de la familia figuraban listados en un cierto orden- jefe, esposa, hijos. En el muestreo sistemático, una vez elegido el número aleatorio, este determina la pauta para la muestra completa. Como ya dijimos, si el número al azar es 1, la muestra será la 1a, 5a, 9a, 13a, etc. persona (todas jefes de familias); si el número aleatorio es 2, la muestra incluirá a la 2a, 6a, 10a, 14a, etc. persona (todas esposas del jefe). Para evitar esta dificultad podemos seleccionar un número al azar distinto dentro de cada grupo de 4 personas a fin de evitar un intervalo constante entre nuestros casos muestrales. El esquema de selección está indicado a continuación:

Número aleatorio (1 a 4)	Grupo de cuatro personas	personas seleccionadas
3	1o.	3a.
1	2o.	5a.
2	3o.	10a.
1	4o.	13a.
4	5o.	20a.

etc.

Este sistema requiere un trabajo mayor que el muestreo sistemático regular pero evita la posibilidad de las pautas indicadas más arriba. No estamos queriendo implicar que pautas como la ya mencionada existan con frecuencia y que el muestreo sistemático debe evitarse. En la mayoría de los casos, el muestreo sistemático produce resultados muy satisfactorios.

5.3 El número de serie como fuente de muestreo

Con frecuencia en los archivos de las oficinas de muestreo los registros llevan un número de serie. Podemos sacar ventajas de este hecho para extraer la muestra; por ejemplo, podríamos elegir todos los registros cuyos números de serie terminen en 5, 7 ó algún otro número elegido de una tabla de números al azar. Sin embargo, antes de decidirnos por este sistema debemos asegurarnos de que el último dígito del número de serie es realmente aleatorio y no representa un cierto código; si así fuera; podríamos obtener, al seleccionar repetidas veces el mismo último dígito, una muestra compuesta únicamente por unidades de un tipo particular. Cuando no existe un miembro de serie se puede, con frecuencia, asignar uno al azar sin aumentar mucho los costos y usarlo para extraer muestras.

6. CONTROLES

Una vez seleccionada una muestra es necesario verificar el número de casos realmente obtenido en relación con el número de casos esperado (e) que se calcula aplicando la tasa de muestreo al número de casos en el universo). La existencia de discrepancias puede indicar que el procedimiento de muestreo no se ha seguido correctamente. Por ejemplo, podría ser que se hubiera olvidado muestrear algunos cajones del archivo que se están utilizando en esos momentos y que así se hubiera omitido una parte de la población y obtenido, por lo tanto, un número de casos menor que el esperado. Verificaciones adicionales para descubrir si la muestra tiene aspectos fuera de lo común pueden servirnos igualmente para comprobar si el muestreo se efectuó realmente en la forma proyectada.

7. USO DE LOS DATOS DE VERIFICACION EN EL MUESTREO

Con mucha frecuencia, cuando se selecciona una muestra para un cierto estudio, se recogen y tabulan, además de los rubros de interés especial para el estudio en cuestión, otros datos muestrales para algunos rubros básicos de los cuales ya se tienen los totales poblacionales. Dichos totales ya conocidos se designan como "datos de verificación" o "información independientes". Si los resultados de la muestra en relación con los rubros conocidos son bastante similares a los totales poblaciones ya disponibles, se suele decir que esta coincidencia da "validez" a la muestra y prueba que la misma proporcionará buenos resultados para los otros rubros.

En realidad, la así llamada "validez" no demuestra que tengamos un "buen" procedimiento de muestreo ni que la muestra dará "buenas" estimacio-

nes para los otros temas que se están investigando mediante la encuesta. Sólo sobre la base de un método aleatorio de selección de una muestra es que se puede agregar a muestras estadísticas un error de muestreo y evaluar la probabilidad de que las estimaciones caigan dentro de límites especificados del valor verdadero; por lo tanto, es obvio, que no podemos tener confianza en aquella "validez".

Existen, no obstante, tres aplicaciones aceptables de los datos de verificación:

1) Se pueden usar para mejorar el método de muestreo, por ejemplo, como base para la estratificación. (Este tema será tratado en las dos conferencias siguientes.)

2) Se pueden calcular los errores estándar de las estimaciones hechas con los datos de la muestra. Si los datos de verificación y las estimaciones muestrales para esos mismos rubros difieren algo más de lo que razonablemente podría esperarse de acuerdo con el tamaño de los errores estándar calculados, entonces puede ser que estemos ante una indicación de que los procedimientos de muestreo no se están cumpliendo en forma correcta o que se ha incurrido en algún otro error en la conducción de la encuesta

3) Pueden usarse para mejorar el método de preparación de las estimaciones con los datos de la muestra; por ejemplo, se pueden ajustar las estimaciones muestrales usando la relación entre el valor verdadero del rubro de verificación y la estimación muestral de dicho rubro (usando una estimación por relativo). En conferencias posteriores trataremos más a fondo este punto.

Las tres aplicaciones mencionadas del uso de los datos de verificación (o información independiente) son satisfactorias ya que nos permiten extraer inferencias estadísticas.

8. ESTIMACION DE LA VARIACION DE UNA POBLACION

Cuando tratamos el problema de la determinación del tamaño de la muestra a usarse. Supusimos que era conocida la variancia poblacional (σ^2) del rubro que íbamos a estimar. Si no se conoce el valor de la variancia se puede, por lo general, estimar. En las secciones siguientes se mencionan algunos de olos métodos más sencillos para hacerlo.

8.1 Uso de datos anteriores

Es posible que, para alguna fecha anterior, se conozca el valor de la variancia de ciertas características por ejemplo, a través de estudios efectuados en el pasado de un censo previo. En tal caso resultará conveniente, en general, usar ese valor ya conocido aun cuando se refiera a la población que existía en ese entonces y no a la presente. La experiencia nos indica que la variancia de un rubro dado tiende a cambiar, con el tiempo, mucho más

lentamente que el valor promedio del rubro en sí. Aun cuando varie el valor promedio, el error relativo puede ser bastante estable.

8.2 Variancia de una proporción

Si la estadística que se mide es una proporción por ejemplo, la proporción de fincas que cultivan maíz la variancia de la población es PQ. Para poder estimar S^2 sólo se requiere hacer una conjetura más o menos acertada con respecto a P (la proporción en la población). En la medida en que esa conjetura sea razonablemente buena, obtendremos una buena estimación de S^2. Por ejemplo, supongamos que el valor verdadero de P es.4; luego, el valor de S^2 = PQ sería .4 x.6 =.24. Si, en cambio, nuestra conjetura (algo pobre) es que P = .3, le estaríamos dando a la variancia un valor estimado de .3 x .7 = .21, cantidad que sólo difiere del valor verdadero en un 10 por ciento.

8.3 Muestra especial para estimar variancias

Si no se conoce el valor de la variancia se puede extraer una muestra con ese objeto. La muestra puede ser de cuestionarios de una encuesta anterior o conectada con una prueba previa a la encuesta y destinada a ese fin. Usando los cuestionarios así obtenidos podríamos luego preparar una estimación de la variancia.

8.4 Muestreo en etapas

En ciertas ocasiones es posible y conveniente realizar una encuesta en dos etapas. En la primera sólo se enumera una submuestra (porción aleatoria de la muestra total), la que se analiza para estimar la variancia y efectuar una revisión del tamaño total de la muestra si fuera necesario. En la segunda etapa se enumera la porción restante de la muestra teniendo en cuenta los cambios realizados si se hubiera introducido alguno.

TAREA DE ESTUDIO

Problema A: Usted tiene una población de 185 personas de la que desea seleccionar una muestra.

Ejercicio 1. Seleccione una muestra simple al azar de 20 personas. Use las columnas 1,2 y 3 de la tabla de números al azar que figura en la página siguiente; use las columnas 4,5 y 6 si no le alcanzan las columnas 1, 2 y 3; continúe con las sucesivas columnas si así lo exige su caso. Liste los números asignados a las 20 personas seleccionadas y describa el procedimiento que utilizó para la selección.

Ejercicio 2. Seleccione una muestra sistemática de 20 personas. Liste los números asignados a las 20 personas y describa el procedimiento que utilizó para la selección.

Problema B : Suponga que en una ciudad existen 25,000 unidades de vivienda y que se dispone de una lista exacta de ellas. Las unidades de vivienda están listadas alfabeticamente según el nombre de la familia que las ocupa. Aparecen, asimismo, las direcciones de las viviendas. Usted desea efectuar una encuesta por muestra de unidades de vivienda alquiladas para estimar la distribución de los alquileres mensuales. Ha decidido que para obtener datos de una exactitud conveniente necesita una muestra total de 400 unidades de vivienda ocupadas por inquilinos. La lista de unidades de vivienda existentes en la ciudad no dice si se trata de unidades ocupadas por dueños o por inquilinos pero usted sabe que alrededor de dos tercios de todas las unidades de vivienda de la ciudad están alquiladas.

Ejercicio 3 : Describa cómo extraería la muestra mencionando en la descripción (a) el método de selección; (b) la tasa de muestreo y (c) el tratamiento dado a las unidades de vivienda ocupadas por propietarios.

TABLA DE NUMEROS ALEATORIOS*

1089	8719	2272	1359	3328	0014	6773	1278	2761	3550
9385	7902	5034	6723	3835	6978	7084	3992	5857	2377
6934	8660	0311	2979	0095	2647	8299	5163	0073	7788
0052	1907	4866	6497	4138	8144	0284	2906	0316	4810
5736	9249	3062	7604	8137	4575	2245	6309	1601	3580
1901	5988	2633	8605	2964	0736	3046	0612	9663	3663
5372	6212	9675	0286	6825	7823	5778	2680	1227	5186
4057	0762	6469	2735	5082	3852	7457	5729	8436	6478
5484	0770	7222	4912	0062	0609	9291	4056	5034	8359
0125	9592	3729	7858	5153	7200	1308	9638	0345	9293
5587	2698	2784	0458	0122	4721	3963	2916	3763	0468
7963	1937	6002	4490	5404	2817	6818	7120	8495	2692
8894	0546	6771	8401	1359	9935	8594	7513	8303	0649
9090	2972	0932	3907	6077	7374	0992	8851	6723	8748
7986	0132	8683	8563	2374	4215	3574	4177	8495	6662
2676	6123	4352	3195	8505	2599	6526	2200	2269	3864
1727	5363	5319	9610	4556	0760	8243	7406	7222	0675
4710	7892	3258	2504	0443	8042	1712	0583	3907	8166
7654	5820	1428	1657	7152	7329	4229	7790	9551	6453
1499	2908	6193	8309	4699	5572	7590	9369	6847	4523
0012	8520	9535	3820	0060	4456	0791	6723	1055	5004
4117	6797	4427	6919	7026	2643	3586	9697	2811	6479
4069	0228	6215	3098	4938	1823	6886	9313	2709	7613
0581	5826	2212	2668	2824	0764	0729	0905	3802	9636
1555	9910	4082	0037	9691	0235	3525	5214	4533	0153
7384	6217	3438	9094	5752	9526	7276	1836	7591	7976
4786	4094	5646	8893	5015	8753	0875	4531	9582	9243
7872	5228	3242	8446	3322	4612	4522	0585	0367	2352
2297	3084	8969	8087	1431	3059	8601	4111	2034	5289
0858	9152	3641	9678	0447	1330	1285	4933	5545	8405
9296	2919	6086	0779	8357	2519	5941	7895	9381	4056
4703	2537	5381	1356	6949	1488	0988	0369	4208	0413
7311	4110	5017	3633	5565	8507	3973	0477	9167	5612
0016	6380	3071	5758	4867	9975	8704	1468	4884	1879
3818	4607	3996	8535	3503	0623	2485	2595	9545	1035
0747	5454	9260	1416	2171	9525	6016	9430	2253	2176
2828	6877	2570	4049	2073	1220	0593	4690	1342	8664
2962	3218	7985	8659	2767	3818	5981	9162	5498	0184
1560	3004	0883	2339	1363	4219	0189	4453	2364	0900
0806	1970	4130	7998	1736	5243	4212	0621	8439	9898

* Tomada de Tracts For Computers, No. XV, Random Sampling Numbers, por L.H.C. Tippett, Cambridge University Press, Londres, 1927.

CONFERENCIA 7. MUESTREO ESTRATIFICADO-TEORIA BASICA

1. DESCRIPCION DEL PROCESO DE ESTRATIFICACION

El muestreo simple al azar no requiere ningún esfuerzo especial para obligar a la muestra a ser representativa de la población; la tendencia a tener este carácter es inherente al proceso mismo. En esa clase de muestreo el único camino para reducir el error de muestreo es aumentar el tamaño de la muestra. Sin embargo, antes de comenzar una encuesta, si se tiene algún conocimiento acerca de la población se puede utilizar esa información en la estratificación y así reducir el error de muestreo. Para ésto puede ser necesario recurrir al juicio de un experto.

El Muestreo estratificado

Es un método de muestreo que consiste en clasificar primero los elementos de la población en grupos y seleccionar luego, en cada grupo, una muestra simple al azar tomando, al menos, un elemento de cada grupo. (Para estimar la media es suficiente tomar un elemento de cada grupo, pero, en cambio, para estimar su confiabilidad se requieren dos elementos. Por lo general, se necesitan más de dos elementos para hacer estimaciones de precisión suficiente). El proceso que se sigue para establecer los grupos ya mencionados se conoce como estratificación; los distintos grupos se denominan estratos. Los estratos pueden reflejar regiones geográficas de un país, áreas densamente o escasamente pobladas, distintos grupos étnicos, o cualesquiera otros grupos.

En la estratificación agrupamos elementos similares a fin de que la varian cia dentro de cada grupo (s^2_i) sea pequeña; al mismo tiempo es deseable que las medias ($\bar{X}_i$) de los distintos estratos sean lo más diferentesposible.

En el muestreo estratificado las probabilidades de selección de un grupo al otro grupo pueden ser iguales o diferentes. No es necesario que todos los elementos tengan una misma probabilidad de selección aunque se debe conocer la probabilidad que corresponde a cada uno. Por lo general todos los elemen tos que forman parte de un estrato dado tienen probabilidades de selección iguales. Si bien no toda combinación de elementos es posible, todas las muestras posibles (es decir, combinación de elementos) que se pueden extraer tienen la misma probabilidad de ocurrir.

2. SIMBOLOS

usamos los mismos símbolos propuestos para el muestreo simple al azar excepto que agregamos un subíndice para indicar el estrato al que se refiere la información. Así, N representa el número total de elementos de la población, como anteriormente; N_1 representa el número de elementos del primer estrato; N_2 del segundo, etc. En forma similar, $\bar{X}$ es el promedio verdadero de una característica; $\bar{X}_1$ el promedio verdadero de esa característica en el primer estrato; $\bar{X}_2$ en el segundo estrato, etc. El subíndice i se utiliza para in-

1 Una excepción aparece tratada en la sección 4 de la conferencia 10

dicar términos que se refieren a uno cualquiera de los estratos (el i- ésimo estrato). Para los valores "verdaderos" o poblacionales (distintos de los muestrales o estimaciones) los símbolos son:

N.... Número total de elementos de la población

N_i... Número de elementos del i-ésimo estrato

X.... Total poblacional para una cierta variable (característica)

X_i... Total poblacional para la misma variable en el i-ésimo estrato.

$\bar{X}$.... Promedio de la variable en la población (promedio para todos los estratos combinados).

$\bar{X}_i$... Promedio de la variable en i-ésimo estrato.

P.... Proporción de casos en la población que poseen cierta característica.

P_i... Proporción de casos en el i-ésimo estrato que poseen esa característica

S^2... Variancia de la variable en la población total

S^2_i... Variancia de la variable en el i-ésimo estrato

K.... Número de estratos.)Por lo tanto, i toma los valores 1 a K en lugar de 1 a n ó de 1 a N como en las conferencias anteriores).

2.1 Ejemplo para la población completa

Supongamos que tenemos un universo de 8 fincas de las cuales conocemos el valor de la tierra y los edificios, a saber:

Finca	Valor de la tierra y los edificios
A	$2026
B	6854
C	1532
D	2180
E	5408
F	9284
G	1438
H	8836

Calculemos el promedio (media) y la desviación estándar de esos valores. Usando los símbolos dados antes tendremos:

$N = 8$

$\bar{X} = \$4,694.75$

$S = \$3,111.00.$

Agrupemos las fincas en dos estratos de modo que los valores en cada uno sean:

Estrato 1	Estrato 2
$1438	$5408
1532	6854
2026	8836
2180	9284

Calculemos el promedio y la desviación estándar en cada grupo de 4 fincas separadamente tendríamos:

Estrato 1	Estrato 2
$N_1 = 4$	$N_2 = 4$
$\bar{X}_1 = \$1,794$	$\bar{X}_2 = \$7,595.50$
$S_1 = \$315$	$S_2 = \$1,559$

2.2 NOTACION PARA LAS ESTIMACIONES MUESTRALES

Al igual que en el muestreo simple al azar, se usan letras minúsculas para los totales muestrales el mismo tipo de los valores dados antes. Así, n será para indicar el tamaño total de la muestra; n_i para el tamaño de la muestra en el i-ésimo estrato; $\bar{x}$ para el promedio muestral por elemento en el mismo estrato, etc. Igualmente se utilizará una forma simple, por ejemplo, x') para indicar las estimaciones de los valores poblacionales obtenidas en la muestra.

3. ESTIMACIONES CON UNA MUESTRA ESTRATIFICADA

La media de la población se puede expresar en función de los totales de los estratos en la forma siguiente:

1) $$\bar{X} = \frac{X_1 + X_2 + X_3 + \ldots + X_k}{N} = \frac{\sum X}{N}$$

Como cada X_1 se puede expresar mediante $N_i \bar{X}_i$; podemos también escribir:

2) $$\bar{X} = \frac{1}{N} \sum_{1}^{k} N_i X_i$$

Dentro de cada estrato se utiliza el muestro simple al azar. Hemos visto antes que en el muestreo simple al azar $\bar{x}$ es una estimación insesgada de $\bar{X}$. Esto sugiere que en el muestreo estratificado se puede obtener una estimación de la media poblacional sustituyendo cada media del estrato por la correspondiente estimación dada por la muestra. Es decir, la media de los elementos muestrales del primer estrato da una estimación de la media verdadera del primer estrato; la media de los elementos muestrales del segundo estrato da una estimación de la media verdadera del segundo estrato, etc. En símbolos, por lo tanto, la estimación de la media poblacional con una muestra estratificada es:

(7.3) $$\bar{X}' = \frac{1}{N} \sum_{1}^{k} N_i \bar{x}_i$$

Otra forma de expresar esta misma fórmula es:

(7.4) $$\bar{X}' = \frac{1}{N} \sum_{1}^{k} \frac{N_i}{n_i} x_i$$

donde x_i es el total muestral en el i-ésimo estrato.

3.1 Ejemplo de estimación media

Se extrae una muestra estratificada de una población de 1.000 fincas para estimar el gasto promedio por productor agropecuario en mano de obra contratada. Existen tres estratos; en el primero el número total de fincas es 300; en el segundo, también 300; y en el tercero 400. Las muestras seleccionadas en cada uno de los tres estratos tienen, respectivamente 30, 30 y 40 fincas, el gasto promedio del primer estrato es \$12.20; en las 30 fincas del segundo estrato, \$25.60; y en las 40 fincas del tercer estrato, \$48.70. La estimación muestral del gasto promedio en el total de fincas en la población sería

$$\bar{X} = \frac{N_1 X_1 + N_2 \bar{X}_2 + N_3 \bar{X}_3}{N}$$

$$= \frac{300(12.20) + 300(25.60) + 400(48.70)}{1.000}$$

$$= \frac{3.660 + 7.680 + 19.480}{1.000} = \$30.82$$

3.2 Estimación del total

Al igual que en el muestreo simple al azar, calculamos una estimación del total poblacional multiplicando la estimación de la media por el número total de elementos en la población:

$$(7.5)\qquad \left.\begin{aligned} X' &= N\,\bar{x}' \\ &= \sum_{i}^{k} N_i\,\bar{x}_i \\ &= \sum \frac{N_i}{n_i}\,x_i \,. \end{aligned}\right.$$

3.3 Estimación de una proporción

Para estimar una proporción en la población se sigue un procedimiento similar al usado para la media:

$$(7.6)\qquad p' = \frac{1}{N}\sum N_i P_i$$

4. ERROR ESTANDAR DE UNA MUESTRA ESTRATIFICADA

Los errores estándar de los tres tipos de estimaciones mencionadas antes se calculan usando la Ecuación (7.7) para la media; la Ecuación (7.8) para el total; y la Ecuación (7.9) para una proporción:

$$(7.7)\qquad S_{\bar{x}'} = \sqrt{\frac{1}{N^2}\sum_{i}^{k}\left(\frac{N_i-n_i}{N_i-1}\right)N^2{}_i\left(\frac{S^2{}_i}{n_i}\right)}$$

$$(7.8)\qquad S_{y'} = N S_{\bar{x}'}$$

$$(7.9)\qquad S_{p'} = \sqrt{\frac{1}{N^2}\sum_{i}^{k}\left(\frac{N_i-n_i}{N_i-1}\right)N^2{}_i\left(\frac{P_i\,Q_i}{n_i}\right)}$$

Nótese que, en el muestreo estratificado, el número de elementos en un estrato (N_i) puede ser bastante pequeño. Por lo tanto es mejor mantener en el denominador de las Ecuaciones (7.7) y (7.9) el término $(N_i - 1)$ a menos que se sepa que $\frac{N_i}{N_i-1}$ es un número próximo a la unidad.

Para el error relativo (V) se pueden derivar fórmulas similares dividiendo las expresiones anteriores por el valor del rubro que se está estimando. Así, por ejemplo,

$$V_{\bar{x}'} = \frac{s_{\bar{x}'}}{\bar{x}}$$

4.1 Ejemplo

Apliquemos la ecuación (7.7) al caso de las 8 fincas del ejemplo de la sección 2. Supongamos que tomamos de las 8 fincas una muestra de 2 fincas - una en cada estrato - y calculamos $\bar{x}$ mediante la ecuación (7.3). ¿Cuál es el error estándar de $\bar{x}'$ como estimación de $\bar{X}$?.

Los valores en los dos estratos serían:

Estrato 1	Estrato 2
$N_1 = 4$	$N_2 = 4$
$N_1^2 = 16$	$N_2^2 = 16$
$n_i = 1$	$n_2 = 1$
$s_1 = 315$	$s^2 = 1,559$
$s_1^2 = 99,550$	$s_2^2 = 2,431,213$

Aplicando la ecuación (7.7)

$$s_{\bar{x}} = \sqrt{\frac{1}{N^2}\left[\left(\frac{N_i - n_i}{N_i - 1}\right)N_i^2\left(\frac{s_1^2}{n_1}\right) + \left(\frac{N_2 - n_2}{N_2 - 1}\right)N_2^2\left(\frac{s_2^2}{n_2}\right)\right]}$$

$$s_{\bar{x}} = \sqrt{\frac{1}{64}\left[\left(\frac{4-1}{4-1}\right)16\left(\frac{99,550}{1}\right) + \left(\frac{4-1}{4-1}\right)16\left(\frac{2,431,213}{1}\right)\right]}$$

$$= \sqrt{\frac{1}{64}\left(1,592,800 + 38,899,408\right)}$$

$$= \sqrt{632,690,75}$$

$$= \$795.40.$$

Es interesante comparar este error estándar con el correspondiente error de la media en una muestra simple al azar de 2 fincas. Con una muestra

Simple al azar de 2 fincas tendríamos:

$$S_{\bar{x}'} = \sqrt{\left(\frac{N-n}{N-1}\right)\frac{s^2}{n}}$$

$$= \sqrt{\left(\frac{8-2}{8-1}\right)\frac{(3,111)^2}{2}} = \$2,037.$$

En este ejemplo el error estándar de la muestra estratificada es mucho más pequeño que el de una muestra simple al azar, en realidad, casi la mitad. En términos más precisos, se necesitaría, usando el muestreo simple al azar, una muestra de 6 fincas para alcanzar la misma confiabilidad (es decir, un error estándar tan pequeño) que la obtenida con una muestra estratificada de 2 fincas.

4.2 Comentarios

En la práctica, por lo general no conocemos los valores verdaderos de S_i^2 y de P_i Q_i. En su lugar sustituimos, en las Ecuaciones (7.7), (7.8), (7.9) dichos valores por estimaciones muestrales a fin de obtener estimaciones de $S_{\bar{x}'}$, etc. Para calcular tales estimaciones con una muestra necesitaríamos, al menos, dos elementos de cada estrato. (En los ejemplos descritos antes, pudimos calcular el error estándar de muestras que tenían un solo elemento por estrato debido a que disponíamos de información acerca de todos los elementos en el universo).

TAREA DE ESTUDIO

Problema A: Usted tiene una población de 12 personas cuyos ingresos por hora son:

A - $ .85	E - $1.80	I - $1.75
B - $1.35	F - $3.10	J - .75
C - $ .60	G - $ .90	K - $2.40
D - $2.20	H - $1.50	L - $2.10

Ejercicio 1 . ¿Cuáles son los ingresos promedios por hora de este grupo?.

Ejercicio 2 . Cuál es el error estándar de la media de una muestra de 4 personas seleccionada como una muestra simple al azar.?

Ejercicio 3. Estratifique esta población en tres estratos de igual tamaño en la mejor forma posible para estimar los ingresos promedio. Liste las personas en cada estrato de acuerdo con sus ingresos por hora.

Ejercicio 4. Seleccione una muestra de 4 personas una en cada uno de los dos estratos con ingresos bajos y dos en el otro estrato:

a) Indique la fórmula que usaría para estimar los ingresos promedio por hora para esta muestra.

b) Indique la fórmula para el error estándar de la media estimada.

Problema B. Refiérase al universo de 8 fincas (sección 2.1 del texto) en el que los valores de la tierra y los edificios son:

A- $2026	E- $5408
B- $6854	F- $9284
C- $1532	G- $1438
D- $2180	H- $8836

Ejercicio 5. Identifique las 28 combinaciones de muestras posibles de 2 Fincas en un muestreo simple al azar (AB, AC, etc.):

a) Calcule la media de cada muestra y verifique que la media de las 28 medias es $4.694,75.

b) Calcule la desviación estándar de las 28 medias y verifique que la desviación estándar es $2.037.

Ejercicio 6. Identifique las 16 combinaciones de muestras posibles de 2 fincas en un muestreo estratificado usando los dos estratos (estrato 1 y estrato 2) que figuran en la sección 2.1 del texto :

a) Verifique que la media de las 16 medias es $ 4,694,75.

b) Verifique que la desviación estándar de las 16 medias es $ 795.40.

CONFERENCIA * 8. MUESTREO ESTRATIFICADO - AFIJACION EN LOS ESTRATOS

1. EL PROBLEMA DE LA AFIJACION

La definición del muestreo estratificado no especifica para la muestra en cada estrato un tamaño determinado. Se puede seleccionar la muestra de modo que en cada estrato tenga el mismo tamaño o distribuir el tamaño total en alguna otra forma. En tanto se seleccione al menos un elemento por estrato se satisface la especificación de una muestra estratificada. A su vez, con 2 elementos por estrato ya se puede estimar tanto la media como su error. Por lo general el tamaño total de la muestra es mucho mayor que dos elementos por estrato. Por lo tanto surge la necesidad de establecer un criterio para afijar el tamaño total de la muestra en los estratos.

Volvamos al ejemplo de una población de 8 fincas agrupadas en 2 estratos. Si deseamos seleccionar una muestra de 2 fincas para estimar la media, el único camino posible es tomar en cada estrato una finca. Si en cambio deseamos seleccionar 4 fincas nos encontraremos frente a una alternativa. ¿Qué sería mejor? ¿seleccionar 2 fincas en cada estrato o tomar una en un estrato y 3 en el otro?.

Para determinar la distribución de la muestra entre los distintos estratos existen dos criterios principales. El primero es la conveniencia, es decir, elegir un procedimiento que sea fácil de aplicar y simple para tabular. Este criterio nos conduce, por lo general, al muestreo proporcional. El segundo criterio es la exactitud: Elegir un procedimiento que proporcione el error estándar más pequeño. Esto nos lleva al uso de la afijación óptima.

2. MUESTREO ESTRATIFICADO PROPORCIONAL

En el muestreo estratificado es muy común seleccionar en cada estrato la misma proporción de elementos. Según este procedimiento para seleccionar una muestra del 10 por ciento de una cierta población tomaríamos en cada extremo una muestra del 10 por ciento de dicho estrato.

En este caso, dado que las tasas de muestreo son iguales en todos los estratos, el número de elementos tomados en cada estrato para la muestra variará de un estrato a otro dependiendo del tamaño del estrato. Dentro de cada estrato el tamaño de la muestra será proporcional a la población total del estrato. Matemáticamente se puede expresar esta situación en la forma siguiente:

$$\frac{n_i}{n} = \frac{N_i}{N} \quad \text{ó sino} \quad n_i = \frac{N_i}{N} n$$

Tratándose de las características de la población en las que usualmente estamos interesados (es decir, $\bar{X}$ y Y), podemos preparar estimaciones con una muestra estratificada proporcional tan fácilmente como si lo hiciéramos con una muestra simple al azar en realidad con la misma fórmula:

$$(8.1) \qquad \bar{x}' = \frac{1}{n}\left(\sum_{1}^{n} x_i\right)$$

En esta fórmula la suma se refiere a todos los elementos muestrales sin prestar atención a los estratos. Como $\frac{N_i}{n_i}$ es constante e igual a $\frac{N}{n}$, la ecuación (7.3) de la Conferencia 7 queda reducida a esta expresión.

Tenemos también:

(8.2) $$X' = N\bar{x}' = \frac{N}{n}\sum^{n} x_j.$$

En las ecuaciones (8. 1) y (8.2) el subíndice j se refiere a las observaciones individuales ya que en el muestreo estratificado hemos reservado el subíndice i para indicar estratos.

El procedimiento de ponderación simple hace que el muestreo proporcional sea muy conveniente dado que los resultados son fáciles de tabular. No es necesario tabular cada estrato separadamente sino que se pueden sumar en forma conjunta todos los datos muestrales y luego aplicar un cierto factor que podría ser $\frac{1}{n}$ o $\frac{N}{n}$. Se dice que una muestra con esta característica está autoponderada. El error estándar de la media estimada con una muestra estratificada proporcional es

(8.3) $$S_{\bar{x}}' = \sqrt{\left(\frac{N-n}{N-n}\right)\frac{1}{N}\sum N_i s_i^2}.$$

Esta fórmula es una aproximación de la fórmula verdadera obtenida asumiendo que $\frac{N_i}{N_i-1}$ es casi igual a la unidad en cada estrato. (alternativamente se puede emplear en lugar de s_i^2 el valor $s_i^2 = \frac{N_i}{N_i-1} s_i^2$, quedando así la ecuación (8.3) transformada en una fórmula exacta daría, por lo general, los mismos resultados.

3. AFIJACIÓN OPTIMA

En el año 1934, Jerzy Neyman investigó matemáticamente el problema de cuál podía ser la distribución de la muestra en los estratos que diera el menor error de muestreo posible, Encontró que la respuesta consistía en dejar que la tasa de muestreo en cada estrato variara con la cantidad de variabilidad de cada estrato- en otras palabras, hacer la tasa de muestreo es un estrato dado proporcional a la desviación estándar en ese estrato. En esa forma el número de elementos a extraer para la muestra en cada estrato dependería no sólo del número total de elementos en el estrato sino también de la desviación estándar de la característica que se va a medir. Para esta afijación óptima el número de elementos que se selecciona en un estrato está dado por la fórmula:

(8.4) $$n_i = n\frac{N_i s_i}{\sum N_i s_i}$$

Con una afijación óptima el error estándar de la media se reduce a:

$$(8.5) \qquad S_{\bar{x}}' = \sqrt{\frac{1}{n}\left(\frac{N_i S_i}{N}\right)^2 - \frac{N_i S_i^2}{N^2}}$$

Para aplicar este tipo de afijación es necesario conocer los valores de S_i en el universo. Si no se conocen, se pueden estimar dentro de cada estrato usando los procedimientos descri-tos en la sección 8 de la conferencia 6.

3.1 Ejemplo

Comparemos los errores estándar que se obtienen en una misma encuesta con la afijación proporcional y con la afijación óptima. En 1942 se levantó un censo de producción de madera aserrada en los Estados Unidos. En 1943 se repitió la misma investigación mediante una encuesta, Antes de seleccionar la muestra se agruparon los aserraderos en estratos de acuerdo con la producción en 1942. Un análisis de los datos mostró la información presentada en el Cuadro 8A.

CUADRO 8A. DATOS BASICOS PARA DETERMINAR LA AFIJACION OPTIMA

(Producción y desviaciones estándar dadas en mil pies de tabla)

Estrato	Producción anual	Número de aserraderos (N_i)	Producción promedio en el estrato	Desviación estándar en 1943 (S_i)
1	5,000 y más	538	11,029.7	9,000
2	1,000 a 4,999	4,756	1,779.6	1,200
3	Menos de 1,000	30,964	203.8	300
Total		36,258	571.2	..1,684

. Estimada con los datos de 1942
.. En el muestreo no estratificado.

Seleccionemos una muestra de 1,000 aserraderos.
El primer punto es considerar cómo determinar el tamaño de la muestra en cada estrato, ya sea con el muestreo proporcional o con el muestreo con

afijación óptima. El segundo es considerar cuál es la confiabilidad de los dos métodos. Atendamos primero el problema del tamaño de la muestra y luego el de la confiabilidad.

3.2 Tamaño de la muestra en cada estrato

En la afijación proporcional dado que la tasa de muestreo es 1,00 en 32,258, se deberá aplicar esta tasa en cada estrato. Por lo tanto los tamaños de muestra serían:

$$n_1 = \frac{1,000}{36,258} \times 538 = 15$$

$$n_2 = \frac{1,000}{36,258} \times 4,756 = 131$$

$$n_3 = \frac{1,000}{36,258} \times 30,964 = 854.$$

En la afijación óptima, los tamaños de muestra en cada estrato se determinarían de acuerdo con el cuadro siguiente:

CUADRO 8B TAMAÑO DE MUESTRA EN LA AFIJACION OPTIMA

Estrato	Número de aserraderos (N_i)	Desviación estándar(S_i)	$N_i S_i$	$\frac{N_i S_i}{\sum N_i S_i}$	Número en la muestra	Tasa de muestreo
1	538	9.000	4.841.000	0.244	244	1/2
2	4.756	1.200	5.707.200	0.288	288	1/15
3	30.964	300	9.289.200	0.468	468	1/60
Total	36.258		19.838.400	1.000	1.000	

* $n_i = 1.000 \times \frac{N_i S_i}{\sum N_i S_i}$

3.3 Errores estándar

¿Cuáles son los errores estándar en los dos diseños de muestra?. En la afijación proporcional el error estándar de la estimación de la media está dado por la Ecuación (8.3).

$$S_{\bar{x}'} = \sqrt{\left(\frac{N-n}{Nn}\right)\frac{\sum N_i S_i^2}{N}}$$

En la encuesta de la producción de madera aserrada,

$$\sum N_i S_i^2 = 538\,(9.000)^2 + 4,756(1.200)^2 + 30.964(300)^2$$

$$= 53.213.400.000$$

y

$$S_{\bar{x}'} = \sqrt{\frac{36.258-1.000}{36.258(1.000)} \times \frac{53.213.400.000}{36.258}}$$

$$= \sqrt{1427} = 37.8 \text{ (mil pies de tabla).}$$

En la afijación óptima el error estándar correspondiente está dado por la Ecuación (8.5):

$$S_{\bar{x}'} = \sqrt{\frac{1}{n}\left(\frac{\sum N_i S_i}{N}\right)^2 - \frac{\sum N_i S_i^2}{N^2}}$$

$$= \sqrt{\frac{1}{1000}\left(\frac{19.838.400}{36.258}\right)^2 - \frac{53.213.400.000}{(36.258)^2}}$$

$$= \sqrt{259} = 16.1 \text{ (mil pies de tabla).}$$

Para completar el análisis podemos comaprar esos resultados con los que hubiéramos obtenido si no se hubieran estratificado los aserraderos sino que se hubieran tomado una muestra simple al azar de 1.000 aserraderos en la población. En este caso,

$$S_{\bar{x}'} = \sqrt{\left(\frac{N-n}{N-1}\right)\frac{S^2}{n}}$$

$$= \sqrt{\left(\frac{36.258-1.000}{36,257}\right)\frac{(1.684)^2}{1.000}}$$

$$= 52.2 \text{ (mil pies de tabla)}$$

4. COMPARACION DE LOS ERRORES DE MUESTREO EN LOS DISTINTOS METODOS DE MUESTREO

Examinando los resultados de los diseños de muestreo anteriores vemos que la afijación óptima de muestreo nos dió 16.1 mil pies de tabla, error

estándar considerablemente más pequeño que el del muestreo proporcional, que fué 37.8. Vemos asimismo que el error de muestreo del muestreo proporcional fué más pequeño que el del muestreo simple al azar, que fué 52.2. Expresando estos mismos resultados en otra forma diríamos que se necesita una muestra proporcional más de 5 veces mayor que una muestra con afijación óptima para alcanzar la misma confiabilidad. Con el muestreo simple al azar se necesitaría una muestra 10 veces mayor. La eficiencia de la afijación óptima se deriva del hecho de que proporciona un muestreo más intenso en los estratos que tienen una desviación estándar mayor que son los que, es de esperar, contribuyen en una medida más acentuada al error total de muestreo.

El ejemplo de la sección 3 anterior ilustra un resultado general que puede demostrarse matemáticamente. Los errores de muestreo de los tres tipos de diseños están aproximadamente relacionados en la forma siguiente (siempre que las tasas de muestreo sean lo suficientemente pequeñas como para poder ignorar los multiplicadores finitos):

(8.6)
$$S^2_{\text{al azar}} = S^2_{\text{óptimo}} + \frac{\sum N_i (S_i - \bar{S})^2}{nN} + \frac{\sum N_i (\bar{X}_i - \bar{X})^2}{nN}$$

(8.7)
$$= S^2_{\text{Proporcional}} + \frac{\sum N_i (\bar{X}_i - \bar{X})^2}{nN}$$

Donde $\bar{S}$ es el promedio ponderado de los valores de S_i; es decir:

$$\bar{S} = \frac{\sum_{k} N_i S_i}{N}$$

Un exámen de esta fórmula muestra que los errores de muestreo obtenidos con la afijación óptima serán al menos tan pequeños, generalmente más pequeños, que los obtenidos con el muestreo estratificado proporcional. Además, los errores obtenidos con uno u otro de los dos métodos serán al menos tan pequeños, generalmente más pequeños, que los obtenidos con el muestreo simple al azar. (Existen unos pocos casos, que casi nunca ocurren en la práctica, en que ésto no es cierto. Cuando la muestra es muy pequeña y la estratificación es completamente ineficaz, ni el muestreo proporcional ni la afijación óptima pueden superar el muestreo simple al azar. Esta posibilidad puede ignorarse en casi todas las situaciones reales.)

Consideremos las condiciones por las que surgen estas importantes diferencias entre los tres procedimientos. Al comparar el muestreo estratificado proporcional con el muestreo simple al azar se puede demostrar que la ganancia de confiabilidad depende de la magnitud de la variación de las me-

dias (en otras palabras, cuanto mayor sea la diferencia entre los estratos) mayor será la reducción del error estándar resultante del uso del muestreo proporcional. Por otra parte, si la variancia entre las medias de los estratos es bastante pequeña comparada con la variancia total no se obtendrá mucha ganancia con la estratificación. En consecuencia, la estratificación es por lo general menos importante cuando se trata de proporciones en lugar de rubros medidos(o de totales o cantidades). Por ejemplo, sería más útil al tratar de estimar los gastos promedio de los productores agropecuarios en mano de obra contratada que para estimar la proporción de productores agropecuarios que emplean mano de obra contratada. Aun para rubros que se expresan en medidas las ganancias serían pequeñas a menos que los estratos se establezcan de modo que las diferencias entre las medias alcancen una magnitud medible (como ocurrió en el ejemplo de los aserraderos). Por ejemplo, en una encuesta para medir ingresos personales, no valdría la pena establecer estratos separados para los diferentes grupos profesionales-por ejemplo, médicos, abogados, etc. Quizá sería más conveniente, en cambio, establecer estratos separados para grupos más generales -obreros, hombres de negocios, profesionales, etc. Como casi siempre el muestreo proporcional es mejor que el muestreo simple al azar, se recomienda usar la estratificación siempre que se pueda efectuar sin mayor trabajo adicional.

Comparando la afijación óptima con la proporcional vemos que si las desviaciones estándar en todos los estratos son iguales los dos métodos resultan idénticos. Cuanto mayor sea la diferencia entre las desviaciones estándar en los estratos mayor será la reducción del error de muestreo que se puede esperar de la afijación óptima. Al menos que las desviaciones estándar tengan una amplitud de variación mayor de 2 ó 3 a 1, las ganancias de la afijación óptima son tan pequeñas que probablemente no justifican el trabajo extra que implica en las tabulaciones. Si existen variaciones mayores en las desviaciones estándar las ganancias son apreciables y conviene aplicar una afijación óptima. En el ejemplo de los aserraderos la desviación estándar en el estrato 1 fué 30 veces mayor que en el estrato 3.

Para (a) aplicar la afijación óptima, o (b) estimar los errores de muestras estratificadas proporcionales, necesitamos conocer en cada estrato los valores de S_i. Por supuesto, en la práctica, nunca conocemos realmente los valores de S_i en cada estrato y debemos por lo tanto estimarlos. Surgen así dos cuestiones: (a) ¿Qué efecto tienen, sobre la exactitud de la muestra, los errores introducidos por el hecho de que se usan estimaciones de S_i en lugar de sus valores vardaderos?; (b) ¿Qué métodos pueden aplicarse para estimar esas cantidades?.

La respuesta a la primer pregunta es que si nuestras estimaciones de la desviación estándar son suficientemente razonables (digamos, exactas, por ejemplo, dentro de un 30% o un 40%), obtendremos casi la ganancia total de la afijación óptima. La razón de ésto es que el error de muestreo no aumenta muy rápidamente cuando la afijación se aparta del óptimo dentro de límites suficientemente amplios. (Debe señalarse que una conjetura pobre acerca de los valores de S_i no introduce ningún sesgo en los resultados; sólo

aumenta los errores de muestreo.) Sin embargo, si las estimaciones de S_i son muy poco confiables, la "afijación óptima" puede tener una variancia más grande que el muestreo proporcional. En este caso es más seguro usar el muestreo proporcional.

Con respecto a la segunda pregunta, podemos usar lo métodos descriptos antes para estimar las desviaciones estándar (sección 8 de la Conferencia 6.) Otro método que se suele utilizar algunas veces es suponer que medios en los estratos: es decir, suponer la misma desviación estándar relativa en cada estrato. (Nótese que en la afijación óptima no es necesario conocer los valores absolutos de las desviaciones estándar; sólo se requiere conocer los valores relativos entre unos y otros.) Esta hipótesis dará, con frecuencia, resultados razonablemente próximos al óptimo. En el caso de los aserraderos ya tratado antes, nos daría una muestra con la siguiente distribución por estrato:

$$n_1 = n\frac{N_i\bar{x}_i}{\sum N_i x_i} = 1.000 \times \frac{538(11.030)}{20.710.570} = 287$$

$$n_2 = n\frac{N_2\bar{x}_2}{\sum N_i\bar{x}_i} = 1.000 \times \frac{4.756(1.780)}{20.710.570} = 409$$

$$n_3 = n\frac{N_3\bar{x}_3}{\sum N_i x_i} = 1.000 \times \frac{30,964(204)}{20.710.570} = 305.$$

Puede verse que esta afijación es mucho más próxima a la óptima que la proporcional. En realidad, si se calcula el error estándar de esta afijación resulta 17.3. No es un valor tan bueno como 16.1 para la afijación óptima pero muy superior al valor 37.8 obtenido con el muestreo proporcional.

5. AFIJACION OPTIMA CON COSTOS VARIABLES

Hasta ahora la discusión de la afijación óptima se ha hecho en función de la obtención de los resultados más confiables para un tamaño dado de muestra. Con frecuencia los costos de obtención de la información varían substancialmente de un estrato a otro. Por ejemplo, supongamos que se han estratificado las familias según residencia urbana o rural; supongamos además que el costo de efectuar una entrevista en la zona rural es 5 veces mayor que en la zona urbana. Lo apropiado sería concentrar más la muestra en el estrato más económico. Otro ejemplo sería una encuesta por muestra de firmas comerciales; podemos enviar los cuestionarios por correo a las compañías pequeñas y visitar personalmente a las grandes cuando son mar-

cadas las diferencias en los costos por unidad.

Un enfoque más general que el descrinto en la sección 4 es considerar la afijación óptima para un costo fijo, en lugar de para un tamaño de muestra fijo. En otras palabras, trataríamos de afijar la muestra entre los estratos en forma tal que obtuviéramos el error estándar más bajo con un presupuesto fijo.

Para ésto necesitaríamos una finción del costo, la cual es una formulación matemática que expresa el costo de levantar la encuesta en función de los tamaños de muestra n_i. Supongamos que el costo promedio por cuestionario en el i-ésimo estrato es C_i. Así, C_i es el costo por cuestionario en el i-ésimo estrato incluyendo el costo de la entrevista, la codificación y perforación, etc. (Puede existir además un costo general de la encuesta que no depende del tamaño de la muestra. No necesitamos tratarlo aquí.) El costo total de la encuesta que puede ser afectado por el tamaño de la muestra es:

$$C = C_1 n_1 + C_2 n_2 + C_3 n_3 + \dots + C_k n_k = \sum_{i}^{k} C_i n_i .$$

Para un costo fijo C_i la afijación óptima de la muestra resulta.[1]

(8.8)
$$n_i = n \times \frac{\frac{N_i S_i}{\sqrt{C_i}}}{\sum \frac{N_i S_i}{\sqrt{C_i}}}$$

Es decir, n_i es directamente proporcional a N_i y C_i, e indirectamente proporcional a $\sqrt{C_i}$.

La fórmula (8.8) lleva a las siguientes reglas:

En un estrato dado, se toma una muestra más grande si:

1) El estrato es más grande que el estrato promedio;
2) El estrato es más variable que el estrato promedio;
3) El costo de obtención y elaboración de la información es menor en ese estrato que en el estrato promedio.

Con respecto al tercer punto, el costo por estrato (C_i) interviene en la fórmula como una raíz cuadrada, Esto tiende a reducir el efecto de las diferencias en el costo unitario. A menos que los costos varíen en un factor de al menos, 2 a 1, el uso de la fórmula anterior dará resultados no muy distintos de la afijación óptima más sencilla espresada en la Ecuación (8.4).

---

1 Para usar esta fórmula debe calcularse primero N. Dicho valor es una función de C y de los C_i, c_i y N_i. Véase Simple Survey Methods and theory-Vol I Methods and Applications por Hansen, M.H. Hurwitz., W.N., y Medow, W.C. John Wiley and Sons, New York 1953 p. 271.

6. AFIJACION OPTIMA PARA VARIOS RUBROS

La fórmula de la afijación óptima de la muestra (Ecuación 8.4 o Ecuación (8.8) está calculada, necesariamente, para una sola característica o variable, es decir X. Si se desea obtener la afijación de la muestra más favorable para varias características, se debe llegar a una cierta clase de compromiso.
Algunas alternativas son:

1) Determinar el rubro más importante (o grupo de rubros altamente correlacionados) y afijar la muestra para obtener su mejor estimación.

2) Seguir el procedimiento (1) y aumentar el tamaño de la muestra en algunos estratos para proporcionar la cobertura adecuada de otros rubros importantes.

3) Establecer una función que asigne una ponderación a cada rubro de acuerdo con su importancia; usar esa función en la afijación para evitar estimaciones muestrales pobres para las características de máxima im - tancia.

La afijación óptima es más efectiva para características que varían marcadamente en las unidades individuales, como podrían ser el ingreso personal, número de pies de tabla producidos en un aserradero, kilogramos de maíz cosechados en una finca, etc.

Sin embargo, en el muestreo para atributos tales como la proporción de población en una clase (por ejemplo, en la clase de ingresos entre $1.000 y $1.999), la mejor afijación puede ser la del muestreo proporcional. Este procedimiento tiene la ventaja adicional de ser autoponderada.

Problema: Se desea estimar el valor total de los productos agropecuarios para una población de 5,900 fincas. Por un censo anterior sobre el valor de los productos agropecuarios se conocen las medias y las variancias clasificadas según tamaño de las fincas y tenencia del productor.

---

2 Véase la sección 13 del capítulo 5 de Sample Survery Methods and Theory, Vol I. Methods and Applications (obra mencionada en la nota al pie # 1 de la Conferencia 8, para una exposición más amplia del problema del muestreo para varias características.

Tamaño y tenencia	Número de fincas (N_i)	Valor promedio de los productos ($\bar{X}_i$)	Variancia (S^2_i)	Desviación estándar (S_i)	$N_i S_i$
Todas las fincas	5,900	$3,500	97.000.000	$9.840	
TAMAÑO DE LA FINCA					
Menos de 10 Acres.	590	1,200	18,000,000	4,240	2,502,000
10 a 49 acres..	1,600	1,500	15,000,000	3,870	6,192,000
50 a 99 acres..	1,,50	2,200	18,000,000	4,240	4,876,000
100 a 179 acres	1,200	3,600	35,000,000	5,920	7,104,000
180 a 259 acres	490	5,500	70,000,000	8,370	4,101,000
260 a 999 acres	650	6,200	200,000,000	14,150	9,198,000
1,000 acres y más	220	18,000	400,000,000	20,000	4,400,000
Suma de productos			*349,620,000,000		38,373,000
REGIMEN DE TENENCIA					
En propiedad total.....	3,300	2,600	35,000,000	5,920	19,536,000
En co-propiedad	660	6,900	110,000,000	10,490	6,923,000
Por administración.....	50	18,000	510,000,000	22,580	1,129,000
Por arrendamiento.....	1,890	3,500	40,000,000	6,320	11,945,000
Suma de productos.....			*289,200,000,000		39,533,000

* Suma de productos = $\sum N_i S_i^2$

Ejercicio 1. Calcular el error estándar del valor total de los productos con una muestra estratificada proporcional de 300 fincas de acuerdo con cada uno de los dos métodos de estratificación (según tamaño y según régimen de tenencia).

Ejercicio 2. ¿Cuál método de estratificación es más eficiente con una muestra proporcional?.

Ejercicio 3. Calcular el error estándar de la estimación del valor total de los productos usando una muestra simple al azar de 300 Fincas.

Ejercicio 4. Usar, con ambos métodos de estratificación, la afijación óptima para una muestra de 300 fincas,y calcular:

a) El número de fincas muestrales en cada estrato.

b) El error estándar de la estimación del valor total de los productos.

Ejercicio 5. Sobre la base de éste análisis, ¿Cuál de los cinco métodos de afijación de la muestra usted recomendaría.

Ejercicio 6. Supongamos que la muestra se estratificó según régimen de tenencia y que la afijación se llevó a cabo mediante el método óptimo. Supongamos también que se obtuvieron las siguientes medias por estrato:

Régimen de tenencia	Valor medio de los productos
En propiedad total	$ 2,900
En co-propiedad	6,400
Por administración	20,000
Por arrendamiento	4,000

Estimar el valor medio de los productos en la población de 5.900 fincas.

Ejercicio 7. Describa la forma como usted calcularía el error estándar de la media calculada en el ejercicio 6 después de que se disponga de los resultados de la encuesta.

CONFERENCIA 9. MUESTREO DE CONGLOMERADOS

1. DESCRIPCION DEL MUESTREO DE CONGLOMERADOS

La exposición hasta este momento se ha referido a métodos de muestreo en los que las unidades de análisis (personas, fincas, firmas comerciales, etc.) se han considerado que estaban ordenadas en una lista (o su equivalente) y de la que se podía extraer directamente una muestra. Consideremos ahora un procedimiento de muestreo en el que las unidades de análisis en la población se consideran agrupadas en conglomerados y se selecciona una muestra de conglomerados (en lugar de una muestra de unidades individuales). Los conglomerados muestrales, por lo tanto, determinan las unidades que se incluyen en la muestra. Para esta determinación existen dos alternativas:

1) La muestra podría incluir todas las unidades en los conglomerados seleccionados. Se denomina, por lo general, este procedimiento como muestreo unietápico de conglomerados.

2) En los conglomerados seleccionados se podría seleccionar una submuestra de unidades y enumerar únicamente esa submuestra de unidades. Este es el muestreo polietápico de conglomerados o simplemente muestreo polietápico.

Existen dos razones principales para utilizar el muestreo de conglomerados. Con frecuencia no se dispone de un marco adecuado (como podría ser una lista) de donde seleccionar una muestra de los elementos de la población y el costo de construcción de dicho marco ser posiblemente elevado. En otros casos puede existir tal marco pero resultar más eficiente el muestreo de conglomerados (sobre cierta base geográfica) que el muestreo simple al azar dadas las economías que reporta el primero en lo que se refiere a los costos de las labores de campo. En muchas situaciones prácticas, una muestra de un cierto número de unidades seleccionadas al azar dará una variancia más pequeña que una muestra de igual tamaño seleccionada en conglomerados. Sin embargo, cuando se hace un balance entre el costo y la precisión, la muestra de conglomerados puede ser más eficiente.

Aun cuando las unidades en las que estamos interesados no se seleccionen directamente, la probabilidad de selección de un conglomerado y de cada unidad dentro del mismo (y, por lo tanto, de cada unidad en la población) está fijada de antemano.

En consecuencia, el muestreo de conglomerados satisface los criterios del muestreo probabilístico.

Consideramos algunos ejemplos que permiten ver como funciona el muestreo de conglomerados.

1.1 Muestreo unietápico de conglomerados

Para extraer una muestra de personas generalmente no sería factible obtener una lista de todas las personas y seleccionar luego una muestra de la lista. Se podría encontrar en cambio una lista de familias. Podríamos luego extraer una muestra de familias y obtener información mediante entrevista acerca de todas las personas en las familias seleccionadas. Este es un ejemplo de muestreo unietápico de conglomerados donde la familia constituye el conglomerado. Nótese que para un número dado de individuos en la muestra sería indudablemente menos costoso, en función tanto de los viajes como del tiempo, considerar todas las personas dentro de las familias seleccionadas que seleccionar al azar, entre todos los individuos que componen la población, el mismo número de personas.

Con frecuencia no se dispone de una lista de familias y es necesario seguir algún otro procedimiento. Uno de ellos podría ser el siguiente. En las grandes ciudades, por lo general, existen mapas donde están señalados los límites de las manzanas lo que permite seleccionar una muestra de manzanas. Para el resto del país se pueden utilizar mapas divididos en áreas pequeñas, que se denominan segmentos, con límites identificables: en ellos se puede seleccionar una muestra de segmentos. Se podría luego incluír en la muestra todas las personas dentro de las manzanas y los segmentos muestrales o, alternativamente, seleccionar una muestra de las personas que viven en las manzanas seleccionadas. La decisión dependerá del número de etapas de muestreo que

se considere más conveniente. Cuando se utilizan mapas se elimina la necesidad de la lista de todas las personas; se la reemplaza por una lista de manzanas y segmentos y una lista de familias dentro de una muestra de las manzanas y segmentos. (En la práctica existe, con frecuencia, una etapa previa de muestreo en la que se selecciona una muestra de ciudades y/u otras divisiones administrativas.) La exposición anterior ilustra una aplicación importante del muestreo de conglomerados: el muestreo de superficies O ÁREAS . Si embargo, el muestreo de conglomerados tiene también otras aplicaciones.

1.2 Muestreo polietápico de conglomerados

Supongamos que deseamos efectuar una muestra de la población escolar para obtener información acerca de su salud o acerca de su conocimiento en relación con un tema dado. Una forma de llevar a cabo esta encuesta es obtener una lista completa de escuelas, seleccionar luego una muestra de escuelas y, por último, elegir una muestra de niños en las escuelas seleccionadas. En forma similar, para tener una muestra de obreros fabriles podríamos seleccionar primero una muestra de fábricas y luego entrevistar una muestra de los obreros de las fábricas seleccionadas. En ambos casos necesitaríamos construír una lista de individuos sólo para las escuelas o fábricas seleccionadas en la muestra. Estos ejemplos ilustran el muestreo multietápico (específicamente, bietápico) de conglomerados. La probabilidad de que una unidad de la población resulte seleccionada en la muestra puede expresarse como el producto de las probabilidades en cada etapa. Así,en el primer ejemplo, la probabilidad de que resulte seleccionado el j-ésimo niño de la i-ésima escuela es igual a la probabilidad de que resulte seleccionada la i-ésima escuela multiplicada por la probabilidad condicional de que resulte seleccionado el j-ésimo niño suouesto que se haya seleccionado la i-ésima escuela es decir:

$$P(\text{j-ésimo niño, i-ésima escuela}) =$$

$$P(\text{i-ésima escuela}) \times P(\text{j-ésimo niño} \mid \text{i-ésima escuela}).$$

2. MUESTREO DE SUPERFICIES

Como el muestreo de superficies es una aplicación frecuentemente usada del muestreo de conglomerados describiremos con más detalles los métodos que se aplican usualmente. El muestreo de superficies es conveniente cuando se presentan una o ambas de las condiciones siguientes:

1) Cuando no se dispone de listas completas de las unidades de vivienda (o de otras unidades de observación deseables) pero existen mapas que incluyen una cantidad razonable de detalles. Tales mapas pueden considerarse como una lista que cubre todas las unidades de vivienda en la región.

2) Cuando al enviar un entrevistador de una unidad de vivienda seleccionada al azar a otra unidad de vivienda seleccionada al azar implica costos

elevados en viajes. Dada una cierta cantidad de dinero, podremos aumentar el número de unidades de vivienda en la muestra en gran medida agrupando las unidades y seleccionando una muestra al azar de los grupos.

Para extraer una muestra de superficies existen tres procedimientos simples. Como ejemplo utilizaremos las manzanas de una ciudad (podríamos también hacerlo con segmentos de tierra con límites bien identificables en las zonas rurales y proceder con dichos segmentos en igual forma que con las manzanas de una ciudad). Supondremos que se ha de extraer una muestra del uno por ciento de las unidades de vivienda.

Procedimiento A:

Para una muestra de superficie que se enumerará completamente:

1) Obtener un mapa razonablemente exacto de la ciudad que muestre el mayor número de detalles acerca de las manzanas. Si el mapa no es reciente, se deben tomar las medidas convenientes, a través de consultas en la misma localidad, para actualizarlo (por ejemplo, marcar nuevas calles que se hayan podido abrir desde la fecha en que se imprimió el mapa).

2) Numerar en serie las manzanas, anotando los números directamente en el mapa; se aconseja adoptar un sitema de numeración serpentino a fin de tener la seguridad de que no se omite alguna manzana.

3) Seleccionar una muestra simple al azar o una muestra sistemática de manzanas utilizando una muestra de uno por ciento. Si se extrae una muestra sistemática, seleccionar un número al azar entre 1 y 100 para determinar la primer manzana muestral e incluir en adelante cada centésima manzana.

4) Entrevistar todas las familias en las manzanas muestrales.

Procedimiento B:

Para una muestra de superficies con submuestreo de superficies más pequeñas:

La muestra del uno por ciento se puede también obtener extrayendo, por ejemplo, una muestra de 1 en 25 manzanas y luego seleccionando una submuestra de un cuarto de la superficie en cada manzana muestral.

1) Proceder como en (1). (2) y (3) del Procedimiento A anterior, salvo - que se toman en lugar de 1 manzana en cada 100, 1 en cada 25.

2) Dividir cada una de las manzanas muestrales en 4 segmentos. Si se cuenta con mapas que muestren la estructura interna de cada manzana (pasajes, edificios, etc.), pueden usarse los mismos. Si no es así, prepa - rar un bosquejo rápido y esquemático de las manzanas muestrales señalando cada edificio; usar ese bosquejo como base de la segmentación. Los 4 segmentos dentro de una manzana deben tener cada uno aproximadamente el

mismo número de unidades de vivienda.

3) Numerar los segmentos en cada manzana de 1 a 4.

4) Seleccionar los segmentos muestrales tomando un número aleatorio entre 1 y 4 para cada manzana.

5) Entrevistar todos los hogares en los segmentos seleccionados.

Nótese que aun cuando en ambos casos se obtiene una muestra del uno por ciento, el procedimiento B incluye más manzanas muestrales y menos unidades de vivienda por manzana. Por lo general, será más costoso obtener un mismo tamaño de muestra con el procedimiento B ya que existe allí el costo adicional del submuestreo que no aparece en el Procedimiento A asimismo, aumentará el número de viajes para visitar una mayor cantidad de manzanas. Este procedimiento de submuestreo es casi equivalente a dividir cada manzana de la ciudad en 4 partes, o segmentos, y tomar uno de cada 100 de esos segmentos. Por lo tanto, el uso del submuestreo descripto en el procedimiento B puede considerarse como esencialmente equivalente a usar una muestra de pequeños conglomerados de unidades de vivienda (en los que se enumeraría cada unidad de vivienda) pero con un muestreo matápico como un recurso para reducir la cantidad de trabajo exigida por la extracción de una muestra de pequeños conglomerados.

Procedimiento C:

Para una muestra de superficies con listado y submuestreo:

Para llevar a cabo el procedimiento B es necesario tener o preparar mapas detallados. Un tercer procedimiento, permite alcanzar aproximadamente los mismos resultados, y se aplica con frecuencia cuando no se tienen mapas detallados y cuando no es fácil prepararlos.

1) Proceder como en el paso (1) del Procedimiento B para seleccionar también una muestra de 1, en 25 manzanas.

2) Visitar cada manzana en la muestra y hacer una lista de todas las unidades de vivienda en la misma. La numeración puede efectuarse (a) separadamente en cada manzana (es decir, comenzando con el número 1 en cada manzana), (b) en una única secuencia a través de todas las manzanas en la muestra, o (3) mediante alguna combinación como podría ser una secuencia separada para varios grupos de manzanas.

3) Seleccionar un cuarto de las unidades de vivienda dentro de las manzanas muestrales ya sea usando números al azar o mediante el muestreo sistemático, utilizando los números de serie asignados a las unidades de vivienda.

4) Entrevistar las familias cuyos números de serie resultaron seleccionados en la muestra.

NOTA:

Si se dispone de información previa acerca del número aproximado de unidades de vivienda en todas las manzanas, se puede utilizar alguna combinación de los procedimientos mencionados, junto con la estratificación de las manzanas por tamaño.

3. ELECCION DE LA UNIDAD DE MUESTREO Y DEL DISEÑO DE LA MUESTRA

Al diseñar una muestra, el especialista en muestreo debe decidir cuantas etapas de muestreo comprenderá dicho diseño. Debe, además, determinar en cada etapa cuál es la unidad de muestreo. En estas decisiones el estadístico tiene ante sí varias alternativas. Supongamos, por ejemplo, que desea estimar el número promedio de cabezas de ganado por explotación. La información, finalmente, se obtendrá de una muestra de las fincas individuales (unidades de análisis o unidades elementales). Para obtener la muestra podría usar, en cambio, cualquiera de los planes siguientes:

1) Tomar una muestra simple al azar, estratificada o sistemática, de fincas individuales si existen listas completas y exactas de las fincas.

2) Usar los mapas para subdividir el país en pequeños segmentos de superficies (por ejemplo, segmentos con un promedio de alrededor de 5, ó 10, fincas); luego seleccionar una muestra de esos segmentos de superficies e incluir en la muestra todas las fincas dentro de cada segmento seleccionado. Para el caso de fincas que cruzan los límites de los segmentos se deberían establecer reglas para asociar las fincas con los segmentos.

3) Seleccionar una muestra de pequeñas subdivisiones administrativas, tales como distritos, e incluir en la muestra todas las fincas de los distritos seleccionados. O seleccionar una submuestra de fincas.

4) Seleccionar una muestra de provincias (subdivisiones administrativas más grandes) y tomar una muestra de superficies y fincas dentro de las provincias seleccionadas usando alguno de los procedimientos A, B y C descritos anteriormente.

Cuando se usa submuestreo, el conglomerado inicialmente seleccionado se denomina unidad de la primera etapa o unidad primaria de muestreo (UPM) y la unidad de submuestreo se denomina unidad de la segunda etapa. Por ejemplo, en (3) anterior, si se selecciona una submuestra de fincas, el "distrito" es la UPM y la finca es la unidad de la segunda etapa; en (4), la UPM es la"provincia", la superficie pequeña es la unidad de la segunda etapa y las superficies todavía más pequeñas o las fincas, serían unidades de la tercera etapa.

---

1 Véase la sección 14 del capítulo 2 de Sample Survey Methods and Theory (obra mencionada en la nota al pie 1 de la Conferencia 8) para una exposición más amplia al respecto.

¿Cómo podemos hacer una inteligente selección entre las varias alternativas? Podemos razonar en la forma siguiente: Si el costo no es importante, el muestreo unietápico, usando como unidad de muestreo la unidad elemental (la finca en el caso anterior), da los resultados más exactos para el número dado de unidades elementales en la muestra. (Existen algunas excepciones pero son casos muy poco usuales.) Sin embargo, cuando el costo y la conveniencia administrativa son importantes, puede ser conveniente el uso del muestreo de conglomerados en una o más etapas. El costo de la enumeración o por unidad elemental es regularmente menor si esas unidades forman conglomerados que si las mismas están aleatoriamente distribuidas a través del país; la conglomeración reduce el tiempo y costo de los viajes requeridos para la entrevista. En consecuencia, para una suma de dinero dada puede ser posible, usando el muestreo de conglomerados aumentar el número de unidades elementales en la muestra por arriba del número que se podría alcanzar, con el mismo presupuesto, usando una selección al azar. Si el mayor número de unidades compensa el hecho de que la muestra de conglomerados tiende a incrementar el error estándar, se habrá obtenido una ganancia neta en la confiabilidad de las estimaciones derivadas de la muestra.

Para efectuar una elección entre unidades de muestreo alternativas, debemos, por lo tanto, comparar los costos esperados con los errores estándar de los distintos diseños posibles y usar el método que proporcione el error estándar menor para un costo fijo. En ciertas situaciones de orden administrativo, la decisión correcta puede ser obvia. Si la encuesta implica costos de viaje reducidos o nulos por ejemplo, si los cuestionarios se envían por correo o si la encuesta se efectúa con personal que cumple el viaje como parte de sus otras actividades normales, como podrían ser policías o carteros y si se tienen listados de las unidades elementales, debe tomarse siempre como unidad de muestreo la unidad elemental. Si los costos de los viajes o de la preparación de listas son elevados, un diseño alternativo usando una muestra de conglomerados será mejor regularmente. Una exposición completa de este punto está fuera del alcance de estas conferencias. Sin embargo trataremos aquí algunos de los puntos importantes.

4. ANALISIS DE COSTOS

Por lo general para una encuesta se tiene un presupuesto fijo y así una de las funciones más importantes del estadístico en muestras es proponer, dentro de ese presupuesto, un método que conduzca al menor error de muestreo. Examinemos primero como intervienen los costos en una encuesta donde se usa el muestreo de conglomerados.

Cuando estudiamos el muestreo estratificado expusimos la posibilidad de que los costos de enumeración y elaboración pudieran variar de un estrato a otro y construimos una función del costo expresando la parte variable del costo total como una suma de los costos unitarios multiplicados por los tamaños de muestra(por ejemplo, $C = C_1 n_1 + C_2 n_2 + \ldots$). En el muestreo de conglomerados se necesita un enfoque similar, si bien los costos unitarios son de tipo diferente. Consideremos, para tomar una situación sencilla, una mues

tra bietápica.

4.1 Componentes del costo

Para analizar los costos de una muestra bietápica de conglomerados es necesario identificar las distintas etapas de la encuesta y distinguir entre tres elementos del costo.

1) Costos generales, es decir, los costos que son fijos cualquiera sea el procedimiento de selección de la muestra.

2) Costos que dependen principalmente del número de conglomerados de la primera etapa en la muestra y de la forma como tales costos varían con las variaciones del número de esas unidades primarias de muestreo en la muestra.

3) Costos que dependen principalmente del número de unidades de la segunda etapa en la muestra y la forma como tales costos varían con este número.

4.11 Costos generales Los costos generales incluyen aspectos tales como el trabajo técnico y administrativo exigido por la encuesta, el alquiler del espacio y de ciertos tipos de equipo, el costo de la impresión de los resultados finales, etc. Esos costos serán aproximadamente los mismos aun cuando se consideren variaciones grandes en el tamaño y diseño de la encuesta. Como estos costos no se ven afectados por el tamaño de la encuesta no interfieren en la decisión acerca del diseño muestral. La única razón de separar estos costos es restarlos del presupuesto total disponible a fin de saber cuales son los fondos que pueden invertirse para atender los costos variables.

4.12 Costos de las unidades de la primera etapa: Ciertos costos varían, por lo general, en proporción con el número de unidades de muestreo de la primera etapa. Estos costos incluyen: (a) El costo de seleccionar cada unidad de la primera etapa, llegar hasta la misma y localizarla; (b) el costo de preparar una lista de unidades de la segunda etapa (dentro de las unidades primarias); y (c) el costo de diseñar la submuestra de unidades de la segunda etapa. Dependiendo de la naturaleza de la organización administrativa y de los materiales disponibles antes de comenzar la encuesta, pueden existir otros costos (costo de preparación de los mapas para las unidades de la primera etapa, contratación de enumeradores especiales, etc.).

4.13 Costos de las unidades de la segunda etapa: Los costos que dependen del número de unidades de la segunda etapa incluirán los costos de entrevista, revisión de los resultados de la encuesta, codificación, perforación, etc.

4.2Una función simple del costo

Supongamos una situación sencilla, en la cual el costo por unidad de primera etapa no cambia cualquiera sea el número de esas unidades en la muestra. En forma similar, el costo por unidad de la segunda etapa no cambia. Luego el costo variable total (que excluye los costos generales) puede representarse

así:

(9.1) $$C = C_1 m + C_2 n = C_1 m + C_2 m\bar{n}$$

donde

C_1..... es el costo por unidad de la primera etapa

C_2..... Es el costo por unidad de la segunda etapa

m y n.. Son los números totales de unidades de la primera y segunda etapa, respectivamente, en la muestra, y

$\bar{n}$..... Es el número promedio de unidades de la segunda etapa en una unidad primaria.

Usando la Ecuación (9.1) podemos determinar combinaciones de m y n que llegarían a un mismo costo. Por ejemplo, supongamos que el costo variable total fijado para una encuesta fue $2,500 y que las estimaciones de C_1 y C_2 fueron, respectivamente, $10 y 12. La tabla a continuación indica distintas combinaciones de tamaños de muestra, todas las cuales costarían exactamente $2,500; la última columna corresponde al tamaño medio del conglomerado ($\bar{n}$) en cada afijación:

Número de unidades en la muestra		Promedio ($\bar{n} = \frac{n}{m}$)
Primera etapa (m)	Segunda etapa (n)	
10	1200	120
20	1150	57.5
50	1000	20
75	875	11.7
100	750	7.5
125	625	5
150	500	3.3

Si se pudiera encontrar el error estándar para cada una de las combinaciones anteriores, sería posible elegir la combinación que diera el error estándar más bajo. En realidad, con este tipo simple de función del costo, se puede, por lo general, determinar matemáticamente la afijación óptima. Sin embargo, no es necesario; si se puede encontrar una fórmula que exprese la variancia en función de m y n puede verse simplemente cuál es la mejor combinación. Además, también puede hacerse esto en casos que implican funciones del costo complejas para las que es más difícil desarrollar una solución matemática del problema de la afijación óptima. La conferencia siguiente versará sobre el análisis de las variancias en las situaciones más

sencillas y comunes.

4.3 Funciones del costo más complejas

Debe hacerse un comentario más acerca de los costos. La formulación de la expresión del costo que hemos presentado antes, es decir $C = C_1 m + C_2 n$, cubre sólo el tipo más sencillo de situación. En la práctica, la función del costo puede ser mucho más compleja. Por ejemplo, puede existir una estratificación ya sea de las unidades de la primera o segunda etapa con diferentes costos unitarios en cada estrato. La función del costo sería:

$$C = \sum C_{1i} m_i + \sum C_{2i} n_i$$

y el problema de la afijación de la muestra representaría una combinación de la afijación óptima en el muestreo de conglomerados con la afijación óptima en el muestreo estratificado. Con frecuencia, los costos unitarios dependen del número de unidades en la muestra. Por ejemplo, supongamos que C_1 incluye una componente que resultó del tiempo destinado a viajes desde una a otra unidad de primera etapa. Con unas pocas unidades primarias en la muestra la distancia promedio de una unidad a la próxima sería algo grande con la consecuencia de aumentar el valor de C_1. Sin embargo, si aumentara el número de unidades en la muestra, la distancia promedio se haría más pequeña y C_1 resultaría menor. En tal situación se utilizaría un tipo diferente de función de costo. En general, cuando se proyecta una encuesta de gran escala e importancia, se debe efectuar un análisis detallado de las variaciones del costo a fin de construir una función del costo que sea realística para esa encuesta en particular.

TAREA DE ESTUDIO

Problema A : En la página siguiente aparece un mapa de un distrito dividido en seis subdistritos. Cada subdistrito está dividido en un número variable de AE)áreas de enumeración). Suponga - mos que desea seleccionar una muestra del 12.5 por ciento (1 en 8) de fincas en ese distrito

Ejercicio 1: Describa cómo seleccionaría una muestra del 12.5 por ciento de fincas usando una muestra del 25 por ciento de AE y una muestra del 50 por ciento de fincas dentro de las AE seleccionadas (usando un enfoque similar al procedimiento C en la Conferencia 9 del texto para seleccionar las fincas dentro de las AE en la muestra).

Ejercicio 2: Describa cómo seleccionaría una muestra del 12.5 por ciento de fincas usando una muestra del 12.5 por ciento de AE.

Ejercicio 3: Suponga que seleccionó primero una muestra del 50 por ciento de subdistritos. ¿Qué haría después para seleccionar su muestra del 12.5 por ciento de fincas en un total de dos etapas?.

¿Y en un total de tres etapas?.

Problema B: Se realiza una encuesta para proporcionar información sobre la producción de un cierto cultivo que puede ser cultivado sólo dentro de un programa de permisos bajo control gubernamental. Los permisos otorgados al comienzo de la estación de cultivo se usarán como una fuente de información. Los referidos permisos son concedidos por las oficinas agrícolas de los distritos. La muestra será una muestra bietápica: primero se seleccionará una muestra de oficinas agrícolas de los distritos, operación que estará a cargo de un técnico en muestras. Luego los enumeradores visitarán las oficinas seleccionadas, prepararán una lista de productores a los que se les ha otorgado un permiso y seleccionarán una muestra de los mismos. Los enumeradores visitarán luego las fincas en la muestra para obtener datos sobre la producción. Debido a que las oficinas que otorgan permisos están algo dispersas, se necesitará un enumerador distinto en cada una.

Ejercicio 4: Listados a continuación aparecen algunos de los rubros que componen el costo de la encuesta. Indicar con una marca en la columna apropiada si el costo debe considerarse principalmente parte del costo general, del costo de una unidad de la primera etapa o del costo de una unidad de la segunda etapa.

	General	Primera etapa	segunda etapa
a. Impresión de los cuestionarios.....			
b. Adiestramiento de los entrevistadores..			
c. Obtención de una lista de oficinas que otorgan permisos.....			
d. Visitas a las oficinas que otorgan permisos para seleccionar una muestra de productores con dichos permisos.....			
e. Selección de la muestra de productores con permisos.....			
f. Obtención de la información en las explotaciones.....			
g. Verificación en el campo del trabajo de los enumeradores por los supervisores..			
h. Crítica de los cuestionarios de obtención de la información.....			

	General	Primera etapa	segunda etapa
i. Preparación de un programa para tabular los resultados de la encuesta.....			
j. Preparación del informe final..			

Problema C: Refiérase a la tabla de la sección 4.2 en la Conferencia 9.

Ejercicio 5: Verificar que el costo de las distintas combinaciones de tamaños de muestra totaliza $2,500; suponer que el costo es $10 para una unidad de la primera etapa y $2 para una unidad de la segunda etapa.

CONFEPENCIA 10. MUESTREO DE CONGLOMERADOS--VARIANCIAS

1. VARIANCIA DE UNA MUESTRA BIETAPICA DE CONGLOMERADOS

Para estudiar la variancia de una muestra bietápica de conglomerados será conveniente revisar algunas ideas relativas al muestreo estratificado. En el muestreo estratificado el error estándar de una estimación muestral depende de las variancias intra-estrato, S_i^2. Para cada estrato, la variancia (S_i^2) se define mediante la misma fórmula que S^2 (variancia total de la población) pero usando sólo los elementos del i-ésimo estrato. Vimos antes que el muestreo estratificado era más útil cuando las medias de los estratos eran muy diferentes. En realidad, la ganancia que reporta el muestreo estratificado puede determinarse calculando la desviación estándar entre las medias de los estratos (es decir, calculando la desviación estándar de los números $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$, etc. ponderados por el número de unidades dentro de cada estrato) si es que se dispone de los datos necesarios. El cuadrado de esta desviación estándar ponderada entre las medias de los estratos se denomina la variancia <u>entre estratos</u>.

En el muestreo de conglomerados pueden considerarse conceptos similares. En realidad, existe una íntima analogía entre el muestreo de conglomerados y el estratificado. En ambos casos agrupamos en conjuntos los elementos individuales antes de seleccionar la muestra. La diferencia es que en el muestreo estratificado es necesario tomar muestras dentro de cada uno de los conjuntos (los estratos); en el muestreo de conglomerados se selecciona una <u>muestra</u> de los conjuntos (los conglomerados) y luego se incluyen ya sea todos o ya sea una muestra de los elementos dentro de los conjuntos seleccionados. <u>En los dos casos son muy diferentes el objetivo y el método de formación de los conjuntos</u>.

1.1 Símbolos

Consideremos un diseño bietápico en el que las unidades de la segunda etapa son las unidades de análisis.

M....es el número de UPM o conglomerados de la primera etapa en el universo (por ejemplo, segmentos de superficies o manzanas)

m....es el número de UPM seleccionadas en la muestra

N_i...Es el número de unidades de la segunda etapa (por ejemplo, explotaciones agropecuarias o unidades de vivienda) en el i-ésimo conglomerado

$\bar{N} = \frac{1}{M} \sum_{i=1}^{M} N_i$....es el número promedio en la población de unidades de la segunda etapa por unidad de la primera etapa.

n_i....es el número de unidades de la segunda etapa en el iésimo conglomerado seleccionadas en la muestra.

$\bar{n} = \frac{1}{m} \sum_{i=1}^{m} n_i$... es el número promedio en la muestra de unidades de la segunda etapa por unidad de la primera etapa

$S^2_{B:x}$....es la variancia entre conglomerados de la característica X (variancia entre los totales de los conglomerados); está definido por:

(10.1) $$S^2_{B:x} = \frac{1}{M} \sum_{i=1}^{M} (X_i - \bar{X})^2$$

donde $\bar{X}$ es el valor promedio de la variable por conglomerado de la primera etapa(no por unidad de la segunda etapa).

$$\bar{X} = \frac{1}{M} \sum_{i=1}^{M} X_i,$$

y X_i es el total en el i-ésimo conglomerado,

$$X_i = \sum_{j=i}^{N_i} X_{ij}.$$

En forma similar,

$S^2_{B:N}$ ...es la variancia entre conglomerados del número de unidades que contienen los conglomerados.

(10.2) $$S^2_{B:N} = \frac{1}{M} \sum_{i=1}^{M} (N_i - \bar{N})^2$$

$S_{B:x}$ .... es la covariancia entre conglomerados (covariancia entre los totales de los conglomerados para la característica X y el número de unidades que contienen los conglomerados).

(10.3) $$S_{B:y,N} = \frac{1}{M} \sum_{i=1}^{M} (X_i - \bar{X})(N_i - \bar{N}).$$

$S^2_{W:i}$... es la variancia dentro del conglomerado en el i-ésimo conglomerado que se define mediante

$$S^2_{w:i} = \frac{1}{N_i} \sum_{j=1}^{N_i} (X_{ij} - \bar{N}_i)^2$$

donde X_{ij} es el valor de la variable en la j-ésima unidad dentro del i-ésimo conglomerado y $\bar{Y}_i$ es el promedio de los totales de unidades en el i-ésimo conglomerado.

$$\bar{X}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} X_{ij}.$$

1.2 Estimación de medias y totales

Las fórmulas presentadas en las Conferencias anteriores para estimar las medias poblacionales son apropiadas cuando la unidad de muestreo es idéntica a la unidad de análisis. Una característica importante del muestreo de conglomerados es que la unidad de muestreo (al menos en la primera etapa) no es coincidente con la unidad de análisis. Así, en los ejemplos de la conferencia anterior, podría ser que no estuviéramos interesados en la media por familia, por escuela, por establecimiento fabril o por manzana, sino que quisiéramos estimar la media por miembro de la familia, por escolar, por trabajador fabril o por unidad de vivienda. Consideremos un diseño bietápico en el que las unidades de la segunda etapa son las unidades de análisis y en el que, primero, se seleccionan, mediante un muestreo simple al azar, m conglomerados entre los M que existen en la población y, segundo, de las N_i unidades en el i-ésimo conglomerado muestral, se extrae, también mediante un muestreo al azar, n_i unidades.

(10.4) $$\bar{\bar{X}}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} X_{ij}.$$

Como las unidades dentro del conglomerado se seleccionaron mediante un muestreo simple al azar, podemos (de acuerdo con la Conferencia 4, sección 2) estimar esta media, en forma insesgada, con la fórmula siguiente:

(10.5) $$\bar{\bar{X}}_{ij} \cong \frac{X_i}{n_i} \quad \text{donde} \quad X_i = \sum_{j=1}^{n_i} X_{ij}$$

Estas estimaciones de las medias por unidad en los conglomerados, calculadas con los m conglomerados muestrales, deben combinarse luego, en alguna forma, para estimar el total general en la población (X) y la media poblacional por unidad ($\bar{X} = \frac{X}{N}$). Al efecto existen varios estimadores que se mencionan en los textos usuales; examinaremos uno sólo de ellos.

Construiremos primero un estimador del total poblacional para la característica X. La fórmula siguiente nos da un estimador insesgado de X_i (total del i-ésimo conglomerado):

(10.7a) $$X' = \frac{M}{m} \sum_{i=1}^{m} X'_i = \frac{M}{m} \sum_{i=1}^{m} \frac{N_i}{n_i} X_i$$

En forma similar, podemos estimar el número total de unidades de análisis en la población (suponiendo que no lo conocemos) mediante

(10.7b) $$n' = \frac{M}{m} \sum_{i=1}^{m} N_i$$

La media poblacional por unid es:

$$\bar{X} = \frac{X}{N} = \frac{\sum X_i}{\sum N_i}$$

Un estimador de esa media es

(10.8) $$\bar{\bar{X}} = \frac{X'}{N'} = \frac{\sum_{i=1}^{m} N_i \bar{\bar{X}}'_i}{\sum_{i=1}^{m} N_i}$$

Como puede verse, este estimador es una media ponderada de las medias por unidad en los m conglomerados muestrales, donde las ponderaciones son los tamaños de los correspondientes conglomerados. Como se indicó antes, éste es sólo uno de los varios posibles estimadores; sin embargo, paraece ser el de uso más general. D do que tanto el numerador como el denominador son variables aleatorias, se trata de un estimador del tipo "por relativos" y tiene, por lo tanto, el sesgo acostumbrado de un estimador de ese tipo. Por lo general, si el número de conglomerados en la muestra es razonablemente grande, el sesgo no adquirirá mucha importancia. En la Conferencia 11 se exponen, con más detalle, los estimadores por relativos.

1.3 Variancias

Las variancias de x' (el estimador de X) y de n' (el estimador de N), son respectivamente

(10.9a) $$S^2_{x'} = \frac{M^2 (M-m)}{m(M-1)} S^2_{B:X} + \frac{M}{m} \sum_{i=1}^{M} N_i^2 \frac{(N_i - n_i)}{n_i (N_i - 1)} S^2_{w:i}$$

(10.9b) $$S^2_{n'} = \frac{M^2}{m} \frac{(M-m)}{(M-1)} S^2_{B:N}$$

La variancia del estimador de X es una suma compuesta de dos términos. El primero representa la contribución de las variancias que se deriva de la selección de las unidades de la primera etapa. El segundo es la contribución de la selección de las unidades de la segunda etapa. Si el diseño incluye tres o más etapas de muestreo, la fórmula de la variancia incluirá, a su vez, términos adicionales, similares a estos en su forma, por cada etapa adicional.

La variancia de $\bar{\bar{x}}$ (el estimador de $\bar{\bar{X}}$ es más compleja. Dicha variancia está dada, aproximadamente, por

$$(10.10)\quad s^2_{\bar{x}'} = \left(\frac{1}{N}\right)^2 \left(\frac{M-m}{M-1}\right) \frac{1}{m}\left[S^2_{B:X} + (\bar{\bar{X}})^2 S^2_{B:N} - 2\bar{\bar{X}} S_{B:X,N}\right]$$

$$+\left(\frac{1}{N}\right)^2 \left(\frac{1}{mM}\right) \sum_{i=1}^{M} N_i^2 \frac{(N_i-n_i)}{(N_i-1)} \frac{1}{n_i} S^2_{w:i}$$

La fórmula de la variancia incluye términos originados por las diferencias entre los conglomerados y un término debido a las diferencias dentro de los conglomerados. Debe notarse que la primera parte de la ecuación no sólo incluye un término por la variación entre las X_i (los totales en los conglomerados para la característica X) sino también términos por la variación entre las N_i (los tamños de los conglomerados en función del número de unidades secundarias) y la covariancia entre las X_i y las N_i. Como la covariancia será casi siempre positiva(es decir, cuanto mayor sea el número de unidades que contenga el conglomerado mayor será el valor de X_i), el término de la covariancia será, casi siempre, negativo. En esa forma sirve para reducir la componente de la variancia que se debe a la variación en el tamaño del conglomerado. O, considerando el mismo tema desde otro punto de vista, la componente "entre conglomerados" de la variancia general está, en último grado, basada en las diferencias entre las medias de los conglomerados por unidad de análisis más bien que en las diferencias entre los totales de los conglomerados. Esto se ve más fácilmente recordando que la primera parte de la variancia general puede escribirse en la siguiente forma:

$$\left(\frac{M-m}{M-1}\right)\left(\frac{1}{m}\right)\left(\frac{1}{M}\right)\sum_{i=1}^{M}\left(\frac{N_i}{N}\right)^2 (\bar{X}_i - \bar{\bar{X}})^2.$$

1.4 Métodos de los grupos aleatorios para aproximación de las variancias

Para estimar las variancias de la muestra existen fórmulas similares a las ecuaciones 10.9a y 10.10 anteriores en las que se usan estimaciones muestrales de las componentes de la variancia. Esas fórmulas son, sin embargo, algo engorrosas. En consecuencia se suelen utilizar aproximaciones abrevia-

das para reducir la cantidad de trabajo, en particular, si las estimaciones de la variancia deben calcularse para un número elevado de características. Una de esas aproximaciones se conoce con el nombre de método de los grupos aleatorios.

El método de los grupos aleatorios consiste en dividir la muestra, al azar, en varios grupos y usar luego cada uno para hacer una estimación total, media, etc. (ésto se efectuaría para cada característica cuya variancia se proyecta calcular.) Cada uno de los grupos aleatorios reflejará los diferentes pasos de la selección de la muestra de modo que la estimación que se derive de cada grupo será una estimación del total obtenida con el mismo diseño muestral que el de la muestra total (si bien, con un tamaño de muestra mucho menor). En una muestra multietápica, los grupos aleatorios se forman, generalmente, ubicando la muestra entera extraída de una misma unidad primaria de muestreo dentro del mismo grupo. Para el caso de diseños más complicados donde se usa la estratificación y/o el muestreo a través del tiempo, existen métodos algo diferentes para dividir la muestra en grupos aleatorios. Sin embargo, el método no es muy útil si el número de unidades de la primera etapa es pequeño.

Al calcular las estimaciones de la variancia, nuestro interés está precisamente en la variancia entre las diferentes estimaciones posibles del total o la media. Por lo tanto, este método que proporciona un conjunto de estimaciones diferentes del total o la media, cada una con cierto grado de estabilidad (es decir, el número de casos en un grupo no debe ser demasiado pequeño) constituye un método de estimación de las variancias que se ajusta a la realidad.[1]

2. FORMULAS LIMITATIVAS DE LA VARIANCIA EN EL MUESTREO BIETAPICO

Si examinamos las ecuaciones de la variancia (10.9a) y (10.10), podemos interpretar fácilmente lo que sucede en dos situaciones sencillas. Primero, si todas las unidades de la segunda etapa entran a formar parte de la muestra, estamos en el caso descripto en la Conferencia 9 como "muestreo unietápico de conglomerados". Dado que $n_i = N_i$, el término debido a la variación dentro de las unidades de la primera etapa resulta igual a cero. En la Ecuación (10.9a), el primer término es igual a la fórmula de la variancia en el muestreo simple al azar, salvo que los tamaños de muestra y los valores de X_i se refieren a las unidades de la primera etapa. Por ejemplo, si las unidades de la primera etapa son segmentos de superficies, N es el número total de esos segmentos y X_i (en la Ecuación (10.1) es el total en el segmento para la variable mencionada. En la Ecuación (10.10), el primer término es igual a la fórmula de la variancia para una estimación por relativos basado en una muestra simple al azar[2] (véase Ecuación 11.2 en la Conferencia 11) salvo que, también aquí, los tamaños de muestra y los valores de Y_i se

[1] Refiérase a la sección 16 del capítulo 10 de Sample Survey Methods and Theory (obra mencionada al pie 1 de la Conferencia 8).

refieren a las unidades de la primera etapa (por ejemplo, segmentos de superficie).

En segundo lugar, consideremos una situación en la que todas las unidades de la primera etapa están en la muestra. En este caso m = M y el primer término resulta cero. La variancia del estimador del total en la población es, por lo tanto, igual a

$$\sum_{i=1}^{M} N_i^2 \left(\frac{N_i-n_i}{N_i-1}\right) \frac{1}{n} S_{w:i}^2$$

La variancia del estimador de la media poblacional por elemento es, entonces, igual a

$$\left(\frac{1}{M}\right)^2 \sum_{i=1}^{M} N_i^2 \left(\frac{N_i-n_i}{N_i-1}\right) \cdot \frac{1}{n_i} S_{w:i}^2$$

Estas son las fórmulas de la variancia para los estimadores de totales y medias en una muestra estratificada. En otras palabras, una muestra estratificada es simplemente un caso especial de una muestra de conglomerados en la que se incluyen en la muestra todas las unidades de la primera etapa y se selecciona, en cada unidad de la primera etapa, una submuestra de unidades de la segunda etapa.

En esta exposición se ha considerado únicamente el caso del muestreo simple al azar, tanto en la selección de la primera etapa como en la selección de la segunda etapa. Se pueden desarrollar fórmulas análogas para el muestreo de conglomerados estratificado en las cuales la única diferencia es que los términos de las ecuaciones son reemplazados por la suma de términos similares a través de todos los estratos.[3] (Ver página siguiente).

3. ANALISIS DE LAS COMPONENTES DE LA VARIANCIA

Un análisis más detallado de la Ecuación (10.9a) y de la ecuación (10.10) mostraría que en el caso de una muestra bietápica que contuviera un cierto número total de unidades de análisis, las variancias de muestreo de las estimaciones computadas con la fórmula (10.7a) y la fórmula (10.8) dependen de varios factores. Dos factores importantes, que el muestrista debe considerar en el diseño de la muestra, son:

---

[2] Esta equivalencia no es directamente obvia debido a las diferencias entre los símbolos para los conglomerados en la Ecuación (10.10) y para el muestreo simple al azar en la Ecuación (11.2). Por ejemplo, Y en (11.2) corresponde a N en (10.10), $\bar{Y}$ corresponde a $\bar{X}$, N corresponde a M, S_Y^2, corresponde a $M^2\left(\frac{M-m}{M-1}\right)\frac{1}{m} S_{B:X}^2$, etc.

1) La variabilidad en el tamaño de las unidades de la primera etapa en función del número de unidades de la segunda etapa que contienen aquellas unidades.

2) La variabilidad entre las unidades de la segunda etapa (las unidades elementales o unidades de análisis) dentro de las unidades de la primera etapa.

3.1 Variabilidad en el tamaño de las unidades de la primera etapa:

Si el tamaño de las unidades de la primera etapa, en función del número de unidades de la segunda etapa (por ejemplo, número de fincas en un segmento de superficie) que contienen, es desigual, tales variaciones de tamaño pueden tener un efecto profundo sobre el tamaño de la variancia del estimador del total poblacional como lo señala el primer término de la Ecuación (10.9a). Como se mencionó en la sección 1.3 anterior, la variancia del estimador de la media poblacional por unidad elemental (Ecuación (10.10) está afectada por la variación entre las medias en la primera etapa por elemento. A pesar de que la estimación por relativos tiende a compensar las variaciones de tamaño de las unidadaes de la primera etapa, siempre puede persistir un efecto adverso sobre la variancia del estimador si las medias en la primera etapa por elemento guardan relación con los tamaños de las unidades de la primera etapa es decir, si las unidades grandes de la primera etapa muestran una tendencia a tener medias grandes (o pequeñas)por elemento. Si la variación del tamaño es muy pronunciada, será necesario usar una muestra grande de unidades de la primera etapa o cambiar el método de muestreo y estimación para mantener el error estándar dentro de límites razonables (véase sección 4 más adelante).

3.2 Variabilidad entre las unidades de la segunda etapa

El segundo factor importante es la variabilidad entre las unidades de la segunda etapa (unidades de análisis) dentro de las unidades de la primera etapa (conglomerados). Para un plan de muestreo dado en el que se seleccionan primero m de los M conglomerados y luego, en cada conglomerado muestral, un promedio $\bar{n}$ de unidades de análisis se puede demostrar que cuanto mayor sea la variabilidad entre las unidades de la segunda etapa dentro de las unidades de la primera etapa, más pequeña resultará la variabilidad de muestreo de las estimaciones que se obtengan. En otras palabras, es conveniente que las unidades de análisis tengan una correlación intraclase relativamente baja. La correlación intraclase es una medida de la similitud de las unidades dentro de un conglomerado en relación con las características que se investigan. 4 (ver página siguiente).

La demostración matemática de este fenómeno escapa al alcance de esta conferencia; sin embargo, podemos tener una interpretación intuitiva del mismo mediante un ejemplo extremo. Consideremos una situación en la que las unida-

3 Refiérase a la ecuación (5.3) en la sección 5 del capítulo 7 de Sample Survey Methods and Theory (obra mencionada en la nota al pie de la Conferencia 8).

dades de análisis dentro de cada conglomerado sean idénticas. Evidentemente, un plan de muestreo como el descripto antes no será eficiente. Una única unidad de análisis dentro de un cierto conglomerado proporcionaría información completa acerca de todas las unidades y, en consecuencia, las n-1 unidades restantes no aportarían a nuestro conocimiento ninguna información adicional. Incluirlas en la muestra sería derrochar los recursos. La ineficiencia de este diseño en una situación como ésa se reflejaría en una elevada variabilidad de muestreo en comparación con una muestra simple al azar con el mismo número de unidades de análisis.

El estadístico cuando prepara el diseño de una muestra debe tomar en cuenta el efecto de la correlación intraclase sobre la variabilidad de muestreo. Esto es particualrmente cierto en el muestreo de superficies ya que las unidades que están geográficamente próximas son, por lo general, bastante similares en lo que se refiere a muchas características como, por ejemplo, ingreso, educación, actitudes, tipo de actividad agropecuaria, etc. El enfoque acostumbrado consiste en limitar el número de unidades de análisis que se toma de las unidades de la primera etapa e incluir en la muestra más unidades de la primera etapa. En una muestra monoetápica ésto puede lograrse haciendo los conglomerados tan pequeños como sea prácticamente posible. Sin embargo, el enfoque más común es introducir etapas adicionales en el proceso de muestreo de modo que el número de unidades de análisis que se selecciona últimamente de cada unidad en la etapa final sea pequeño. El estadístico debe, por supuesto, hacer un balance entre la precisión y el costo antes de decidir acerca del plan de muestreo.

Nótese que en el muestreo de conglomerados obtenemos una ganancia teniendo, dentro de los conglomerados, unidades tan diferentes como sea posible mientras que en el muestreo estratificado la ganancia proviene de tener, dentro de los estratos, unidades tan similares como sea posible. La razón de esta diferencia resulta clara si se recuerda, de acuerdo con la sección 2 anterior, que en el muestreo estratificado la componente "entre-conglomerados" de la variancia desaparece totalmente de la ecuación.

4. CONTROL DE LA VARIABILIDAD DEL TAMAÑO DEL CONGLOMERADO

En toda esta exposición, se ha supuesto que el único camino posible que podíamos tomar para producir, con una población dada, algún efecto sobre la variancia de muestreo era tomar más o menos casos en la muestra en la primera o en la segunda etapa o variar el tamaño de las unidades de la primera etapa. Por supuesto, si la variancia de muestreo puede reducirse mediante una estratificación apropiada esto debe hacerse primero. Existen también varios procedimientos especiales para controlar el efecto de la variabilidad en el tamaño del conglomerado. Más abajo se describen los procedimientos más importantes al respecto.

---

[4] Véase la ecuación 8 del capítulo 6 de Sample Survey Methods and Theory (obra mencionada en la nota 1 de la Conferencia 8) para una exposición más detallada del efecto de la correlación intraclase sobre la variancia.

Si bien la exposición se refiere a una muestra bietápica, se podría hacer un análisis similar para tres o más etapas. Los procedimientos descriptos a continuación para controlar la variabilidad en el tamaño se aplican igualmente a las etapas primera, segunda o subsiguientes siempre que se use muestreo de conglomerados.

4.1 Definir conglomerados de igual tamaño

Un método obvio es intentar definir los conglomerados de modo tal que sean de tamaño aproximadamente iguales, en función del número de unidades de análisis, con la esperanza de que esto tienda a hacerlos también iguales en función de las características que se investigan. Si ésto es posible usando los materiales e información disponibles no se requiere ninguna otra acción. Por ejemplo, si se dispone para las ciudades y aldeas de recuentos del número de unidades de vivienda en las manzanas es posible agrupar manzanas pequeñas para hacer conglomerados que contengan aproximadamente el mismo número de unidades de vivienda.

En algunos casos es posible que se puedan definir los conglomerados directamente en función de una de las características que se investigan. Por ejemplo, en una encuesta agropecuaria , los conglomerados se pueden construir de modo que sus superficies sean casi iguales. Si se cuenta con fotografías aéreas recientes, podrían hacerse casi iguales aún en términos de la extensión de tierra cultivada.

4.2 Estratificar los conglomerados según el tamaño

Si se tiene información sobre el tamaño de todos los conglomerados de la primera etapa en el universo antes de efectuar la encuesta (son adecuadas aproximaciones razonablemente buenas, es posible estratificar los conglomerados según grupos de tamaños. El efecto de la estratificación es reemplazar una variancia total por una suma de variancias dentro de los estratos. Dentro de cada estrato, los conglomerados deben ser casi de igual tamaño; por lo tanto, la estratificación por tamaño del conglomerado tendrá casi el mismo efecto que hacer casi iguales los tamaños de los conglomerados en la población total.

Si no se tiene información acerca del tamaño puede ser conveniente destinar una pequeña parte de los recursos disponibles, por ejemplo, en efectuar un "Recuento Rápido" de las manzanas y obtener en esa forma los tamaños aproximados de las unidades de la primera etapa (en función del número de unidades de vivienda que contienen). Los errores en los recuentos no originan sesgos en las estimaciones ya que éstas se basan en el número verdadero de unidades de vivienda encontradas en la encuesta misma.

Se puede efectuar un muestreo proporcional u óptimo dependiendo de cuál se considera el más apropiado para el caso particular. Si se trata de estimar más de una característica, el muestreo proporcional puede ser más conveniente que la afijación óptima, puesto que cada característica podría tener una afijación óptima distinta. Además, por lo general el muestreo proporcio-

nal puede ser más conveniente que la afijación óptima, puestoque cada característica podría tener una afijación óptima distinta. Además por lo general el muestreo proporcional es más seguro, a menos que se tengan medidas muy buenas del tamaño, ya que el uso de la fórmula de la afijación óptima con pobres medidas del tamaño puede actualmente aumentar la variancia.

4.3 Usar estimaciones por relativos

Un tercer método para reducir el efecto de la variabilidad en el tamaño de los conglomerados es usar estimaciones por relativos. Las estimaciones por relativos se expondrían con mayor detalle en la Conferencia 11; daremos aquí un ejemplo del método. Una estimación por relativos hace uso de una cantidad de la forma $\frac{x'}{y'}$ donde tanto x' como y' son estimaciones de totales calculadas con los datos muestrales. Y, el total en el universo de la cantidad de la que y' es una estimación, debe ser un valor conocido. Se puede hacer una estimación por relativos del total en el universo X- estimación que, con frecuencia, es muy eficiente -usando

$$X'' = \frac{x'}{y'} Y$$

en lugar de x' solamente. La nueva estimación X'' difiere por lo tanto significativamente de x' ya que implica dos rubros que poseen variancias de muestreo enlugar de uno. Las estimaciones por relativos son en general mucho menos sensibles a la variación en el tamaño de los conglomerados que las estimaciones del tipo.

$$X' = \frac{N}{n} \sum_{i=1}^{n} X_i$$

y su uso reducirá frecuentemente los errores estándar en forma apreciable.

4.31 Relación con el número aproximado de unidades de análisis.

Nos referimos aquí a dos usos diferentes de las estimaciones por relativos. En el primero, y es una variable íntimamente relacionada con el número total de unidades de análisis en los conglomerados, e y' es una estimación del total poblacional Y, basada en los conglomerados de la muestra únicamente. Por ejemplo, consideremos un diseño de muestra en el que las unidades de la primera etapa son las manzanas y las unidades de la segunda etapa, y unidades de análisis a la vez, las viviendas. Como resultado de censos anteriores o de recuentos especiales hechos con ese propósito disponemos de recuentos aproximados (Y_i) del número de unidades de vivienda en cada manzana. Para obtener Y se pueden totalizar esos recuentos en todas las manzanas de la ciudad. Luego y',estimación muestral de Y, se puede obtener sumando los recuentos aproximados en las manzanas muestrales únicamente y multiplicando la suma por $\frac{M}{m}$ (siendo M el número total de manzanas en la ciudad y m el co-

5 Podría ser una proyección o alguna otra cifra que se considere muy próxima al valor verdadero

rrespondiente número en la muestra). Luego, $X'' = \frac{x'}{y'}, Y$ es una estimación por relativos de X.

Si dentro de las unidades de la primera etapa se aplicara un submuestreo sería necesario modificar el procedimiento. A fin de obtener la mayor ganancia con este tipo de estimación por relaticos se aconseja no hacer un submuestreo independiente dentro de los conglomerados sino tratar las unidades de la segunda etapa dentro de los conglomerados como si formaran parte de una lista continua y tomar una muestra sistemática a través de la lista completa.

4.32 Relación con una estadística correlacionada.

El segundo uso de las estimaciones por relativos se refiere al caso en que se conoce el valor verdadero de algún total en el universo Y y se puede obtener en la encuesta una estimación muestral y' (de Y). Si las características X e Y están correlacionadas positivamente, luego $\frac{x'}{y'}$ Y, reducirá también el efecto de la variabilidad en el tamaño de los conglomerados (y, posiblemente, también otros tipos de variabilidad). Por ejemplo, supongamos que se proyecta una encuesta para medir las ganancias totales en jornales y salarios de los trabajadores fabriles (X). Podemos tomar una muestra de fábricas (los conglomerados) e incluir todos los trabajadores en la muestra de fábricas. Supongamos que, de alguna otra fuente - digamos, los registros de impuestos- se puede extraer el valor total de las ventas en todas las fábricas (Y). Podríamos luego incluir, en los cuestionarios presentados a las fábricas que integran la muestra, una pregunta sobre las ventas totales (y_i) y sobre los pagos en jornales y salarios (x_i) y preparar estimaciones de los totales poblacionales para ambas características usando la muestra en la forma acostumbrada. La estimación por relativos de los jornales y salarios sería entonces $\frac{x'}{y'}$ Y.

4.4 Uso de probabilidades proporcionales al tamaño

Un cuarto método para controlar los efectos de la variabilidad en el tamaño de los conglomerados es seleccionar los conglomerados muestrales con probabilidades proporcionales al tamaño en lugar de hacer una muestra simple al azar de conglomerados. La expresión probabilidad proporcional al tamaño se abrevia frecuentemente con las letras PPT. La selección con PPT implica que un conglomerado que es, por ejemplo, 5 veces más grande que otro, tendrá una probabilidad 5 veces mayor de figurar en la muestra. A primera vista podría parecer que esto introduciría un sesgo en los resultados muestrales con la sobrerepresentación de algunos conglomerados y la subrepresentación de otros. La estimación insesgada del total, cuando se use PPT y no existe submuestreo, es

$$x' = \sum_{i=1}^{m} \frac{X_i}{P_i}$$

donde x_i es el total en el i-ésimo conglomerado en la muestra y P_i es la probabilidad de selección de ese conglomerado. Puede verse fácilmente que ésto proporciona una estimación insesgada de $\bar{Y}$.[6]

4.41 Muestreo bietápico

Una aplicación común del muestreo con PPT es el uno de PPT para la selección de las unidades de la primera etapa en una muestra bietápica. Cuando se hace así, las tasas de submuestreo son por lo general fijadas en forma inversamente proporcional al tamaño (es decir, igual a $\frac{K}{P}$). Como resultado, la probabilidad de que una unidad de la segunda etapa sea incluída en la muestra es el producto de la probabilidad de selección en la primera y en la segunda etapa, es decir, $P_i \left\{\frac{m}{P_i}\right\} = k$. Todas las unidades de la segunda etapa tienen, por lo tanto, probabilidades idénticas (iguales a k) resultando la muestra autoponderada. En este caso la preparación de las estimaciones es muy simple; por ejemplo,

$$x' = \frac{1}{k} \sum^{m} x_i.$$

Este tipo de procedimiento de selección tiene muchas otras ventajas por ejemplo, la carga de trabajo puede hacerse aproximadamente igual en todas las unidades de la primera etapa seleccionadas; además, las estimaciones tendrán variancias más pequeñas que las de una muestra proporcional en la que las unidades de la primera etapa están seleccionadas con probabilidades iguales,.

4.42 Pérdidas del tamaño

Para hacer una selección por PPT es necesario tener las medidas del tamaño de cada conglomerado en la población, en la misma forma que se necesitaba para la estratificación según tamaño o para la primera forma de estimaciones por relativos descrita antes (sección 4.31 de esta conferencia). Si no se tienen las medidas del tamaño será conveniente hacer todos los esfuerzos para preparar estimaciones aproximadas del tamaño (las aproximaciones serán casi tan efectivas como medidas más exactas). Asumiendo que se tienen dichas medidas. El mecanismo para seleccionar una muestra con PPT se podrá exponer mejor a través de un ejemplo.

4.43 Ejemplo:

Supongamos que los conglomerados son manzans y que deseamos extraer una muestra de unidades de vivienda en un universo compuesto por las 10 manzanas listada en la columna 1 del cuadro 10/. En la columna 2, anotaríamos las medidas del tamaño de cada una de las manzanas (podrían ser estimaciones aproximadas del número de unidades de vivienda) y en la columna 3 los valores acumulados de tales medidas. La última cifra en la columna 3 es el número total (estimación aproximada) de unidades de vivienda en las 10 manzanas. Digamos que deseamos incluir en la muestra 5 de las 10 manzanas y que la muestra va ha incluir el 10 por ciento de todas las unidades de vivienda.

---

6 P_i se usa sólo en esta sección para indicar probabilidades. En otras secciones el símbolo P ha indicado una proporción.

CUADRO 10A, SELECCION DE LAS MANZANAS QUE INTEGRAN LA MUESTRA

Manzana número (UPM) (1)	Medida del tamaño (2)	Medida Acumulada (3)	Designación en la muestra (4)	Probabilidad de selección(P_i) (5)	tasa dentro del conglomerado($\frac{K}{P_i}$) (6)
1	50	50	22.5	50÷60.2	60.2÷500
2	12	62			
3	20	82			
4	31	113	82.7	31÷60.2	60.2÷310
5	10	123			
6	60	183	142.9	60÷60.2	60.2÷600
7	55	238	203.1	55÷60.2	60.2÷550
8	13	251			
9	30	281	263.3	30÷60.2	60.2÷300
10	20	301			

Una vez completadas las tres primeras columnas del cuadro 10A se procede en la forma siguiente:

1) Dividir la media final acumulada (301) por 5 ya que existirán en la muestra 5 manzanas; el valor 60.2 que se obtiene es el"intervalo de muestreo" para la selección de las manzanas.

2) Elegir un número al azar comprendido entre 00.1 y 60.2; supongamos que el número es 22.5.

3) Usar ese número aleatorio como número de arranque y anotar el mismo en la columna 4 en la línea en que por primera vez la media acumulada resulta igual o mayor que 22.5.

4) Sumar el intervalo 60.2 al número aleatorio de arranque; anotar el valor 82.7 así obtenido en la línea de la manzana a la que le corresponde una medida acumulada igual o mayor que dicho valor. Repetir sucesivamente la suma del intervalo 60.2; los números obtenidos, 142,9.....,etc., anotarlos en la columna 4 siguiendo la regla anterior hasta llegar a un número mayor que la última medida acumulada.

5) Integrar la muestra con las manzanas que figuran con anotaciones en la columna 4. En este ejemplo, las manzanas 1, 4, 6, 7, 7 y 9.

6) Anotar en la columna 5 la probabilidad (P_i) de selección de cada manzana. Para cada manzana, esta probabilidad se calcula dividiendo la medida del tamaño que figura en la columna 2 por el intervalo de muestreo 60.2.

7) Calcular y anotar en la columna 6 la tasa de muestreo $\left(\frac{k}{P_i}\right)$ que se usará dentro de cada manzana seleccionada. Para cada manzana esta tasa es la probabilidad deseada general de selección -exactamente $\frac{1}{10}$ dividida por la anotación que figura en la columna 5. Así, en la manzana 1, la tasa sería $\frac{1}{10} \div \frac{50}{60.2} = \frac{60.2}{500}$ ó también $60.2 \div 500$.

En ciertas ocasiones ocurre que algunas de las manzanas son tan grandes que las medidas del tamaño resultan mayores que el intervalo de muestreo. En esos casos figurarán en la columna 4 dos o más anotaciones para la misma manzana. De ser así se ajusta la tasa de submuestreo dentro de la manzana para hacer que la probabilidad general de la selección de las unidades de vivienda sea igual a k.

Problema A: Una población está compuesta de cuatro conglomerados. Las unidades de la segunda etapa, que son también en esta caso unidades elementales, son viviendas a las que les corresponden los alquileres siguientes:

	Conglomerado 1	Conglomerado 2	Conglomerado 3	Conglomerado 4
	\$100	\$100	\$10	\$50
	100	100	20	90
	200		40	
	400		50	
Totales	800	200	120	140

Ejercicio 1. ¿Cuál es el valor de $S^2_{B:Y}$ (variancia entre conglomerados)?.

Ejercicio 2. ¿Cuál es el valor de la variancia "dentro de los conglomerados" en el primer conglomerado?.

Ejercicio 3. Se selecciona una muestra de dos conglomerados con probabilidades iguales; dentro de cada conglomerado seleccionado pasan a integrar la muestra la mitad de las unidades elementales que forman aquél.

a) ¿Cómo calcularía usted $\bar{x}'$, es decir, la estimación de $\bar{Y}$?

b) ¿Cuál es la variancia de la estimación muestral del total ($S^2_{\hat{Y}'}$)?

c) ¿cuál es la probabilidad de una unidad elemental de ser incluida en la muestra?.

Problema B: Considere una ciudad compuesta por las 12 manzanas listadas a continuación en la primer columna. Las cifras en la segunda columna son las medidas del tamaño (número aproximado de unidades de vivienda en cada manzana). Sobre la base de esta información deseamos seleccionar, con probabilidades proporcionales al tamaño, una muestra de 4 manzanas y luego, dentro de éstas, seleccionar unidades de vivienda en forma tal que se tenga una muestra autoponderada con un número esperado de 10 unidades de vivienda.

Número de la manzana (UPM)	Número aproximado de unidades de vivienda (medida del tamaño)	Medida acumulada	Número verdadero de unidades de vivienda*	No. de serie de las unidades de vivienda verdaderas
1	10	10	9	1 a 9
2	5	15	6	10 a 15
3	2	17	2	16 a 17
4	5	22	6	18 a 23
5	5	27	6	24 a 29
6	10	37	8	30 a 37
7	10	47	8	38 a 45
8	2	49	2	46 a 47
9	2	51	4	48 a 51
10	5	56	6	52 a 57
11	5	61	6	58 a 63
12	10	71	9	64 a 72
Total	71		72	

* El número verdadero de unidades de vivienda que se encontraría en la realidad en la manzana en una operación de campo si la manzana fuera seleccionada en la muestra.

Ejercicio 4. Prepare una hoja de trabajo similar al cuadro 10A (en la Conferencia 10) y seleccione la muestra de manzanas. Suponga que el número aleatorio de arranque para determinar las manzanas muestrales es 3.7.

Ejercicio 5. Suponga que ha visitado las manzanas seleccionadas en su muestra y que ha establecido el número de vivienda en la forma que figura en la columna anterior. Las unidades de vivienda realmente existentes en cada manzana están señaladas mediante "números de serie" como se ve en la columna quinta. Efectúe los cálculos necesarios para seleccionar la muestra de unidades de vivien

da y liste los números de serie de las unidades de vivienda seleccionadas en la muestra.

CONFERENCIA 11. ESTIMACIONES POR RELATIVOS

1. RAZONES PARA CONSIDERAR EL USO DE ESTIMACIONES POR RELATIVOS

En las conferencias anteriores tratamos el problema de cómo diseñar la muestra más eficiente (desde el punto de vista de la minimización del error estándar) usando la mayor cantidad posible de información confiable acerca de la población que pudiéramos obtener. Hemos visto de qué modo se usa la información para la estratificación, ya sea en el muestreo proporcional o en la afijación óptima, cómo tomar en consideración las unidades de costo y de qué modo se decide una elección entre diferentes clases de unidades de muestreo. Hemos visto cómo usar el conocimiento anterior que tengamos sobre los costos y sobre las variancias de dos diferentes métodos de muestreo a fin de producir la cantidad máxima de información con los recursos disponibles a nuestro alcance. Todos estos análisis se han efectuado en función de estimaciones bastante simples, como por ejemplo $\bar{x}$, x, p', preparadas utilizando únicamente los datos de la muestra, el número total de unidades (N) en la población y las probabilidades de selección. Así, en el muestreo simple al azar,

$$\bar{X}' = \frac{\sum x_i}{n}$$

$$X' = \frac{N}{n} \sum x_i .$$

En el muestreo estratificado,

$$\bar{x}' = \frac{1}{N} \sum_i \frac{N_i}{n_i} \sum_j x_{ij}$$

$$x' = \sum_i \frac{N_i}{n_i} \sum_j x_{ij}$$

Fórmulas similares vimos en el muestreo de conglomerados o en la estimación de proporciones (p').

Sin embargo, existen, para estimar esas mismas estadísticas, métodos más complejos que, bajo ciertas circunstancias, pueden resultar en reducciones considerables en los errores estándar.

Además deseamos medir otros tipos de estadísticas -como ser, relaciones de dos características, cambios a través del tiempo de una sola caracterís-

le

tica, etc. Por ejemplo, podremos obtener información sobre pago en jornales y salarios y sobre el número de horas trabajadas. Sin embargo, nuestro interés puede estar, no tanto en los salarios y jornales totales u horas trabajadas totales, como en la estimación de las ganacias promedio por hora. En el caso de encuestas que abarcan dos períodos distintos, podríamos estar más interesados en descubrir si los salarios totales han subido o bajado que en medir el nivel de uno cualquiera de los salarios en una fecha. El análisis de los errores estándar de relaciones estimadas también ayuda a resolver el problema de producir estimaciones más eficientes de medias y totales.

Investigaremos el método más sencillo y más frecuentemente utilizado para mejorar la confiabilidad de una media o un total estimado consistente en usar una técnica especial de estimación que produce una "estimación por relativos". En situaciones particulares es útil emplear un conjunto de otras herramientas muy poderosas; por ejemplo, las estimaciones por regresión, el muestreo doble (en el que la muestra final se selecciona de una muestra mayor seleccionada previamente la que proporciona información para mejorar la selección final o el procedimiento de estimación), y los métodos especiales para la estimación de series cronológicas. La conferencia 11 se dedicará sin embargo al tratamiento de las estimaciones por relativos únicamente.

2. ESTIMACIONES POR RELATIVOS AGREGADOS

Las estimaciones por relativos son las más frecuentemente usadas dentro del conjunto de las técnicas de estimación más complejas a disposición de los estadísticos. Son a su vez las más fáciles de aplicar. Estas estimaciones son apropiadas si las unidades de la población poseen dos características correlacionadas positivamente- cuanto mayor es la correlación, mayor es la ganancia que se deriva de esta técnica. El tipo más sencillo de estimador por relativos de la forma X'', dado por la ecuación (11.1), es una estimación de X (el total en la población).[1]

(11.1)

$$x'' = \frac{x'}{y'}\ Y.$$

Aquí, x' e y' son las estimaciones acostumbradas de los totales de las dos características X e y; para estimar el total Y debe conocerse el total Y.

Para calcular X" no es necesario calcular x' e y' ya que, en una muestra autoponderada

$$\frac{x'}{y'} = \frac{\bar{x}'}{\bar{y}'} = \frac{\sum^{m} x}{\sum^{n} y}$$

1. El estimador por relativos de una media $\bar{x}''$ (como una estimación de $\bar{X}$) se obtiene dividiendo x'' por N tiene el mismo coeficiente de variación que x''.

en cambio, la fórmula (11.1) es útil para derivar la variancia de x".

Las estimaciones por relativos de agregados se aplican ordinariamente a las tres situaciones que se exponen en las secciones 2.1 a 2.3 siguientes.

2.1 Relación con respecto a la misma característica o alguna otra afin en un período de tiempo anterior.

Las características X e Y son de tipo similar si bien Y se refiere a un período de tiempo anterior en el que se efectuó un censo completo. Por ejemplo, en un cierto año se puede haber tomado un censo completo de manufacturas y deseamos realizar una encuesta por muestra al año siguiente. Supongamos que deseamos estimar el valor total de los embarques. Para cada establecimiento fabril en la muestra obtenemos no sólo el valor x_i, valor de los embarques en el año de la encuesta, sino también y_i, valor durante el año anterior en que se efectuó el censo. x' e y' serían, por lo tanto, estimaciones obtenidas con la muestra de los embarques totales en los dos años preparadas mediante los métodos expuestos antes. Y es el valor total de los embarques tabulado en el censo completo. En esta aplicación, la encuesta se usa realmente para medir la tasa de cambio entre los dos años, usando una misma muestra de establecimientos. La tasa de cambio se multiplica luego por el total CENSAL para el año anterior.

2.2 Relación de dos características afines en el mismo período de tiempo

X e Y son dos características diferentes para el mismo período de tiempo que se sabe de antemano están correlacionados positivamente. Se conoce además el valor verdadero del total Y. Por ejemplo, para la i-ésima finca en una muestra, y_i puede ser el total de hectáreas de la finca y x_i los pagos por mano de obra; el número total de hectáreas de todas las fincas Y, se conoce a través de otra fuente. Si las fincas de mayor tamaño pagan, en general, salarios totales por mano de obra agropecuaria superiores a los de las más pequeñas, la estimación por relativos puede reducir drásticamente el error de muestreo. En esta aplicación, la encuesta se usa para medir una relación (tal como el pago promedio por hectáreas) que se multiplica por el número conocido de hectáreas.

2.3 Relación de un subconjunto con respecto al total

La característica X es un subconjunto de Y, variando aproximadamente en proporción a Y. Por ejemplo, y_i puede ser el número total de acres en la i-ésima finca de la muestra y x_i los acres cultivados con un producto dado en esa finca. Otra aplicación sería el caso en que Y es el número total de unidades de análisis y X el número de esas unidades que tienen un atributo particular. Por ejemplo x_i podría ser el número de personas en la fuerza de trabajo en el i-ésimo conglomerado; y_i el número total (2, ver pág. siguiente) de personas en ese conglomerado; e Y el número total de personas en la población, valor conocido. En esos casos, la encuesta se usa para medir una relación $\frac{x'}{y'}$ que se multiplica luego por el total poblacional (Y) de la característica que figura en el denominador de la relación.

3. VARIANCIA Y SESGO DE UNA ESTIMACION POR RELATIVOS

Al examinar $\frac{x'}{y'}Y$, se ve claramente que Y no se deriva de la muestra. El error de muestreo de la estimación $x'' = \frac{x'}{y'}Y$, depende, por lo tanto, del error de muestreo de la relación $r' = \frac{x'}{y'}$; Y actúa como un factor constante. En esa forma el análisis del error de muestreo de x'' está íntimamente relacionado con el de la relación $r' = \frac{x'}{y'}$ como una estimación de $R = \frac{X}{Y}$.

La fórmula matemática de la distribución, de muestra en muestra, de la relación de dos variables aleatorias es mucho más complicada que la distribución de las estimaciones más sencillas ya descriptas. Implica la relación de dos variables cada una de las cuales acusa errores de muestreo. De allí que se requiere un cuidado mayor para decidir cuando se usan tales relaciones. Los siguientes hechos acerca de la variancia de una relación y de las estimaciones por relativos indicarán en qué casos se usa un estimador por relativos para estimar una media o un total. Nos dirán asimismo el error que debe esperarse cuando se usa la estimación.

3.1 Variancia de relaciones y estimaciones por relativos

La variancia de una relación estimada $r' = \frac{x'}{y'}$ es aproximadamente

(11.2)
$$S_{r'}^2 = R^2 \left(\frac{S_{x'}^2}{X^2} + \frac{S_{y'}^2}{Y^2} - 2p \frac{S_{x'} S_{y'}}{XY} \right)$$

donde R es la relación en la población $\frac{X}{Y}$ (una relación de totales). En términos similares, la variancia de una estimación por relativos de un total, $x'' = \frac{x'}{y'}Y$, es

(11.3)
$$S_{x''}^2 = Y^2 S_{r'}^2 = X^2 \left(\frac{S_{x'}^2}{X^2} + \frac{S_{y'}^2}{Y^2} - 2p \frac{S_{x'} S_{y'}}{XY} \right)$$

Las Ecuaciones (11.2) y 11.3) resultan algo más simples si se expresan en función del coeficiente de variación, es decir V. El cuadrado del coeficiente de variación (es decir, la variancia relativa) de r' es igual al del de x'' y puede expresarse así:

(11.4)
$$V_{x''}^2 = V_{r'}^2 = V_{x'}^2 + V_{y'}^2 - 2p\, V_{x'}\, V_{y'}.$$

---

2 En el muestreo de conglomerados, la estimación del número total de unidades de análisis (y') será una variable aleatoria que, por lo general, no es exacta e igual al valor verdadero (Y). Por lo tanto, la proporción de unidades que poseen el atributo deben ser considerada como una relación de variables aleatorias.

En las fórmulas anteriores, p es el coeficiente de correlación entre las variables Y eY. Representa la correlación de Y eY, no para las unidades elementales de análisis sino para las unidades usadas en el muestreo. Por ejemplo, si Y eY representan los ingresos de las personas en dos años distintos, pero la muestra es una muestra de conglomerados, luego el coeficiente de correlación (p) será la correlación entre los valores X_i e Y_i donde X_i es la suma de los ingresos de <u>todas</u> las personas en el i-ésimo conglomerado en el año de estimación e Y_i la suma correspondiente en el año base. Frecuentemente, P $S_{x'}$, $S_{y'}$ se denomina como la <u>covariancia de muestreo</u> entre x' e y' y se representa con el símbolo $S_{x',y'}$. Puede calcularse exactamente como la variancia pero reemplazando los cuadrados $(Y_i - \bar{Y})^2$ por los productos cruzados $(Y_i - \bar{X})\ (Y_i - \bar{Y})$ en los casos en que así ocurra.
Así, en el muestreo simple al azar tenemos:

(11.5)
$$PS_{x'}\ S_{y'} = S_{x'y'} = N^2\left(\frac{N-n}{N-1}\right)\frac{S_{y,y}}{n}$$

donde

(11.6)
$$S_{X,Y} = \frac{1}{N}\sum_{i}^{N}\ (y_i - \bar{y})\ (Y_i - Y).$$

Puede calcularse una estimación de $S_{X,Y}$ con la muestra usando

(11.7)
$$S_{X,Y} = \frac{1}{n-1}\sum_{i}^{n}\ (x_i - \bar{x})\ (y_i - \bar{y}).$$

La estimación correspondiente de p, que se representa mediante p', se obtiene con la ecuación (11.5) reemplazando los valores poblacionales por los valores muestrales de $S_{x'}$, $S_{y'}$ y S_{xy} y despejando luego p que pasa a ser p'. También puede calcularse p' directamente con la fórmula siguiente:

(11.8)
$$p' = \frac{\sum_{i}^{n}(x_i - \bar{x})\ (y_i - \bar{y})}{\sqrt{\sum_{i}^{n}(x_i - \bar{x})^2 \sum_{i}^{n}(y_i - \bar{y})^2}}$$

En una muestra estratificada, con estimaciones de los totales dados por x' e y'.

(11.9)
$$S_{x',y'} = \sum_{i} N_i^2\left(\frac{N_i - n_i}{N_i - 1}\right)\frac{S_{i\ x,y}}{n_i}$$

donde $S_{i:\,x,y}$ son las covariancias dentro de los estratos que se calculan exactamente en la misma forma si bien sólo con los valores dentro de cada estrato.

3.2 Ganancia que se obtiene con una estimación por relativos

Si examinamos la Ecuación (11.4), fórmula para calcular la variancia relativa de una estimación de un total,

(11.4) $$V^2_{x''} = V^2_{x'} + V^2_{y'} - 2pV_{x'}V_{y'},$$

Vemos que V^2, en el caso de la estimación por relativos ($V^2_{x''}$), puede expresarse como V^2, en el caso de la estimación más simple ($V^2_{x'}$), más el término ($V^2_{y'}$), menos el término ($2p\,V_{x'}V_{y'}$). Que obtengamos una ganancia o que suframos una pérdida con el uso de una estimación por relativos, en comparación con la estimación más simple (x'), depende de si $V^2_{y'} - 2pV_{x'}V_{y'}$ es mayor o menor que cero. Otra forma de expresar este mismo concepto es la siguiente:

1) Si $p > \frac{1}{2}\left(\frac{V_{y'}}{V_{x'}}\right)$, una estimación por relativos es más eficiente.

2) si $p < \frac{1}{2}\left(\frac{V_{y'}}{V_{x'}}\right)$, una estimación por relativos es menos eficiente.

3) Si $p = \frac{1}{2}\left(\frac{V_{y'}}{V_{x'}}\right)$, ambas estimaciones tienen el mismo error estándar.

3.21 Correlación alta:

Para ver las consecuencias de estos hechos en algunas situaciones comunes, consideremos el ejemplo de un censo de manufacturas que se efectuó en un cierto año seguido al año siguiente por una muestra. Deje x_i e y_i representar los valores de los embarques para la misma firma en la muestra en los dos años consecutivos. En este caso, $V_{y'}$ y $V_{x'}$ son casi iguales y $\frac{V_{y'}}{V_{x'}}$ es un número casi igual a la unidad.

Además, existirá una correlación muy alta entre X e Y, probablemente casi 0.90 ó 0.95. En consecuencia, una estimación por relativos resultará en una ganancia substancial en exactitud. La magnitud de la ganancia puede determinarse en la forma siguiente:
Si $V_{x'} = V_{y'}$, la ecuación (11.4) resulta

$$V^2_{x''} = 2V^2_{x'}(1-p)$$

y si p= .90, tenemos

$$v^2_{x'} = 0.20\ v^2_{x},$$

En otras palabras, el uso de una estimación por relativos produce una reducción del 80 por ciento en la variancia. Si p= .95, $v^2_{y'}$ es igual a 0.10 $v^2_{x'}$ y la reducción es del 90 por ciento. Interpretando los resultados en otra forma, la estimación por relativos es tan efectiva como usar una muestra 5 (ó 10) veces más grande.

3.22 Correlación baja:

Consideremos ahora la situación descripta en la sección 2.3 donde X es un subconjunto de Y. En tal caso, la correlación es posiblemente algo baja, al menos que $\frac{X}{Y}$ sea considerablemente grande por ejemplo, mayor que $\frac{1}{2}$. En la práctica, si $\frac{X}{Y}$ es menos del 20 por ciento aproximadamente, una estimación por relativos puede aumentar el error de muestreo aunque, por lo general, no mucho. Si $\frac{X}{Y}$ es más grande que un 40 ó un 50 por ciento, una estimación por relativos mejorará, por lo general, la eficiencia; cuanto más próximo sea ese valor a 100 mayor será la ganancia. Entre 20 y 40 por ciento, las diferencias entre ambos tipos de estimaciones serán muy pequeñas. Así, por ejemplo, en una encuesta de fuerza de trabajo, el uso de las estimaciones por relativos proporcionará probablemente un mejoramiento importante en la estimación del número de empleados (que representa una proporción considerablemente alta de la población adulta) pero resultará, posiblemente, en un cierto aumento en el error estándar de la estimación del número de desempleados.

3.3 Sesgo de una estimación por relativos

La estimación por relativos es una estimación sesgada. Puede demostrarse este hecho fácilmente construyendo una pequeña población con los valores x_i e y_i en cada elemento, tomando luego todas las muestras posibles de dos o tres elementos, y calculando por último los valores x' en cada muestra. Se verá inmediatamente que el promedio de las relaciones y' no es el promedio verdadero. Sin embargo, el sesgo tiende a ser despreciable en muestras moderadamente grandes. En la mayoría de las aplicaciones prácticas el sesgo es tan pequeño, comparado con la ventaja ganada en la reducción del error de muestreo, que se prefiere la estimación por relativos en lugar de una estimación insesgada.

3.4 Estimaciones consistentes

Una estimación por relativos es, a pesar de ser sesgada, una estimación consistente. Esto quiere decir que, si usamos una muestra suficientemente grande, podemos estar seguros de que la estimación resultará tan próxima al valor verdadero como queramos. Aumentando el tamaño de la muestra, no solo disminuye el error estándar sino que también se reduce el sesgo.

3.5 Límites de confianza

Con muestras razonablemente grandes, la estimación por relativos se disminuye normalmente (en el caso de poblaciones del tipo de las que frecuentemente encontramos en la práctica). En consecuencia, si podemos calcular el error estándar de la estimación por relativos, podemos construir para $\bar{x}''$ y x'' límites de confianza como los ya tratados para $\bar{x}'$ o x', o sea, podemos decir que tenemos una probabilidad del 68 por ciento de que la cifra verdadera estará en un intervalo de la estimación más o menos una vez el error estándar; una probabilidad del 95 por ciento de que la cifra verdadera estará dentro de más o menos dos veces el error estándar, etc.

3.6 Tamaño mínimo de muestra requerido

Las secciones 3.3 y 3.5 anteriores se refieren al hacho de que se necesitan muestras moderadamente grandes para que el sesgo sea despreciable y para proporcionar una distribución razonablemente normal de las estimaciones muestrales. ¿Cuándo es una muestra suficientemente grande? Se han sugerido las siguientes reglas de trabajo al respecto: Si el tamaño de la muestra es mayor de 30 y si los coeficientes de variación de $\bar{x}'$ e $\bar{y}'$ son, ambos, menos de 10 por ciento, luego el sesgo es despreciable y podemos asumir que corresponde aplicar la teoría de la distribución normal. La primera condición no implica que una estimación por relativos es necesariamente mejor que una estimación simple insesgada siempre que $n > 30$, sino que se requiere tener este tamaño de muestra antes de que las fórmulas del error de muestreo tengan el significado usual en función de los intervalos de confianza.

3.7 Fórmula del sesgo

Una aproximación del sesgo de una estimación de una relación de dos variables $r' = \frac{x'}{y'}$ es

$$\text{sesgo} \doteq \frac{1}{n}\left(V_{y'}^2 - \rho\, V_{x'}\, V_{y'} \right)$$

donde ρ y R se definen de acuerdo con lo expresado en la sección 3.1. Para la estimación de un total $x'' = \frac{x'}{y'}Y$, el sesgo es

$$\text{sesgo} \doteq \frac{X}{n}\left(V_{y'}^2 - \rho\, V_{x'}\, V_{y'} \right).$$

Aun con valores bajos de ρ este sesgo será pequeño comparado con el error estándar de x' siempre que la muestra sea razonablemente grande como para que $V^2_{y'}$ sea pequeño.

Estas fórmulas del sesgo se presentan aquí con fines analíticos. Nunca se usan para ajustar estimaciones. En los casos en que se esperara que el -

sesgo pudiera ser considerablemente grande, se aumentaría el tamaño de la muestra o se aplicaría un método distinto de estimación.

3.8 Peligro en el uso de estimaciones por relativos

Si se aplican separadamente estimaciones por relativos a un número grande de subgrupos de la población, con una muestra pequeña en cada subgrupo, el sesgo en el sugrupo puede acumularse y llegar a ser demasiado grande como para ignorarlo. Por ejemplo, supongamos que se clasifica una muestra relativamente pequeña de personas por grupos separados sexo-edad 300 personas divididas en grupos quinquenales de edad por sexo. Existirían unos 30 grupos de ese tipo. Supongamos que conocemos el total poblacional verdadero en cada uno de los 30 grupos. Para cualquier estadística en que estuviéramos interesados podríamos calcular una estimación por relativos separada para las personas en cada uno de los 30 grupos y luego obtener una estimación final sumando los 30 resultados. El tamaño promedio de la muestra en cada grupo sería 10. Dado que sólo existiría en cada uno de los grupos de edad una muestra pequeña para calcular una estimación por relativos, la acumulación de 30 estimaciones por relativos diferentes podría acarrear un serio sesgo. En tal caso, no es recomendable usar una estimación por relativos grupo por grupo.

TAREA DE ESTUDIO

Problema: Se ha seleccionado una muestra simple al azar del 10 por ciento de las unidades de vivienda de una aldea obteniéndose las 12 unidades de vivienda listadas a continuación. En cada unidad de la muestra se obtuvo información sobre el número de personas en la familia y las ganancias anuales totales; los resultados figuran abajo. Se sabe también, a través de otras fuentes independientes, que la población total de todas las familias en la aldea es 600 personas.

Unidad en la muestra	Total de personas	Ganancias totales
1	6	$7,000
2	6	8,000
3	5	3,000
4	8	10,000
5	4	2,000
6	2	1,000
7	4	2,000
8	5	3,000
9	1	1,000
10	7	8,000
11	4	1,000
12	5	6,000
Total	57	$52,000

Ejercicio 1. Estimar las ganancias totales en todos los hogares en la aldea usando un factor directo de expansión.

Ejercicio 2. Estimar las ganancias totales en todos los hogares en la aldea usando una estimación por relativos.

Ejercicio 3. Usar los resultados muestrales para estimar el coeficiente de variación de cada una de las estimaciones anteriores.

CONFERENCIA 12. EL MUESTREO EN LAS ENCUESTAS AGROPECUARIAS DE MEDICIONES OBJETIVAS

1. NECESIDAD DE LAS MEDICIONES OBJETIVAS

Los principios del muestreo expuestos en las conferencias anteriores se pueden aplicar extensamente, por lo general, en los programas de encuestas. Sin embargo, ciertos tipos de encuestas pueden requerir técnicas especiales de muestreo y recolección de los datos debido a la naturaleza de la encuesta o a la capacidad de los informantes para proporcionar respuestas correctas. En la Conferencia 12 se describen algunas técnicas especiales utilizadas en las encuestas agropecuarias.

En la mayoría de los países las estadísticas sobre superficies con cultivos individuales y sobre rendimientos de dichos cultivos se basan en informes periódicos preparados por los llamados reporteros de cosechas. En algunos países actúan como tales los productores agropecuarios u otros individuos que residen en las zonas rurales y tienen conocimiento de la agricultura local. Sus informes, de carácter voluntario, se remiten, por lo general, por correo. En otros países, los reporteros son funcionarios o agentes del gobierno. Por lo común los informes de estos últimos son menos exactos que los de los individuos que actúan a título personal debido, en parte, a que aquéllos deben casi siempre informar sobre áreas más extensas y, en parte, a que no tienen un contacto muy íntimo con la actividad agropecuaria. De cualquier modo, ya sea que estén preparados por agentes del gobierno o por individuos voluntarios, los informes mencionados están sujetos a sesgos que son frecuentemente de magnitud considerable y siempre difíciles de evaluar. Por ejemplo, investigaciones efectuadas en diversos países muestran que en la estimación de los rendimientos los reporteros (sobre todo, los agentes) acusan un sesgo hacia el rendimiento normal; en otras palabras, en los años buenos tienden a subestimar el rendimiento mientras que en los años malos tienden a sobre-estimarlo. Aunque los reporteros privados tienen, en cierta medida, la misma tendencia, se inclinan más a la subestimación con la creencia de que tal criterio les reportará alguna ventaja. En cambio tienden a sobreestimar las superficies cultivadas debido a la dificultad de hacer estimaciones apropiadas en relación con los áreas no cultivadas alrededor de los bordes del campo y las áreas, dentro de los campos, que no pueden ser cultivadas.

Para evaluar los sesgos en las estimaciones de la producción calculadas sobre la base de los informes de los reporteros se pueden usar datos de verificación de años anteriores. En el caso de cultivos como el tabaco o el algodón, que se elaboran antes de su uso, se puede obtener información sobre la producción en otras fuentes, por ejemplo, los elaboradores, y compararla con las cifras correspondientes dadas por los reporteros. Para otros cultivos se pueden usar, en forma similar, datos obtenidos de los mercados o embarcadores. Si esa información está completa (seguridad que, por lo común, no se tiene) y si los sesgos relativos se mantienen más o menos constantes de un año al otro, se pueden ajustar las estimaciones en el año actual sobre la base de esa experiencia anterior. Para otros cultivos, que en parte, al menos, se destinan al consumo local, alimento de ganado, etc., no existen datos de verificación. Si se dispone de datos censales, éstos pueden usarse como datos de referencia para ajustar los informes sobre tales cultivos. Sin embargo los datos del censo están también sujetos a sesgos de información. Además, los ajustamientos, tomando como referencia los datos censales, pierden progresivamente confiabilidad a medida que aumenta el lapso entre el último censo y el año corriente.

La experiencia en muchos países diferentes bajo condiciones distintas ha mostrado que los métodos subjetivos para estimar la producción, aun cuando exista otra información para ajustar las estimaciones, no pueden proporcionar resultados confiables. Si se necesitan estimaciones exactas e insesgadas, la única alternativa es establecer algún tipo de programa en el que se utilicen métodos objetivos de observación aplicados sobre una base de muestreo aleatorio. Tales encuestas son las llamadas "encuestas de mediciones objetivas" ya que los datos se recogen mediante la observación y medición o recuento verdaderos y no mediante métodos que dependen del juicio, buena memoria o educación de las personas que dan la información requerida. A pesar de que tal programa de encuestas de mediciones objetivas resulta relativamente costoso y difícil de llevar a cabo, los resultados que produce justifican generalmente el esfuerzo que se haga.

2. DISEÑO DE LA MUESTRA

Las consideraciones teóricas que afectan el diseño de la muestra, expuestas en las conferencias anteriores, son tan pertinentes al diseño de una encuesta de mediciones objetivas como a cualquiera otra encuesta.

2.1 Tipos de estimaciones requeridas

El especialista en muestras debe conocer si se requieren estimaciones para todo el país, para las provincias o distritos individuales o para algunas otras áreas administrativas. La afijación de la muestra se debe proyectar de modo que se obtengan estimaciones para las áreas deseadas con un nivel aceptable de confiabilidad. Asimismo, al proyectar el diseño de la muestra, debe considerarse si se necesita además una estimación del número de fincas (ya sea total o en relación con un cultivo dado).

2.2 Estratificación

Los estratos de primer nivel están formados, frecuentemente, por las -

áreas más pequeñas para las que se requieren estimaciones separadas. Se puede lograr una ganancia adicional en eficiencia haciendo una estratificación más profunda dentro de las áreas geográficas que rienen tasas de rendimiento del cultivo dado relativamente homogéneas. También se pueden usar otras bases de estratificación como tierras irrigadas y no irrigadas, variedades de cultivo, etc.

2.3 Afijación en los estratos

El estadístico debe tomar una decisión sobre la afijación de la muestra en los estratos. Una práctica acostumbrada es hacer la afijación proporcional a la superficie con el cultivo o grupo de cultivos que se investigan. Cuando se tiene algún conocimiento sobre las variancias relativas y/o los costos relativos de las labores de campo en los distintos estratos debe utilizárselo asimismo para la afijación de la muestra.

2.4 Muestreo dentro de los estratos

Debe decidirse el método de muestreo dentro de los estratos. Como se indicó en la sección 3 de la Conferencia 9, existen, por lo general, varios posibles diseños de muestra y unidades de muestreo. Para decidirse acerca del plan de muestreo el estadístico en la materia necesitará conocer qué materiales hay para construir el marco de muestreo y los diferentes tipos de datos que se requieren. Su elección podrá también verse afectada por otros factores como la disponibilidad de prsonal capacitado para cumplir el trabajo. Sin embargo, aun con las restricciones impuestas por estas consideraciones, en la mayoria de los casos tendrá un conjunto de posibles alternativas.

2.41 Etapas de muestreo y tipos de unidades de muestreo

En casi todas las aplicaciones prácticas se usarán, dentro de los estratos, varias etapas de muestreo. Por ejemplo, si los estratos son divisiones administrativas grandes, por ejemplo provincias, se podría seleccionar en la primera etapa una muestra de distritos y extraer en la segunda etapa una muestra de subdistritos dentro de los distritos muestrales. Cuando existen "aldeas" con límites bien identificables y que dan razón de toda la tierra, pueden servir convenientemente como unidades en alguna etapa del muestreo. La unidad última de análisis será, por lo común, una finca individual, el campo o terreno individual o (en algunos estudios que implican estimación de rendimientos) pequeñas parcelas dentro de los campos. Si la unidad de análisis es el terreno, podría seleccionarse las fincas en la etapa precedente.

2.42 Métodos para la selección de fincas y campos

Los siguientes ejemplos ilustran algunos procedimientos que pueden usarse para seleccionar fincas o campos en las etapas finales del diseño muestral. La selección de parcelas dentro de los campos aparece tratada en la sección 4.4 de esta conferencia.

1) Las fincas se pueden seleccionar de listas, si es que éstas existen o pueden prepararse sin mucha dificultad. Se necesitarían listas de fincas sólo en las unidades (aldeas, subdistritos, etc.) realmente seleccionadas en la muestra en la etapa precedente; se podrían, si fuera necesario, compilarlas como parte de la labor de campo. La selección de fincas puede hacerse con probabilidades iguales o con probabilidades proporcionales al tamaño (suponiendo que se tiene o se puede obtener esa información acerca del tamaño). La medida del tamaño podría se la superficie total declarada de la finca, la superficie total con un cultivo o grupo de cultivos particulares, etc.

   En forma similar, dentro de cada finca seleccionada se podría vcompilar una lista de campos y seleccionar una muestra. Una vez más, la selección podría hacerse con probabilidades iguales o con probabilidades proporcionales al tamaño.

2) Si existen mapas o fotografías aéreas se pueden usar para seleccionar los campos directamente sin tener que seleccionar primero las fincas. Un procedimiento para hacer esta selección es suponer sobre el mapa o la foto una cuadrícula donde se han colocado, al azar o siguiendo una pauta sistemática, algunos puntos, Cada campo sobre el que cae un punto se incluye luego en la muestra dando así a los campos probabilidades de selección proporcionales a sus tamaños. El procedimiento requiere, por supuesto, que los mapas o fotografías sean lo suficientemente detallados como para que el punto y el campo que le corresponde puedan localizarse en el terreno. (Este procedimiento no es fácil de adaptar si lo que se desea es estimar el número de fincas.)

3) Los segmentos de superficies son unidades de muestreo útiles para determinar cuales son las fincas y/o campos que se incluirán en la muestra. Esos segmentos pueden construirse ya sea con límites naturales que pueden localizarse en el terreno, ya sea con límites imaginarios dibujados en la fotografía o el mapa; la decisión al respecto depende de la situación particular de que se trate. Las fincas y/o los campos pueden asociarse con segmentos de superficie en una de las siguientes formas:

   a) Se podrían usar como unidades de muestreo de la primera etapa segmentos de superficie con límites imaginarios y seleccionar una muestra de segmentos para extraer luego los campos, dentro de los segmentos muestrales, como unidades de la segunda etapa en la forma descrita antes en (2).

   b) Un procedimiento alternativo sería incluir en la muestra todos los campos (o fincas) en los que cae, dentro de los límites del segmento, un punto definido de manera única. Con este procedimiento, los campos (o fincas) no se seleccionarían con probabilidades proporcionales al tamaño; las probabilidades de selección serían las mismas que las probabilidades de selección de los segmentos en los que caen los puntos. Este enfoque se conoce con el nombre de segmento abierto. Los segmentos determinan las unidades que se incluirán en la muestra si bien

los datos se tabulan para algunos campos (o fincas) que parcialmente están dentro del segmento.

El punto único debe ser definido con cuidado. Por lo común se designa como tal una esquina particular del campo (o finca). Debido a que los campos (o fincas) pueden no ser rectangulares, es posible que se necesite una regla especial para ubicar dicha esquina. Por ejemplo, si el punto único designado fuera la esquina noroeste, se lo podría definir ya sea (1) identificando los puntos de límite del campo (o finca) ubicado más al oeste y eligiendo entre todos ellos, como esquina noroeste, el situado más al norte; o (2) identificando los puntos del límite ubicados más al norte y eligiendo entre ellos el que está más al oeste. Si la unidad de análisis fuera la finca, la residencia del productor (suponiendo que tales residencias tuvieran una probabilidad de ser incluidas en la muestra) sería generalmente preferida como punto único por ser el punto más fácil de localizar. Tal vez sea aun de mayor utilidad combinar estas reglas. Por ejemplo, se podría utilizar la residencia del productor cuando el mismo vive en la finca y una esquina particular cuando no es así. En cualquier caso el punto debe ser definido de modo que sea realmente único (es decir, cada unidad debe tener un, y sólo un punto, asociado con ella y, en consecuencia, tener una, y sólo una, probabilidad de ser incluida en la muestra). Además debe ser bastante fácil de identificar.

c) Si la unidad de análisis es la finca, el enfoque del segmento ponderado resultará, por lo general, más eficiente que el del segmento abierto. Según este procedimiento se incluyen en la muestra todas las fincas que tengan alguna tierra en el segmento. En la estimación se ponderan los datos obtenidos en cada finca mediante un factor basado en la proporción de la finca completa que cae dentro de los segmentos. En casi todas las aplicaciones, el enfoque del segmento ponderado requiere que los segmentos tengan límites naturales que puedan identificarse en el terreno.

d) Otra posibilidad más es usar el llamado enfoque del segmento cerrado según el cual sólo se incluyen en la muestra los campos o partes de los campos que están dentro de los segmentos. Una ventaja de este procedimiento es que evita la dificultad de tener que definir la explotación. Por su puesto que si desea información basada en la exposición no resulta apropiado el enfoque del segmento cerrado ya que algunas explotaciones se extienden ciertamente más allá de los límites del segmento.

3. PROCEDIMIENTOS DE MEDICIONES OBJETIVAS PARA LA ESTIMACION DE SUPERFICIES

Dado que se sabe que los datos sobre superficies de tierra obtenidos a través de preguntas planteadas a los individuos en un cuestionario pueden ser muy inexactos, se han investigado otros medios para recoger esa información. El enfoque usual en las encuestas de mediciones objetivas consiste en seleccionar

Una muestra de superficies y trasladarse luego a las misma para proceder a medirlas directamente. Existen además métodos para obtener estimaciones objetivas de superficies que no requieren la medición directa de la tierra; por ejemplo, se puede medir la superficie en fotografías aéreas. Además de las mediciones se puede obtener otra información; por ejemplo, clasificación de la tierra en varias categorías según su uso (cultivo, pastoreo, virgen, etc., identificación del cultivo particular en cada fracción de tierra, etc.

3.1 Medición de la superficie de tierra

El primer paso para hacer mediciones directas de un terreno es preparar un plano o esquema de acuerdo con una escala. Para ésto debemos poder medir distancias y ángulos. Un plano preparado por un agrimensor usando equipo técnico resultará muy preciso. Por otra parte, un esquema dibujado por un individuo inexperto que midiera las distancias a pasos y los ángulos a simple vista no sería muy exacto. Entre ambos extremos existen muchos otros procedimientos para cumplir la tarea. Debemos hacer un balance entre el costo relativo y la exactitud relativa de los distintos procedimientos y seleccionar el método que proporcione un nivel aceptable de confiabilidad al costo más bajo.

Una vez preparado el plano se deberá establecer la superficie que cubre el mismo. Si el campo que se midió es de forma regular, acorde con una figura geométrica, por ejemplo un rectángulo, trapecio, etc., resulta relativamente fácil establecer la superficie que encierra el plano aplicando las fórmulas matemáticas conocidas. Usando el factor de expansión que corresponda se tendrá la superficie de tierra representada en el plano. Sin embargo, con frecuencia, la superficie es de forma "irregular"y es necesario usar otros métodos, por ejemplo, triangulación, planimetría, cuadriculado, recuento de puntos y pesaje de mapas.

3.11 Triangulación

En la triangulación se convierte el polígono dibujado en el plano en triangulos más simples. Existe un principio geométrico que dice que esto es siempre posible. (Los límites curvos se reemplazan, en forma aproximada, antes de la triangulación, por una serie de líneas rectas). Cada triángulo se mide y se calcula su superficie con las fórmulas estándar. Se trata de un procedimiento largo y tedioso que ha sido en gran parte reemplazado.

3.12 Planimetría

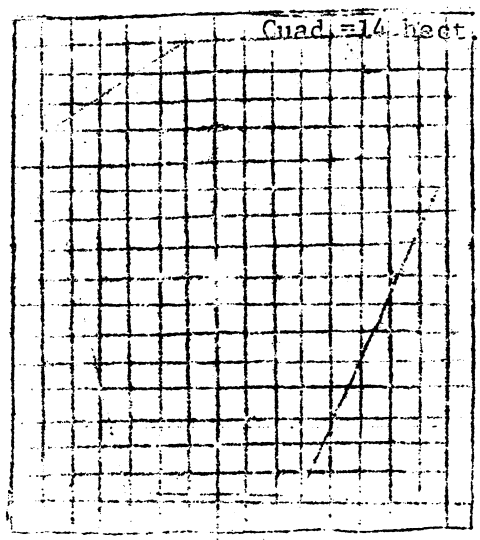
Un planímetro es un instrumento con el que se puede medir la superficie de una figura cerrada recorriendo los límites de la misma con un dispositivo similar a un lápiz. Un buen planímetro dará resultados muy exactos. Requiere, sin embargo, un operador adiestrado y mucho tiempo para cumplir el trazado.

1 Véase Estadística Agrícola! Estimación de superficies, S.S. Zarkovich (ed) vol. XXIV, No. 90 y 91 de Estadística, Instituto Interamericano de Estadística, para una exposición sobre las técnicas y la experiencia en varios países.

3.13 Cuadriculado

Básicamente una cuadrícula es un plano dividido en pequeños cuadros (por ejemplo, una hoja de papel común para gráficos). Para la medición de superficies los cuadrados se construyen de modo que cada uno sea equivalente a una porción determinada de superficie de acuerdo con la escala del trazado del plano. Sobre éste se puede colocar una cuadrícula transparente o imprimirse la cuadrícula sobre el papel y hacer el dibujo del plano directamente en ese papel. Para estimar la superficie representada por el plano se cuentan todos los cuadrados enteros y partes de cuadrados dentro del perímetro del dibujo construído según la escala convirtiéndose después ese número a su equivalente en función de la unidad de superficie apropiada.

Figura 1: MEDICION MEDIANTE CUADRICULADO



Si bien el cuadrado no es tan exacto como la planimetría se puede hacer en menos tiempo. Requiere sólo que el individuo sea capaz de contar exactamente y saber hacer la conversión de los cuadrados parciales a su equivalente en número de cuadrados enteros. Véase la figura 1 que presentamos arriba donde aparece una ilustración del método. Dentro del dibujo construído de acuerdo con una escala existen aproximadamente 159 cuadrados (inclusive los cuadrados parciales que se sobreponen a los límites); luego, como cada cuadrado representa 1/4 de hectárea, el campo mide al rededor de 40 hectáreas.

3.14 Recuento de puntos

El recuento de puntos, es en esencia igual al cuadriculado salvo que, en lugar de pequeños cuadrados, la cuadrícula contiene puntos espaciados uniformemente. Cada punto representa una superficie unitaria de acuerdo con la escala

del dibujo. Sólo se necesita contar los puntos que caen dentro del perímetro del dibujo para saber cual es la superficie. Si algunos puntos están sobre los límites sólo se cuenta la mitad de ellos.

3.15 Corte y pesajes de mapas

En este procedimiento se cortan cuidadosamente en pedazos los mapas o fotografías del área. Dichos pedazos representan categorías distintas de campos a lo largo de líneas trazadas por el encargado del trabajo de campo. Luego se pesa con cuidado cada porción. La estimación se basa en el peso del papel en cada categoría en relación con el peso de la superficie completa. Este procedimiento no es muy práctico; requiere tiempo y un instrumento para medir el peso, de alta precisión; además se debe usar, para dibujar el mapa, un papel de calidad uniforme.

3.2 Observación del uso de la tierra en una muestra de puntos o líneas

Algunos métodos de medición objetiva de superficies no requieren la medición directa de la tierra en sí. En su lugar se estima la proporción de tierra que pertenece a varias categorías mediante algunos medios objetivos y se multiplica por la superficie total conocida de tierra en el universo (provincia, distrito, etc.) para estimar así la superficie total en cada categoría Todos los procedimientos expuestos en la sección 3.2, salvo el último (el método último descrito en el párrafo 3.22) requieren mapas ó fotografías aéreas exactas y actualizadas; en consecuencia su utilidad está algo limitada en estos momentos. Sin embargo, dado el progreso que se observa en la toma de fotografías aéreas, es posible que éstos u otros métodos similares resulten cada vez más útiles en el futuro.

3.21 Observaciones en una muestra de puntos

Se selecciona una muestra de puntos y éstos se marcan en los mapas o fotografías aéreas. En la selección de la muestra de puntos se deben usar técnicas apropiadas para la estratificación y conglomeración a fin de hacer máxima la eficiencia del diseño. Por ejemplo, si el interés principal es estimar las superficies cultivadas, se deben aplicar tasas más altas de muestreo en aquellas partes del universo que se sabe están compuestas principalmente por tierras cultivadas.

Si se van ha estimar únicamente categorías amplias de uso de la tierra y se dispone de fotografías aéreas convenientes, es posible hacer las observaciones necesarias en forma directa en las fotografías. Sin embargo, para localizar cada punto muestral y registrar la cosecha en cultivo u otro uso de la tierra en ese punto.

Un autor ha sugerido que en las encuestas periódicas se identifiquen, de manera permanente en el terreno, los puntos mediante marcas al efecto para que resulten más fáciles de localizar. Las marcas no se colocarían en las ubicaciones exactas de los puntos muestrales ya que interferirían con las operaciones agrícolas; deberían, sin embargo, ubicarse en lugares cercanos y estar

provistos de un dispositivo óptico dirigido hacia los puntos muestrales. Este método no ha sido ensayado aún en el terreno. (Refiérase a "fixed-Point Sampling -A New Method of Estimating Crop Areas," por Thomas B. Jabine, Estadística No. 96/97, P. 501, Instituto Interamericano de Estadística -- 1967).

Una vez hechas las observaciones en la muestra de puntos se puede calcular una estimación insesgada de la superficie dedicada a un uso particular:

1) En cada estrato en que se muestrearonpuntos a una tasa constante, contar el número de puntos muestrales en cada categoría de uso de la tierra.

2) Multiplicar la superficie total conocida del estrato por la proporción de puntos muestrales dedicados a ese uso.

3) Sumar a través de todos los estratos.

3.22 Observaciones en una muestra de líneas

Se selecciona una muestra de líneas y se marcan las mismas en los mapas o fotografías aéreas. Como en el caso de los puntos, se deben utilizar técnicas apropiadas de estratificación y conglomeración para aumentar la eficiencia del diseño. El procedimiento acostumbrado dentro de las unidades últimas de muestreo es seleccionar una muestra de líneas paralelas espaciadas a intervalos iguales.

Mediante fotografías aéreas o mediante el corrido real de las líneas, el investigador determina la proporción de cada línea que cae en cada categoría de uso de la tierra. Se calculan luego estimaciones a partir de esas observaciones mediante un procedimiento totalmente análogo al descrito antes para la muestra de puntos.

Una forma relativamente económica aunque sesgada del muestreo de líneas implica la sustitución de una muestra probabilística de líneas por caminos. El investigador recorre luego con su automóvil la ruta indicada. El auto está equipado con un instrumento medidor de distancias. Cuando hace el recorrido el investigador anota y registra la distancia que bordea el camino según cada categoría de uso de la tierra que se está midiendo (cultivos específicos, tierras de labranza en general, pastos, bosques, etc.). Las estimaciones se preparan luego en la forma acostumbrada en el muestreo de líneas.

Esta última técnica es posible que resulte seriamente sesgada, en particular en las áreas donde la red de carreteras está espaciada, ya que la pauta de uso de la tierra a lo largo de los caminos es probable que difiera, en forma substancial, de la pauta general a través del área dada. De ser posible deben utilizarse preferiblemente las técnicas basadas en el muestreo probabilítico.

3.3 Uso de la estimación por relativos y del muestreo doble para mejorar la eficiencia /2.

Una vez completadas las mediciones de superficies en las fincas (o algunas otras unidades de análisis) muestrales, podemos estimar totales directamente con esos datos mediante el procedimiento de estimación que sea apropiado para el particular diseño muestral. Este procedimiento puede, sin embargo, mejorarse generalmente si además de hacer mediciones de superficies en una muestra de la población también se dispone de datos de superficie menos exactos y menos costosos (por ejemplo, obtenidos mediante entrevistas directas) relativos a la población total. Tales datos podrían ser, por lo común, los de un censo completo. Mediante la estimación por relativos podemos con frecuencia obtener estimaciones de totales poblacionales que serán más confiables que los que obtendríamos usando únicamente, ya sea las mediciones objetivas, ya sea las respuestas a entrevistas. El procedimiento es en esencia igual al expuesto en la sección 2.3 de la Conferencia 11. La característica X sería en este caso la medida verdadera de la tierra obtenida en un subgrupo de la población; la característica Y, el dato recogido en la entrevista.

Aún más útil y práctica es una técnica denominada muestreo doble (3) en el que se utiliza la técnica más económica para obtener los datos de una muestra relativamente grande de la población y la técnica más costosa para obtener los datos de una submuestra de la muestra básica. Una vez más se usa la estimación por relativos pero aquí la característica Y es la respuesta que se obtiene mediante la técnica menos costosa y se usa la estimación muestral del total en la población para la característic. Y en lugar de un total basado en una cobertura del 100 por ciento. 4 (ver página siguiente).

Comparado con el método basado únicamente en la medición de la superficie los métodos que aplican la estimación por relativos deberán preferirse si la ganancia en eficiencia más que supera el costo de obtener las observaciones complementarias mediante la técnica menos costosa (ya sea en la población total o, en el caso del muestreo doble, en una muestra más grande de la población). Los factores que deben considerarse son:

1) La fuerza de la relación entre los datos obtenidos mediante los dos métodos. La respuesta en la entrevista debe tener una alta correlación positiva con la medida de la superficie si se quiere obtener un mejoramiento significativo. Podemos razonablemente esperar que éste sea el caso.

2) El costo relativo de los dos métodos. Suponiendo que la correlación es suficientemente grande, la estimación por relativos reducirá el número de fincas que requieren medición de superficies para alcanzar un nivel dado de confiabilidad. El que esta reducción compense o no el costo de obtener respuestas mediante entrevistas depende, en parte, de las diferencias de costo entre los dos tipos de observaciones.

2 La estimación por regresión, que no ha sido tratada en estas conferencias, podría usarse también en forma similar.

Comparado con el método basado únicamente en las respuestas obtenidas en entrevistas, se preferirá el uso de la estimación por relativos si se considera que el sesgo en las respuestas recogidas en las entrevistas es suficiente como para justificar el gasto adicional de efectuar mediciones de las superficies. Para comprender la situación más claramente se necesita el concepto del error cuadrático medio (ECM). Recordamos, por las conferencias anteriores, que la variancia está basada en las diferencias entre las estimaciones (x'), derivadas de muestras, y el valor X que se obtendría si se hubieran recogido los datos de todos los miembros de la población usando las mismas técnicas. El error cuadrático medio, por su parte, está basado en las diferencias entre las estimaciones derivadas de la muestra, y el valor verdadero de la cantidad que se está midiendo (X_T). Si la técnica de recolección de los datos es insesgada, $X=X_T$ y el ECM resulta equivalente a la variancia; si la técnica es sesgada, el ECM es igual a la variancia más el cuadrado del sesgo ($X - X_T$), o sea.

$$ECM = S_{x'}^2 + (X - X_T)^2. \qquad (12.1)$$

Para un costo establecido, los datos se pueden obtener mediante entrevistas en una muestra de un tamaño dado. Para el mismo costo, los datos se pueden obtener mediante entrevistas en una muestra más pequeña combinados con mediciones objetivas en una submuestra de dicha muestra. Las estimaciones basadas en la muestra más grande de entrevistas tendrán un ECM específicado que incluya como componentes un sesgo y además una variancia. Las estimaciones por relativos basadas en la combinación de datos de entrevistas y mediciones objetivas tendrán un sesgo más pequeño pero una variancia mayor. El ECM puede ser ya sea mayor, ya sea menor, que el ECM basado sólo en una muestra más grande de entrevistas dependiendo de la variabilidad en la población, el costo relativo de los dos procedimientos (que determina los tamaños relativos de las muestras), el tamaño relativo de los sesgos (o la efectividad del procedimiento de estimación por relativos para reducir el sesgo), etc. El especialista en muestras deberá considerar todos estos factores al distribuir los recursos disponibles entre los dos procedimientos. Su meta es hacer mínimo el ECM para un costo dado (o hacer mínimo el costo de obtener un nivel aceptable de confiabilidad).

4. MEDICION OBJETIVA DEL RENDIMIENTO

La meta de la medición objetiva del rendimiento es, por lo general, estimar el rendimiento de un cierto cultivo sobre una base unitaria (por ejemplo, quintales por hectáreas, bushels por acre, etc.) Para estimar la producción total es necesario tener también una estimación de la superficie total cultivada con el producto en cuestión. En algunos casos solamente se estima el -

3 El muestreo doble es una técnica estadística útil en una variedad de situaciones en las que una característica de interés, que es difícil o costosa de determinar, está marcadamente correlacionada con otra característica que puede determinarse, en relación, más fácilmente y a un costo menor.

4 Véase nota al pie número 2 en la sección 3.3 de la Conferencia 12.

rendimiento mediante medios objetivos aun cuando las estimaciones de las dos características, rendimiento y superficie, deberían estar basadas en mediciones objetivas.

El procedimiento general para hacer mediciones objetivas del rendimien to (llamado frecuentemente "corte de cultivos") consiste en usar un proceso aleatorio para seleccionar superficies (conocidas comúnmente como parcelas) cultivadas con el producto en cuestión y segar y pesar el producido de cada cada una de esas parcelas en la misma época o en una cercana a la que se cosecha el resto del campo. 5 Cada cultivo tiene características diferentes y aún el mismo cultivo puede comportarse diferentemente en distintas regiones del mundo. En consecuencia no existe un conjunto específico de reglas que pueda aplicarse a todos los cultivos ni siquiera al mismo en diferentes lugares. Trataremos, sin embargo, en términos generales, algunos de los factores que se deben tener en cuenta al proyectar un programa como éste y describiremos algunas de las técnicas aplicadas en el pasado.

4.1 Estudios piloto

Debido a que la información recogida acerca de otros cultivos o acerca del comportamiento del cultivo en cuestión en otros países no es transferible en forma directa a nuestra propia situación, deben llevarse a cabo estudios piloto antes de establecer un programa de mediciones objetivas del rendimiento. Estos estudios pueden proporcionar importante información sobre la mayoría de las cosas que necesitan ser consideradas, tales como la variabilidad de muestreo, el tamaño óptimo y forma de la parcela, los procedimientos de cosecha, problemas de personal y materiales necesarios para cumplir el trabajo, etc. Son útiles asímismo como recurso de adiestramiento para el personal que eventualmente tendrá a su cargo la operación completa. Sobre la base de los estudios piloto el investigador puede desarrollar un plan de muestreo y procedimientos de campo apropiados a las condiciones a que estará sujeta la encuesta..

Una vez decidido el procedimiento es conveniente, por lo general, ponerlo en operación sólo gradualmente y, cuando ha llegado a su implantación completa, mantenerlo durante unos pocos años simultáneamente con el procedimiento que va a reemplazar.

El programa ya existente, no importa lo inadecuado que sea, no debe suspenderse hasta tanto haya sido probado suficientemente el método nuevo y juzgado, en términos claros, superior y factible de realizar. 6 Después de que haya establecido la superioridad y factibilidad del nuevo método el mismo puede servir como una base para evaluar el sesgo del método anterior, operación que no sería posible a menos que se hubieran mantenido simultáneamente ambos durante algunos años.

---

5 Las mediciones objetivas se utilizan también para preparar pronósticos de rendimientos sobre la base de observaciones hechas en períodos anteriores de la estación agrícola. Como los procedimientos de muestreo que se utilizan en los pronósticos de rendimientos son bastante similares a los usados en la estimación de los rendimientos, sólo se tratarán estos últimos en la presente Sección.

Esto es particularmente importante para los usuarios de los datos que están interesados en examinar las diferencias o tendencias a lo largo de varios años; los usuarios deben saber en que medida las diferencias observadas en los datos son simplemente el resultado de diferencias en las técnicas de medición.

4.2 Variabilidad

Debemos tener alguna idea de la variabilidad en el rendimiento del cultivo que se mide a fin de preparar un plan inteligente. Se mencionan a continuación dos aspectos de la variabilidad que son de interés:

1) La variabilidad relativa de los rendimientos en parcelas de diferentes tamaños y formas.

2) La magnitud relativa, para una parcela de tamaño y forma dados, de la variación entre parcelas dentro de un campo.

Para decidirse con el tipo de parcela que usará el investigador debe hacer un balance entre la variabilidad y el costo. Tratará de seleccionar la parcela que dé el grado deseable de confiabilidad al costo más bajo si bien algunos otros factores (por ejemplo, problemas de personal) pueden obligarlo a elegir una que no es precisamente la mejor en función de los costos y las variancias.

La experiencia ha mostrado en casi todos los casos que la variación entre terrenos es considerablemente mayor que la variación dentro de los terrenos. Como consecuencia, el número de parcelas seleccionadas dentro de cada terreno muestral debe ser pequeño de modo que los recursos disponibles puedan destinarse más eficientemente al muestreo de tantos campos diferentes como sea posible. En realidad, en algunas investigaciones, el número óptimo de parcelas ha sido una por campo. 7 Por su puesto que si deseamos estimar con la muestra la variabilidad dentro del terreno se necesita un mínimo de dos parcelas; aun así, el investigador podría decidirse a tener sólo una parcela por terreno si la componente dentro del terreno de la variancia es muy pequeña comparada con la componente entre terrenos.

4.3 Tamaño y forma de la parcela

Para cultivos que están esparcidos en el terreno o plantados en filas muy cercamente espaciadas (por ejemplo, granos pequeños o heno), se ha sabido usar en estudios anteriores parcelas circulares, triangulares, cuadradas y rectangulares.

---

6 En realidad es posible que en cualquier caso sea necesario tener que mantener el programa existente, en particular, si se requieren datos para áreas administrativas diferentes a las que se utilizan para hacer estimaciones con los datos objetivos. Además, el programa existente puede permitir la recolección de información sobre una serie de cultivos que no son económicamente importantes en un grado suficiente como para justificar un costoso programa de mediciones objetivas.

Para cultivos en filas ampliamente espaciadas (por ejemplo, maíz o algodón) la elección lógica han sido las parcelas rectangulares: el ancho está casi siempre expresado en filas y el largo en pies (o metros, etc.).

Juntamente con la forma de la parcela se debe establecer algún método para su señalamiento. Para ésto se han usado con éxito, en el caso de parcelas pequeñas, los marcos rígidos u otros dispositivos. Las cuerdas, cadenas, etc., son fáciles de transportar pero su colocación en el terreno es más difícil si el encargado de hacerlo tiene que medir y clavar estacas en las esquinas,etc. Si la parcela es de forma triangular puede utilizarse, bastante fácilmente, una cadena cerrada con aros en los tres vértices; el mismo recurso se puede adoptar también, suponiendo que la cadena toma la forma de un triángulo rectángulo, para señalar parcelas rectangulares mediante una combinación apropiada de triángulos. Las parcelas grandes, por lo general, se jalonan con estacas o clavijas, socas y una cinta de medir.

A medida que aumenta el tamaño de la parcela la variabilidad entre las parcelas disminuye; sin embargo, como la contribución a la variancia total de la componente dentro del terreno es usualmente despreciable en relación con las otras fuentes de la variancia, se prefieren, por lo general, desde un punto de vista práctico, las parcelas pequeñas. En ellas un sólo hombre, casi siempre , puede hacer el trabajo, puede colocar más rápidamente un marco transportable que jalonar una parcela grande, puede cosechar el cultivo en menos tiempo, y está obligado a manejar menos material.

Desafortunadamente la experiencia ha mostrado que casi siempre las parcelas pequeñas producen estimaciones seriamente sesgadas. Las razones de ésto no son suficientemente claras aunque aparentemente existen dos factores que son mayormente responsables:

1) En la localización de la parcela en el campo es más fácil que el encargado de la operación se deje afectar por la condición del cultivo al establecer la ubicación precisa de una parcela pequeña.

2) En una parcela pequeña se agudiza el problema de decidir si se deben contar o no dentro de la parcela las plantas que están sobre los límites ya que el perímetro de una parcela pequeña es mayor en relación con la superficie que el perímetro de una parcela grande. La tendencia general parece ser incluir plantas que deberían excluirse y, en consecuencia, sobre estimar en forma consistente el rendimiento. En una parcela pequeña,aun una sóla planta incluída erróneamente puede afectar seriamente los resultados.

4.4 Localización de la parcela en el campo

Para localizar las parcelas en el terreno se han propuesto diferentes métodos. Cualquiera sea el que se use, es importante que el personal de campo

7 Teóricamente el número óptimo de parcelas no necesita ser un número entero. Por supuesto que, en el orden práctico, el resultado teórico debe reducirse a un entero.

comprenda claramente cómo debe hacerse y se deben realizar verificaciones a fin de comprobar si dicho personal está cumpliendo las instrucciones. De otro modo, es casi seguro que se introducirá en el procedimiento un sesgo subjetivo debido al trabajo del personal de campo.

Desde un punto de vista teórico sería conveniente dividir el terreno entero en parcelas del tamaño y forma decididos antes y seleccionar al azar el número requerido de parcelas. Sin embargo, ésto no es por lo general práctico. Un método aplicado que en general resulta práctico si el terreno es rectangular (o si se puede cerrar convenientemente en un rectángulo) es localizar puntos al azar dentro del terreno y establecer luego al rededor de esos puntos las parcelas muestrales en una forma pre-establecida. Para cada parcela que se va ha localizar el procedimiento es el siguiente:

1) El encargado de la operación en el terreno selecciona un número al azar "x" entre cero y n, siendo n_1 la longitud total de una dimensión del terreno (o del rectángulo en el que se ha encerrado el terreno); selecciona asimismo otro número al azar "y" comprendido entre cero y n_2, siendo n_2 la longitud total de otra dimensión. Si se tratara de un cultivo en filas, la primera dimensión estaría por lo general expresada en función de filas.[8] En otros casos, las dimensiones se expresarían en función de unidades, por ejemplo, metros, o en función de pasos.

2) Comenzando en una esquina pre-establecida, el trabajador de campo mide directamente o a pasos (o contando las filas) la distancia x a lo largo del lado apropiado del terreno (o del rectángulo que lo encierra); luego, perpendicularmente a ese lado, mide directamente o a pasos la distancia y dentro del terreno.

3) Si el trabajador de campo está todavía dentro de los límites del terreno procede a marcar el punto aleatorio (por ejemplo, haciendo un hoyo con el tacón y clavando un jalón). Si ha salido del terreno (por supuesto que seguiría estando dentro del rectángulo que encierra el terreno). Utiliza otro par de números aleatorios y repite el proceso.

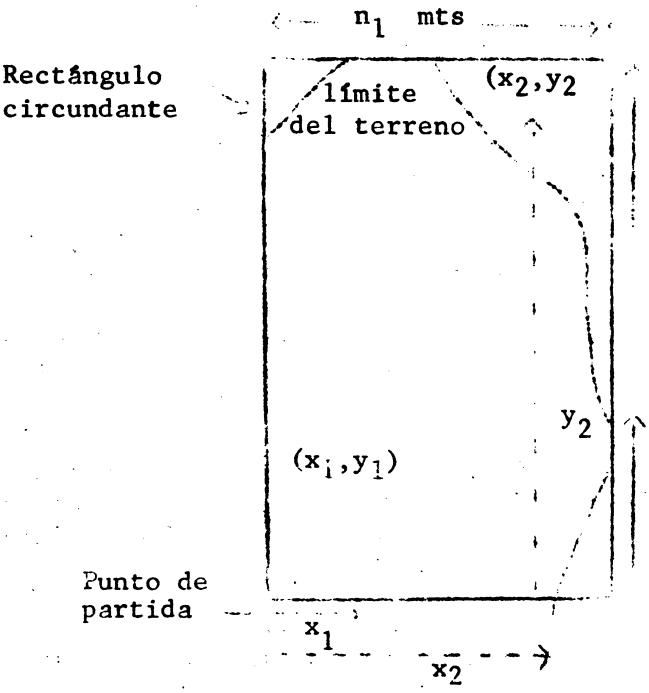
4) Partiendo de ese punto el encargado de la operación establece la parcela. Si es de forma circular debe usar el punto aleatorio como centro del círculo. Si es rectángular o triangular, debe usar el punto para ubicar un vértice o ángulo determinado precisamente, por lo común, en forma tal que la parcela se extienda lejos del punto aleatorio en la dirección en que ha estado encaminado.

La figura 2 ilustra este procedimiento. En este ejemplo el punto (x_1 y y_1) cae dentro del terreno y es, por lo tanto aceptado. El punto (x_2 y y_2 cae fuera del terreno y es, por lo tanto, rechazado. La parcela tendría que extenderse, por lo general, hacia la derecha y hacia arriba en relación con el punto muestral.

---

8 Se seleccionaría luego el número al azar entre 1 y el número total de filas en el terreno (n_1).

Figura 2. LOCALIZACION DE PUNTOS AL AZAR DENTRO DE UN TERRENO



Una dificultad de este esquema es que permite que las parcelas se superpongan a los límites del terreno; cualquiera de las reglas distintas posibles que podrían aplicarse en estos casos plantea algún problema. Consideremos, por ejemplo, un campo sembrado de maíz de 200 hileras de ancho y 108 metros de largo. Supongamos que la parcela ha de medir 4 hileras de ancho por 6 metros de largo. Supongamos además que la coordenada hilera seleccionada es 198 y que la coordenada de longitud es 95. Partiendo de la intersección de ambas coordenadas la parcela sobrepasaría en un metro y una hilera los límites del campo (la parcela se inicia al terminar el 95 i-ésimo metro pero incluye la fila 198). Las reglas posibles que se podrían adoptar para salvar esta situación incluyen:

1) Dar instrucciones al encargado de la tarea de que cultive únicamente la parcela (parcial de 3 hileras por 5 metros y que por supuesto registre esas medidas en el formulario. Se podría luego calcular una estimación insesgada del rendimiento de este terreno utilizando el factor de expansión apropiado. En este ejemplo se podría llevar a la práctica el procedimiento bastante fácilmente; sin embargo, si el terreno fuera de forma irregular o la parcela circular o triangular, el encargado podría encontrar dificultades para estimar la fracción de la parcela en el terreno.

2) Dar instrucciones al encargado de la tarea de que considere a las hileras como si estuvieran numeradas en forma circular y similarmente la longitud. Así, en este ejemplo, la hilera 1 del campo sería la cuarta de la parcela y el primer metro de cada hilera completaría la longitud de la parcela. Este procedimiento es también insesgado, sin embargo no resultaría práctico en ningún caso salvo cuando las parcelas fueran rectangulares dentro de campos de forma rectangular. Además, sería difícil de explicar al trabajador de campo común. Por último no se acomoda al ususal concepto de una parcela como un pedazo contiguo de tierra.

3) Dar instrucciones al encargado de la tarea de limitar la selección al azar de modo- que sólo puedan resultar elegidos números que no determinen esta situación, o, en otras palabras, rechazar las parcelas que sobrepasen los límites y seleccionar en ese caso otro par de coordenadas. De hacer así, en el ejemplo, el encargado restringiría la selección de las hileras a los números entre 0 y 94. Este procedimiento es evidentemente sesgado ya que los bordes del terreno(en el ejemplo, las cuatro primeras y las cuatro últimas hileras y los seis primeros y los seis últimos metros) tienen una probabilidad menor de estar en la muestra que el resto del terreno. Si el rendimiento tiende a ser mayor o menor que el promedio alrededor de los bordes del terreno, las estimaciones del rendimiento basadas en este método serán sesgadas. Sin embargo, éste es el procedimiento más sencillo. Si las superficies relativas en los bordes son más pequeñas que el resto del campo o si no existen razones para creer que el rendimiento es diferente en esos lugares, se puede recomendar este método con preferencia a los procedimientos insesgados aunque más difíciles.

4.5 Procedimiento para recoger la cosecha

Si las parcelas son pequeñas el trabajador de campo probablemente efectuará la tarea él mismo, segando el producto y pesándolo en el terreno. Tomará luego una pequeña submuestra para enviarla a la oficina central para secado. (Constituye siempre una buena práctica devolver el resto del producto al finquero.) Si las parcelas son suficientemente grandes, puede ser conveniente cosecharlas con el mismo método que el productor agropecuario usará en la cosecha regular y, si es posible, al mismo tiempo. Esto exigirá la cooperación y ayuda del finquero.

4.6 Ajustamiento a la producción verdadera

El método que aplica el técnico para cosechar las parcelas pequeñas y procesar el producto da por lo común una tasa más alta de rendimiento que la resultante de los procedimientos de cosecha usuales usados por el productor. Esto se debe a que en la cosecha por los métodos normales las pérdidas son mayores. En el caso de algunso cultivos dichas pérdidas pueden ser importantes. Además no es posible cosechar todas las parcelas al mismo tiempo o inmediatamente después de la fecha de la cosecha. Si el encargado espera demasiado tiempo para iniciar el corte de las plantas, casi ciertamente encontrará algunos campos cosechados antes de que él llegue; en consecuencia necesita comenzar a cosechar las parcelas en algunos campos cuando todavía no ha

madurado el producto. Estos dos factores harán que las estimaciones sean sesgadas si no se introducen algunos ajustamientos. (La cosecha de parcelas pequeñas mide lo que se conoce generalmente como rendimiento biológico.)

Un método de ajustamiento consiste en seleccionar una submuestra de campos de superficie conocida y cosecharlos a nombre del productor usando los procedimientos normales. Esto proporciona una base para ajustar los datos recogidos en las parcelas cosechadas. Un método similar, apropiado para algunos cultivos (por ejemplo cultivos para forraje que se sacan del campo en balas), es hacer los arreglos para pesar la totalidad del cultivo en una submuestra de campos cuando el productor lo saca del campo pero permitiendo que él mismo haga la cosecha cuando quiera y como quiera.

Otro método de ajustamiento es recoger las espigas o granos que ha dejado caer la cosechadora para estimar directamente las pérdidas en el terreno. Las pérdidas en el terreno estimadas por unidad de superficie se restan luego del rendimiento biológico estimado para obtener el rendimiento real. Este procedimiento tiene la ventaja de no necesitar que el trabajador de campo esté presente al efectuarse la cosecha - una consideración importante ya que varios productores de distintos campos muestrales pueden todos decidir cumplir la cosecha en el mismo día. Desafortunadamente la experiencia ha mostrado que los problemas de estimar las pérdidas en el campo son mayores que los de estimar la producción biológica original.

Como ya se ha mencionado es conveniente que las parcelas muestrales se cosechen lo más cercanamente a la fecha en que se cosecha el resto del campo; sin embargo, ésto no puede hacerse siempre en todos los campos. Un objeto de un estudio piloto sería determinar qué ajustamiento deberían hacerse, si alguno, por las diferencias de fechas entre esas cosechas. Para muchos cultivos no se necesitan ajustamientos debido a que los cultivos han llegado esencialmente a su completo crecimiento antes de una u otra fecha y por lo tanto sólo están en el proceso de pérdida de humedad.

Debe efectuarse un ajustamiento adicional por el contenido de humedad. Un procedimiento común consiste en secar el material de las parcelas (o una submuestra del mismo) hasta que ha llegado a un contenido nulo o casi nulo de humedad y pesarlo. Esta "pesada en seco" puede ajustarse luego a cualquier contenido de humedad deseable. Para muchos cultivos, se dispone de especificaciones sobre un contenido estándar de humedad. Si el material secado es sólo una submuestra de la parcela, deben pesarse separadamente la parcela entera y la submuestra en el campo enseguida después del corte. Luego se seca la submuestra y se pesa. El peso en seco de la parcela entera puede a continuación estimarse usando la relación del peso en seco con respecto al peso húmedo de la submuestra.

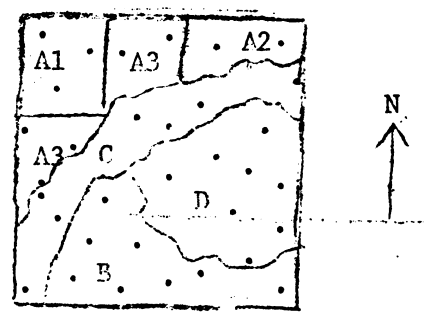
4.7 Consideraciones operacionales

Antes de poner en marcha un extenso programa de mediciones objetivas del rendimiento deben resolverse numerosos problemas prácticos. En éstos se in-

cluyen la disponibilidad de mano de obra, la disponibilidad de medio para el secado de los cultivos, necesidades de equipo, la necesidad de coordinar las actividades de los trabajadores de campo con los planes de los productores para la cosecha de sus campos, etc. El problema de conciliar los calendarios de trabajo puede ser muy difícil, en particular, cuando es posible que el cultivo esté listo para su cosecha al mismo tiempo en una superficie amplia. Como se dijo antes, una razón importante para efectuar estudios piloto es obtener información acerca de esos problemas de orden práctico.

TAREA DE ESTUDIO

Problema A: El dibujo que aparece a continuación simula un segmento delineado en una fotografía aérea.



El segmento contiene un total de 100 hectáreas divididas en cuatro categorías de acuerdo con el uso de la tierra. Las categorías son:

Tierra de labranza

A1	Maíz	B	Pastos
A2	Trigo	C	Bosques
A3	Otra tierra de cultivo	D	Tierra virgen

SD ha colocado una cuadrícula de tres puntos sobre el segmento a utilizarse al estimar la cantidad de terreno por categorías de uno.

Ejercicio 1. Estimar el número de hectáreas en este segmento que se usan como tierra de labranza.

Ejercicio 2. Estimar el número de hectáreas ocupadas por pastos.

Ejercicio 3. Estimar el número de hectáreas ocupadas por bosques y de tierra virgen.

Ejercicio 4. Estimar la proporción de la tierra de labranza dedicada al cul-

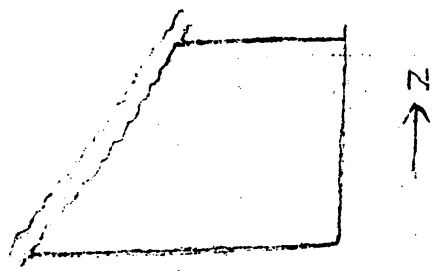
tivo de maíz. ¿En qué forma difiere básicamente esta estimación de las de los ejercicios 1 a 3?.

Problema B. En el esquema anterior las marcas sobre los límites este y oeste del segmento subdividen los límites en 40 unidades. Usando como guía dichas marcas ubique dos líneas al azar a través del segmento que sean paralelas a los límites norte y sur.

Ejercicio 5. Use esas líneas paralelas para estimar las cantidades ya estimadas en el problema A.

Ejercicio 6. Prepare, para cada cantidad, la distribución de las estimaciones obtenidas mediante varios ensayos o varias personas.

Problema C. El dibujo siguiente muestra un campo bordeando un rio.



Ejercicio 7. Trace un círculo alrededor de la esquina correspondiente al punto único de acuerdo con cada una de las definiciones dadas abajo. Ubique la letra apropiada (a,b,c) al lado de cada círculo.

a) Esquina noroeste - Identifique los puntos de los límites que están ubicados más al norte. La esquina noroeste es el punto más al oeste de esos puntos.

b) Esquina noroeste- Identifique los puntos de los límites que están ubicados más al oeste. La esquina noroeste es el punto más al norte de esos puntos.

c) Esquina suroeste - Identifique los puntos de los límites que están ubicados más al sur. La esquina suroeste es el punto más al oeste de esos puntos.

Problema D. Mediante entrevistas en una muestra al azar (seleccionada sin reposición) de 24 fincas extraídas de una población de 96 fincas se han recogido datos sobre la superficie total de tierra de labranza cosechada. En una submuestra de 8 de esas fincas seleccionadas al azar sin reposición se han efectuado mediciones objetivas. Los datos aparecen en el cuadro siguiente.

Unidad	Hectáreas de tierra de labranza cultivada	
	Entrevista (Y)	Mediciones Objetivas (X)
1	14	14.4
2	79	-
3	46	-
4	112	116.1
5	46	-
6	92	-
7	29	-
8	40	41.9
9	12	-
10	78	80.4
11	66	-
12	43	-
13	39	-
14	91	93.9
15	17	16.8
16	68	-
17	100	-
18	87	-
19	74	75.4
20	64	-
21	78	-
22	40	42.6
23	22	-
24	55	-

Ejercicio 8. Estimar el total de tierra de labranza cosechada usando únicamente los datos de las entrevistas. Estimar la variancia de este total estimado.

Ejercicio 9. Estimar el total de tierra de labranza cosechada usando únicamente los datos de las mediciones objetivas. Estimar la variancia de esa estimación.

Ejercicio 10. Usando las fórmulas siguientes estimar el total de tierra de labranza cosechada y la variancia de esa estimación utilizando ambos tipos de datos y estimación por relativos.

$$x'' = N \frac{\bar{x}_2}{y_2} \bar{y}_1$$

$$s^2_{x''} = N^2 \left[\left(1 - \frac{n_2}{n_1}\right) \frac{s^2_{r'}}{n_2} + \left(1 - \frac{n_1}{N}\right) \frac{s^2_x}{n_2} \right]$$

donde n_1 = tamaño de la muestra grande de entrevistas

n_2 = tamaño de la muestra de mediciones objetivas

$$\bar{x} = \frac{1}{n_2} \sum^{n_2} x_i$$

$$\bar{y}_2 = \frac{1}{n_2} \sum^{n_2} y_i$$

$$s^2_{r'} = y^2_x + (r')^2 \; s^2_y - Pr' \; s_{xy}$$

$$s^2_x = \frac{1}{n_{2-1}} \sum_{i=1}^{n_2} (x_i - \bar{x}_2)^2$$

$$s^2_y = \frac{1}{n_{2-1}} \sum_{i=1}^{n_2} (y_i - \bar{y}_2)^2$$

$$s_{xy} = \frac{1}{n_{2-1}} \sum_{i=1}^{n_2} (x_i - \bar{x}_2)$$

$$r' = \frac{\bar{x}_2}{y_2}$$

LISTA SELECCIONADA DE REFERENCIAS

1. Cochran, William G. Sampling Techniques. Segunda edición. John Wiley and Sons, Nueva York, 1963.

2. Dirección del Censo de los Estados Unidos. The Current Population Survey Reinterview Program, Some Notes and Discussion. U..S. Government Printing Office. Washington, D.C., 1963. (Documento técnico No. 6)

3. Dirección del Censo de los Estados Unidos. The Current Population Survey- A Report on Metho-dology. U.S. Government Fринting Office, Washington, D.E. 1963. (Documento Técnico No. 7).

4. Hansen, Morris H.; Hurwitz, William N.; y Madow, Willian G. Sample Survey Methods and theory. vol. 1: Methods and Applications; vol. II: Theory. Hohn Wiley and Sons, Nueva York, 1953.

5. Instituto Interamericano de Estadística (IASI) . Estadística Agrícola: Estimación de Superficies. S.S. Zarkovich (ed.). vol. XXIV, No. 90 y 91 de Estadística. Washington, D.C., 1966. (Publicado originalmente en Inglés por la Organización de las Naciones Unidas para la Agricultura y la Alimentación bajo el título Estimation of Areas in Agricultural Statistics, Roma, 1965).

6. Kish, Leslie. Survey Sampling. John Wiley and Sons, Nueva York, 1965.

7. Naciones Unidas. Oficina de Estadística. Manual de Encuestas sobre hogares - Guía práctica para Investigación del nivel de vida. Nueva York 1964. (Estudios de métodos, Serie F., No. 10)

8. Neter, John y Wasserman, William. Fundamental Stadistics for Business and Economics. Allyn and Bacon, Boston, Massachusetts, Estados Unidos 1961.

9. Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO). Calidad de los datos Estadísticos, por S.S. Zarkovich. Roma, 1968.

10. Organización de las Naciones Unidas para la Agricultura y- la Alimentación (FAO). Estimación de Rendimientos Agrícolas, V.G. Panse. Roma, 1954

11. Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO). Los Métodos de Muestreos y los Censos, por S.S. Zarkovich. Roma, 1967.

12. Sampford, M.R. An Introduction to Sampling Theory. Oliver and Boyd, Edimburgo y Londres, 1962.

13. Sukhatme, Pandurang V. Teoría de Encuestas por Muestreo con Aplicaciones. Fondo de Cultura Económica, Méjico, D.F., 1956.

14. The RAND Corporation. A Million Random Digits. The free Press, Glencoe Illinois, Estados Unidos, 1955.

15. Yates, Frank. Sampling Methods For Censuses and Surveys. Tercera edición Hafner Publishing Company, Nueva York, 1960.

GLOSARIO DE TERMINOS

Afijación de la muestra:

Método para determinar cómo se distribuirá la muestra. En el muestreo estratificado se refiere, por lo general, a la determinación del número de unidades que se selecciona en cada estrato. En el muestreo de conglomerados, se refiere a la decisión acerca del número de conglomerados que se selecciona y al tamaño de la muestra que se extrae en cada conglomerado. ((Conferencias 8 y 10).

Afijación óptima de la muestra:

Sistema de selección de una muestra de manera que se produzca el error estándar mínimo para un tamaño de muestra constante o para un costo constante. Se usa tanto en el muestreo estratificado como en el muestreo de conglomerados. (Conferencias 8 y 9)

Característica:

Una variable que toma diferentes valores posibles en las distintas unidades individuales de muestreo o análisis. En una encuesta por muestra, Observamos o medimos los valores de una o más características en las unidades que forman la muestra. Por ejemplo, observamos) la superficie de tierra cultivada con arroz , o el número de cabezas de ganado en una finca. (Conferencia 2).

Coeficiente de Variación:

El error estándar relativo; es decir, el error estándar como una proporción de la magnitud de la estimación. En estas conferencias hemos representado el coeficiente de variación (Conferencia 4). Con la letra V (en otros textos se usa esta letra para la variancia).

Desviación Estándar:

El error estándar en una muestra simple al azar de tamaño 1. (Conferencia 3).

Elemento: Véase unidad de Análisis

Error cuadrático medio:

Medida que expresa al alcance de la diferencia entre las estimaciones derivadas de la muestra y el valor poblacional verdadero que se trata de estimar. Si las estimaciones son insesgadas, el error cuadrático medio y la variancia son equivalentes.

Error de Muestreo: Véase Error Estándar.

Error Estándar:

Una medida de la magnitud en que difieren las estimaciones de varias muestras con respecto al valor esperado Con un tamaño de muestra razonablemente grande, la distribución de los resultados muestrales en todas las muestras posibles es aproximadamente normal y se pueden hacer afirmaciones, en términos de probabilidad, acerca de la medida en que se espera que la muestra se aproxime al valor esperado - expresándose las probabilidades en términos del error estándar. El error estándar se representa generalmente con la letra griega S (conferencia 3). Véase Variancia.

Estadística:

Una cantidad calculada usando las observaciones muestrales de una característica, con el propósito, por lo general, de extraer alguna inferencia acerca de la población La característica puede ser cualquier variable asociada con un miembro de la población, por ejemplo, edad, ingreso, condición de empleo, etc. La cantidad puede ser un total, un promedio, una media, o algún otro percentilo, o también una tasa de cambio, un porcentaje, una desviación estándar o cualquier otra cantidad cuyo valor se desea estimar en la población. (Conferencia 2).

Estimación:

Cantidad numérica calculada con los datos de la muestra y que tiene por fin proporcionar información sobre un valor poblacional desconocido.

Estimación consistente:

Una estimación de un tipo que (mientras posiblemente sesgada) se acerca cada vez más y más al valor verdadero que se está estimando cuando aumenta el tamaño de la muestra; el ejemplo más común es una estimación por relativos. (Conferencia 11).

Estimación insesgada:

Un tipo de estimación que tiene la propiedad de que el promedio de tales estimaciones calculadas en todas la muestras posibles de un tamaño dado sea igual al valor verdadero. (Conferencia 3).

Estimación por relativos:

Un método de estimación con los datos de la muestra usando la relación $\frac{x}{y}$ donde tanto x como y estan basadas en los datos muestrales, o una variación de esta fórmula. Se le usa generalmente en una de dos formas. Primero para producir una estimación de una relación cuando ésta es la estadística de interés. Segundo, para producir una estimación de una cantidad mediante la fórmula $\frac{x}{y}$ Y, donde x e y son estimaciones muestrales e Y un valor conocido independientemente. Versiones más elaboradas de las estimaciones por relativos pueden tomar la forma $\sum \frac{x_i}{y_i} Y_i$, donde i representa los estratos o grupos diferentes de población. (Conferencia 11).

Estimador: Véase fórmula de estimación.

Estratificación:

El proceso de dividir una población en grupos con el objeto de seleccionar una muestra separada en cada grupo. Cada grupo se hace por lo general lo más homogéneo internamente. Los grupos se llaman estratos. Conferencias 7 y 8)

Fórmula de estimación:

Una fórmula matemática usada para calcular una estimación. (Conferencia 2).

Función del costo:

Una expresión matemática que muestra el costo de efectuar una encuesta en función de los tamaños de las muestras y los costos unitarios. (Conferencia 8).

Información independiente:

Datos conocidos antes o simultáneamente a la realización de la encuesta que no están basados en la encuesta pero que pueden utilizarse para mejorar el diseño de la encuesta. Esos datos pueden usarse para la estratificación, para decidir las probabilidades de selección, o para estimar los resultados finales con los datos de la muestra. (Conferencia 9).

Intervalo de confianza: Un intervalo por arriba y debajo del valor estimado que puede esperarse que contenga el valor verdadero con una probabilidad conocida, suponiendo que no exista sesgo. (Conferencia 3).

Lista:

Una población en la que las unidades de muestreo han sido numeradas o

identificadas en alguna forma; la lista de unidades puede ser la base para la selección de una muestra. (Conferencia 2). Véase también Marco de muestreo .

Marco de muestreo:

El conjunto de todas las unidades de muestreo de donde se selecciona la muestra. El marco puede ser una lista de personas o unidades de vivienda; un archivo de registros o tarjetas perforadas; un mapa subdividido, etc.

Muestra:

Un subconjunto de una población. En la forma usada en estas conferencias, se refiere siempre a una muestra de probabilidad, es decir, una muestra en la que cada elemento de la población tiene una probabilidad conocida de selección. (Conferencia 1).

Muestra autoponderada:

Una muestra en la que cada elemento de la población tiene la misma probabilidad de selección, aun cuando se podrían haber usado probabilidades desiguales en las distintas etapas del muestreo. Por ejemplo, podrían haberse seleccionado conglomerados con PPT y, luego, dentro de un conglomerado seleccionado, extraída la muestra en forma tal que se diera a cada elemento en ese conglomerado la misma probabilidad de selección que la de los elementos seleccionados en otros conglomerados (conferencias 8 y 10).

Muestra de conglomerados:

Un sistema de muestreo donde las unidades de análisis de la población se consideran agrupadas en conglomerados, seleccionándose luego una muestra de conglomerados. Los conglomerados seleccionados son los que determinan las unidades que finalmente formarán la muestra. La muestra puede incluir todas las unidades de los conglomerados seleccionados o una submuestra de unidades en cada conglomerado seleccionado. (Conferencias 9 y 10).

Muestra de superficies:

Un tipo de muestra (por lo general, una muestra polietápica) en la que las unidades de muestreo son superficies individuales de terreno (segmentos) que pueden definirse en un mapa. Los segmentos cubren toda la superficie de Tierra incluida en la encuesta; los segmentos no se superponen; y los límites de cada segmento deben estar claramente definidos de modo que puedan ser reconocidos e identificados por los entrevistadores en el terreno. Con frecuencia los segmentos son conglomerados de las unidades de análisis; por ejemplo, conglomerados de fincas o de unidades de vivienda. Cada unidad de análisis debe estar asociada a uno y sólo un segmento. (Conferencia 9).

Muestra simple al azar:

(Llamada también muestra al azar sin restricciones): El tipo más simple de muestreo. Dado un tamaño de muestra n, cada una de las posibles combinaciones de n unidades elementales que puede formarse con una población de N unidades tiene la misma probabilidad de selección que cualquier otra combinación de n unidades. Además, cada elemento tiene la misma probabilidad de selección que cualquier otro elemento. (Conferencias 2,3,4 y 5)'

Muestreo con reposición:

Un procedimiento de selección de una muestra que consiste en seleccionar primero un elemento de la población, reemplazarlo en la misma, hacer una segunda selección, reemplazarlo nuevamente antes de la tercera selección y continuar así hasta efectuar n selecciones. Con este método de selección una unidad determinada puede estar incluida en la muestra más de una vez en realidad hasta n veces. (conferencia 3).

Muestreo de conglomerados proporcional:

Un sistema en el que se seleccionan conglomerados con probabilidades variables y se submuestrea dentro de los conglomerados seleccionados. La muestra que se obtiene es una muestra autoponderada (Conferencia 10).

Muestreo estratificado:

El método de muestreo que se aplica cuando el universo ha sido estratificado. Al menos debe seleccionarse una unidad en cada estrato. La probabilidad de selección puede variar de un estrato a otro. (Conferencia 7 y 8).

Muestreo estratificado proporcional:

Sistema de selección de una muestra estratificada en el cual se usa la misma probabilidad de selección de cada estrato. (Conferencia 8).

Muestreo polietápico:

El tipo más común de muestreo de conglomerados. En este método se selecciona una muestra de conglomerados y luego, dentro de cada conglomerado seleccionado, se toma una submuestra de unidades. Si la submuestra de unidades constituye la última etapa de selección de la muestra, el diseño se denomina muestra bietápica (aun cuando cada una de esas unidades puede contener más de una unidad de análisis, como ocurre en el muestreo de superficies). Si la submuestra es a su vez un conglomerado de unidades donde se realizará una nueva selección, el diseño es trietápico, o cuatrietápico, etc. (Conferencias 9 y 10).

Muestreo sin reposición:

Un sistema de selección de una muestra que consiste en seleccionar una -

unidad, y sin reemplazarla, seleccionar las restantes continuando así el proceso hasta llegar a tener seleccionadas n unidades diferentes. Con este procedimiento, una unidad puede estar incluída en la muestra una sola vez. (Conferencia 3).

Muestreo sistemático:

Un método de selección de la muestra en el cual la población está listada en un cierto orden seleccionándose para la muestra cada k-ésimo elemento. (Conferencia 6).

Multiplicador finito:

El término en la fórmula de la variancia de una muestra simple al azar que refleja el efecto de la proporción de la población que integra la muestra. Este factor es igual a $\frac{N-n}{N-1}$ y aproximadamente igual a $1-\frac{n}{N}$. (Conferencia 5).

Población:

Cualquier conjunto de unidades (o elementos) claramente definido para el que se calculan las estimaciones. Los elementos pueden ser personas, fincas, semillas, manzanas, condados, firmas comerciales, etc. En su mayor parte, nuestra exposición se ha referido al muestreo de poblaciones finitas - que contienen un número finito de elementos. (Conferencia 2).

Población finita: Véase población

Probabilidad de selección:

La probabilidad que tiene cada unidad de ser incluida en la muestra. (Conferencia 2).

Probabilidad proporcional al tamaño (PPT

Un método de selección de una muestra en el que las unidades se seleccionan con probabilidades desiguales de selección, siendo la probabilidad de cada unidad proporcional a una medida del tamaño. La medida del tamaño de una unidad es un número asignado previamente a esa unidad, antes de la selección, que se considera está altamente correlacionado con la estadística que se estima. La expresión probabilidad proporcional al tamaño se abrevia con las letras PPT. (Conferencia 10).

Sesgo:

Diferencia entre el valor esperado de un estimador y el valor poblacional verdadero que se trata de estimar.

Unidad primaria de muestreo (UPM):

Las unidades que constituyen el marco de muestreo para la primera etapa

de una muestra multietápica.

Unidad de análisis:

Una unidad para la que deseamos obtener datos estadísticos. Las unidades pueden ser personas, familias, fincas o firmas comerciales; pueden ser también tarjetas perforadas, productos fabricados mediante algún proceso mecánico, etc. (Conferencia 2).

Unidad de muestreo:

Las unidades que se seleccionan. Pueden ser iguales o diferentes a las unidades de análisis. Por ejemplo, para obtener información acerca de las personas se podría usar un listado completo en un censo o un registro y seleccionar directamente una muestra de personas. Sin embargo, también se podría seleccionar una muestra de familias e incluir en la encuesta todas las personas de las familias seleccionadas. En forma similar, se podrían seleccionar edificios completos e incluir a todas las personas que residen en los edificios que forman parte de la muestra. La elección de la unidad de muestreo más eficiente es una consideración importante dentro del diseño de una encuesta. (Conferencia 2 y 9).

Universo: Véase población.

Valor esperado:

El valor promedio de las estimaciones muestrales a través de todas las muestras posibles.

Variancia:

El cuadrado del error estándar representado por lo general mediante el símbolo S^2 con un subíndice para indicar la estadística a la que se refiere. El mismo término se usa también sin subíndice para el cuadrado de la desviación estándar. Cuando existe alguna posibilidad de confusión, se usa la denominación variancia de muestreo para el cuadrado del error estándar, y variancia de la población para el cuadrado de la desviación estándar. (Conferencia 3).

Variancia relativa:

Cuadrado del coeficiente de variación que se representa generalmente como v^2. (Conferencias 5 y 11).