DIRECCION GENERAL DE ESTADISTICA Y CENSOS

MINISTERIO DE INDUSTRIA Y COMERCIO

Nº 10

OCTUBRE
1971

REVISTA DE ESTUDIOS Y ESTADISTICAS

SERIE ECONOMICA Nº 5

SAN JOSE, COSTA RICA

Revista de ESTUDIOS y ESTADISTICAS N° 10
Octubre de 1971
Serie Económica N° 5

Publicación de la

DIRECCION GENERAL DE ESTADISTICA Y CENSOS

República de Costa Rica

América Central

San José Apartado 10163

D I R I G E:

LIC. RENE SANCHEZ BOLAÑOS
Director General de Estadística y Censos

COLABORAN EN LA SELECCION DE LOS ARTICULOS Y PREPARACION DE LA REVISTA:

ARTURO MAYNARD DE CESPEDES	Subdirector de Estadística y Censos
ELADIO CORDERO DIAZ	Jefe, Departamento Estadísticas Sociales
FELIPE CHIN FONG	Jefe, Departamento Estadísticas Económicas
ADRIAN CARTIN CAMBRONERO	Jefe, Departamento de Censos
MARIO MURILLO MURILLO	Jefe, Departamento Técnico

LIZARDO GARCIA VALVERDE
Jefe Sección de Publicaciones

La Dirección recibe con agrado los artículos que se le remitan para su publicación y se reserva el derecho de incluirlos o no; en caso de publicarlos no implica necesariamente la conformidad de la Dirección General de Estadística y Censos con los criterios o puntos de vista expresados por el autor.

Queda autorizada la reproducción de los artículos publicados en esta Revista, si se menciona su origen y con el ruego de que se envíe a la Dirección ejemplares de la publicación que los reproduzca.

DISEÑO DE LA MUESTRA PARA ENCUESTA DE GRANOS BASICOS 1970-1971

Por :

J. G. BAPTISTA (FAO-GAFICA)
MARIO MURILLO M. (D.G.E.C.)
JESUS JIMENEZ G. (D.G.E.C.)

I - INTRODUCCION

1. - Esta encuesta forma parte del programa mínimo de encuestas agropecuarias anuales, que la Dirección General de Estadística y Censos inició en 1964. De este programa se han realizado ya las siguientes encuestas: Encuesta de café (1964); Encuesta de granos básicos (1965); Encuesta de ganado y aves de corral (1968); Encuesta de granos básicos (1969); y Encuesta de ganado y aves de corral (1970).

2. - La estructura de la muestra para la presente encuesta fue prácticamente la misma que se adoptó para las encuestas anteriores, es decir, una estructura mixta con enumeración de áreas de muestreo y fincas individuales. Sin embargo se introdujeron algunas alteraciones en la estratificación de fincas y en los métodos de selección de áreas que parece conveniente dejar señaladas en un nuevo informe metodológico. Así, por ejemplo se consideró un estrato nuevo de fincas medias y la selección de áreas de muestreo se hizo con probabilidad proporcional al número de viviendas y no al número de fincas como en encuestas anteriores, ya que no se disponía al momento del listado censal de fincas por segmentos.

II - DEFINICION DE LA POBLACION

3. - El universo o población de la presente encuesta se compone de todas las explotaciones dentro de las fronteras de Costa Rica, con las tres excepciones siguientes:

(a). Explotaciones con área total inferior a una manzana.

(b). Explotaciones localizadas en zonas urbanas que de acuerdo con el censo de población de 1963 tenían menos de 40% de población rural.

(c). Todas las explotaciones localizadas en el distrito de Talamanca habitado por población indígena y de muy difícil acceso.

Estas exclusiones reducirán grandemente el costo total de enumeración mientras que sólamente tendrían un pequeño efecto en las estimaciones de totales. De acuerdo con el censo agropecuario de 1963 la

contribución de las fincas menores de una manzana, al área y producción de cada uno de los tres cultivos era:

Cultivo	TOTAL DEL PAIS		FINCAS MENORES DE 1 MANZ. %			
	Area (Ha)	Produc. (TM)	Area (Ha)	Produc. (TM)	Area	Produc.
Arroz	50.959	40.296	403	362	0.8	0.9
Maíz	54.041	56.794	912	900	1.7	1.6
Frijol	44.422	16.235	554	274	1.2	1.8
TOTAL	149.422	113.325	1.869	1.536	1.2	1.4

III - DISEÑO DE LA MUESTRA

(a) TAMAÑO DE MUESTRA

4. - Consideraciones al respecto del coeficiente de variación del universo, del presupuesto asignado a la encuesta, de los costos promedios de empadronamiento, del tiempo para completar el trabajo de campo, y de las disponibilidades de empadronadores y principalmente de supervisores permitieron llegar a un tamaño aproximado de la muestra de cerca de 2.000 fincas, distribuidas de la siguiente manera:- (a). 500 fincas clasificadas como grandes (b). cerca de 500 fincas de tamaño medio; y (c), aproximadamente 1.000 fincas pequeñas. El criterio adoptado para la clasificación de las fincas en grandes, medias y pequeñas se da más adelante en capítulo sobre estratificación.

5. - Con este tamaño y distribución de la muestra se espera obtener errores de muestreo de cerca de 6-7 por ciento para renglones tales como el de producción total de arroz y maíz. Para la producción de frijoles el error de muestreo será un poco más alto y aún más elevado para otros renglones asociados con un pequeño número de fincas, como por ejemplo ciertos items sobre el uso de tecnología.

6. - El razonamiento que permitió llegar a la cifra aproximada del error relativo de muestreo se da a continuación para el arroz, como ejemplo.

7 - Estimaciones aproximadas de la producción nacional de arroz colocan la cifra, en el orden de 1.500.000 quintales de los cuales aproximadamente 30% serían producidos en fincas grandes, cerca de 40% corresponderían a fincas medias y los restantes 30% a fincas pequeñas.

8. - De acuerdo con estudios anteriores se estima que el coeficiente de variación del universo de fincas pequeñas es aproximadamente igual a 2.00, o sea que para una muestra de 1000 fincas arrojaría un error relativo de la estimación de producción total correspondiente a este universo, de cerca de 200/ 1.000= 200/33= 6.6%. Sin embargo dado que la muestra de fincas pequeñas no es seleccionada absolutamente al azar sino que está condicionada por la selección anterior de segmentos censales, el error relativo de muestreo viene aumentando de acuerdo a la siguiente expresión: $-V = V\,1 + (\bar{n}-1)P$ en donde $\bar{n}$ representa

el número promedio de fincas seleccionadas en cada segmento censal y P es el coeficiente de correlación interno. Para el presente caso se consideró que ñ=6 y P=0.2.

Sustituyendo estos valores en la expresión se obtiene V = 13% en vez del valor apuntado de 6.6%.

9. - Para el universo de fincas medias se espera que el error relativo correspondiente a producción de arroz no sea superior al 6% y como fácilmente se entiende será cero para el universo de fincas grandes.

10. -Los valores absolutos correspondientes a aquellos errores relativos serán respectivamente·

(a) Fincas Medias: -(0.06) (0.40) (1.500.000) = 36.000 qq.
(b) Fincas Pequeñas: -(0.13) (0.30) (1.500.000) = 58.500 qq.

El error absoluto total será, por lo tanto, igual a 36.000 + 58.500 = 95.000 quintales, lo cual representa en términos relativos (95.000)(¢1.500.000) x 100 = 6.3%.

(b) ESTRATIFICACION DE FINCAS

11. -Dado que la distribución de frecuencia de las fincas de acuerdo al área sembrada de arroz y maíz era bastante asimétrica en 1963 (censo agropecuario) se procedió en primer lugar a hacer una estratificación de las mismas en los tres grupos siguientes:

(a) Estratos de fincas grandes (a ser enumeradas individualmente y 100 por ciento).
(b) Estratos de fincas medias (a ser enumeradas en 100 por ciento en U. P. M. seleccionadas).
(c) Estratos de fincas pequeñas (a ser enumeradas en áreas de muestreo seleccionadas).

El criterio adoptado para definir el tipo de finca se basó en el área bajo cultivo de cada uno de los tres productos, así:

12. -Se consideró como fincas grandes todas aquellas que cumplian cualquiera de los siguientes requisitos:

(i) Arroz: - Area sembrada de 50 manzanas o más.
(ii) Maíz: - Area sembrada de 50 manzanas o más.
(iii) Frijol: - Area sembrada de 20 manzanas o más.

Estos límites fueron fijados usando la fórmula (4) de la nota: "Sobre la determinación aproximada del límite inferior de tamaño para identificar fincas grandes en la selección de muestras", en que los tres valores de Y correspondientes a las áreas totales de maíz, arroz y frijol se tomarán iguales a las estimaciones obtenidas en la encuesta de 1965.

13. -Previo a la enumeración de campo se identificaron y localizaron las fincas grandes en todo el territorio nacional por medio de un listado obtenido por comparación de listados facilitados por agencias Bancarias regionales, las agencias de extensión agrícola y el Consejo Nacional de Producción (C. N. P.).

14. -El mismo criterio de área bajo cultivo sirvió para definir las fincas medias, de manera que se consideraron englobadas en este grupo todas aquellas fincas que cumplían cualquiera de los siguientes requisitos:

(i) Arroz: - Area sembrada de 10 a 49 manzanas.
(ii) Maíz: - Area sembrada de 10 a 49 manzanas.
(iii) Frijol: - Area sembrada de 5 a 19 manzanas.

De las mismas fuentes mencionadas para fincas grandes se obtuvieron listados de fincas medias localizadas dentro de U. P. M., seleccionadas.

15. -Como fincas pequeñas se consideraron todas aquellas con una área total igual o superior a una manzana que tenían siembras de cualquiera de los tres cultivos y no estaban todavía incluídos en los dos estratos anteriores. Estas fincas fueron seleccionadas en varias etapas de muestreo con probabilidad uniforme e igual a 1/30.

(c) FORMACION DE UNIDADES PRIMARIAS DE MUESTREO (U. P. M.)

16. -Las U. P. M. fueron formadas por distritos administrativos o cuando estos eran de tamaño reducido en términos del número de fincas, por la unión de dos o más distritos. Se elaboró primero un listado de todos los distritos del país con exclusión de aquellos netamente urbanos.

Este listado que fue organizado por regiones agrícolas contenía además la siguiente información: - (i). Número de fincas con cultivos anuales y respectiva área bajo esos cultivos de acuerdo al censo agropecuario de 1963; (ii). Localización geográfica del distrito por provincia y cantón y (iii). Una estimación de la relación población rural/población urbana.

17. -De la lista así formada se eliminaron después el distrito de Talamanca y todos aquellos en que la relación población rural/población urbana era inferior a 60/40 = 1. 5. Con los restantes distritos se formaron entonces las U. P. M. imponiendo la condición de que cada una debía contener por lo menos 30 fincas con uno o más de los tres cultivos en estudio. Para llenar este requisito fue necesario en varios casos combinar dos o más distritos administrativos.

18. -Las U. P. M. formadas de acuerdo a lo descrito anteriormente fueron después clasificadas en dos tipos:

(i) U. P. M. auto representadas y (ii). U. P. M. no auto-representadas. Tomando en cuenta el número de fincas que en 1963 cultivaban por lo menos uno de los tres granos básicos. Fueron también consideradas como U. P. M. auto = representados, tres distritos de la Provincia de Puntarenas que dado el incremento del cultivo de arroz en esta región se creyó que ameritaba ser estudiados separadamente.

19. -Para identificar las U. P. M. auto-representadas se procedió como sigue:

(i) Tomando en cuenta los costos de trabajo de campo, el presupuesto asignado a la encuesta y las

disponibilidades en personal de campo se estimó que era posible empadronar la muestra de fincas medias y fincas pequeñas en cerca de 70 U. P. M. lo que representa una fracción de muestreo aproximadamente igual a 1/3.

Después de algunos ajustes durante el proceso de selección de la muestra el número de U. P. M. seleccionados quedó finalmente en 66.

(ii) Dado que el número de fincas pequeñas del universo es aproximadamente de 33. 000, el límite inferior para que un distrito sea clasificado como U. P. M. auto-representada es, por lo tanto, 470=33. 000/ 70 fincas pequeñas. En una primera aproximación se identificarán entonces todos los distritos con 470 o más fincas pequeñas para entrar en la muestra con probabilidad igual a uno.

(iii) La operación fue repetida con las fincas y U. P. M. restantes bajando cada vez el límite de fincas y aumentando el número de U. P. M. auto-representadas. Se llegó finalmente al límite de 300 fincas con el cual se verificó que aunque la operación fuera repetida no conduciría a la identificación de nuevas U. P. M. auto-representadas.

20. -Por lo tanto todos los distritos administrativos con un número estimado de fincas pequeñas, igual o superior a 300 sembrado por lo menos uno de los tres cultivos, fueran identificados e incluídos en la muestra con certeza. Se separarán así 32 distritos a los cuales, con una fraceión de muestreo de 1/30, correspondían aproximadamente 564 fincas dejando 1. 101-564=537 para ser seleccionadas en U. P. M. no auto-representadas.

21. -La constitución de la muestra se quedó finalmente como se indica en el euadro siguiente:

Región	N° de fincas Universo	Muestra	Fracción Muestreo	N° de U. P. M. Seleccionadas		N° de Fincas Seleccionadas	U. P. M. N. A. R.
				A. R.	N. A. R.	U. P. M. A. R.	
L L	3. 470	116		0	10	0	116
C P	6. 730	224	f-1/30	6	8	120	104
C G	22. 832	761		26	16	481	280
TOTAL	33. 022	1. 101		32	34	604	500

1. LL;- Región lechera; CP:- Región de cultivos permanentes; CG:- Región de cereales y ganado.

2. A. R. - U. P. M. auto-representadas; N. A. R. - U. P. M. no auto-representadas.

(d) ESTRATIFICACION DE LAS U. P. M., NO AUTO-REPRESENTADAS

22. -Tomando en cuenta la importancia relativa del cultivo de cada uno de los tres granos básicos por U. P. M. y la localización geográfica de estas se formarán dentro de cada región agrícola estratos de U.

P. M. conteniendo aproximadamente el mismo número de fincas pequeñas. El número de estratos formado fue igual a la mitad del número de U. P. M. a seleccionar en cada región, permitiendo así la estimación insesgada del error de muestreo. Dentro de cada estrato se organizaron las U. P. M. por orden creciente del número de fincas lo que combinado con la selección aleatoria sistemática adoptada tiene el efecto de una estratificación por tamaño.

23. -El número de estratos formados y su composición se indica en el cuadro siguiente:

Cuadro III. - NUMERO DE ESTRATOS DE LA MUESTRA POR REGION AGRICOLA

Región Agrícola	N° Distritos Administrativos	N° de U. P. M.	Promedio Dis/U. P. M.	N° de Estratos	Promedio U. P. M. /Estrato	U. P. M. Seleccionadas/Estrato
L L	65	45	1.44	5	9.0	2
C P	78	51	1.53	4	12.8	2
C G	79	77	1.03	8	9.6	2
TOTAL	222	173	1.28	17	10.2	2

Continuación

Región Agrícola	N° de Fincas	Promedio Fincas/U. P. M.	Promedio Fincas/Estr.	N° Fincas Seleccionadas	Promedio Seleccionadas/Estrato
L L	3470	77.1	694.0	116	11.6
C P	3122	61.2	780.5	104	13.0
C G	8381	108.8	1.047.6	280	17.4
TOTAL	14.973	86.5	880.8	500	14.7

(e) SELECCION DE LAS U. P. M. DENTRO DE LOS ESTRATOS

La selección de las U. P. M. se hizo independientemente dentro de cada estrato y en cada región agrícola, con probabilidad proporcional al número de fincas. El proceso que se siguió se indica en el cuadro IV el cual corresponde al estrato I de la región agrícola C. G.

Cuadro IV. - SELECCION DE U. P. M. EN EL ESTRATO I DE LA REGION AGRICOLA C. G.

Código del Distrito	U. P. M. N°	N° Fincas por U. P. M.		N° de Selección	Probabilidad de Selección
		Individual	Acumulado		
21009	1	114	114		
50903	2	126	240		
50905					
60702	3	151	391		
11205	4	185	576	401. 3	185/544. 5=0. 33976
50504	5	251	827		
50302	6	262	1. 089	945. 8	262/544. 5=0. 48118
TOTAL	-	1. 089	-	-	-

Interval de selección = 1089/2=544. 5

N° de arranque al azar = 401. 3

25. -Dado que el número total de fincas del estrato es 1. 089 y se pretende seleccionar 2 U. P. M. la fracción de muestreo es 2/1. 089=1/544. 5. Para identificar la primera U. P. M. seleccionada se obtuvo de las tablas de Números al Azar un número de arranque comprendido entre 000. 1 y 544. 5. Este fue 401. 3 lo cual confrontado con la columna del número acumulado de fincas del Cuadro IV permitió identificar como primera U. P. M. seleccionada la N° 4 o sea el distrito 11205. Para identificar la segunda U. P. M. se adicionó 401. 3 al interval de muestreo, obteniéndose el número 945. 8=401. 3+544. 5 el cual está dentro del intervalo de fincas correspondientes a la U. P. M. N° 6.

26. -Las probabilidades de selección están indicadas en el cuadro. Se calcularán usando la fórmula correspondiente a la selección con probabilidad proporcional al tamaño y con reposición, es decir P=m N/ N, en donde m es el número de U. P. M. a seleccionar del estrato, (en nuestro caso m=2), N es el número de fincas de la iésima U. P. M. seleccionada y N=S(N) es el número total de fincas del estrato.

(f) SELECCION DE FINCAS PEQUEÑAS DENTRO DE LAS U. P. M.

27. -La selección de fincas pequeñas no se hizo directamente del listado de fincas de cada U. P. M. visto que no disponíamos de este listado y saldría muy caro elaborarlo en el campo. Se prefirió concentrar un poco más la muestra, introduciendo una etapa adicional de muestreo que consistió en seleccionar previamente segmentos censales dentro de cada U. P. M. Dentro de los segmentos seleccionados se obtuvo entonces la muestra de fincas. El proceso adoptado fue el mismo para U. P. M., auto-representadas y no auto-representadas y se describe a continuación usando como ejemplo el distrito 11205 seleccionado en el cuadro IV.

Cuadro V. - SELECCION DE SEGMENTOS CENSALES EN EL DISTRITO 11205 DE ESTRATO I DE LA REGION AGRICOLA C. G.

Código del Segmento	N° Viviendas de Segmento	N° de Unidades Ultimas de Muestreo			N° de Selección	Fracción de Muestreo dentro del Segm.
		Aproximadas	Usadas	Acumuladas		
001	36	3.7	4	4	01.5087	1/4
002	39	4.1	4	8		
003	40	4.2	4	12	11.7015	1/4
004	42	4.4	4	16		
005	48	5.1	5	21		
006-007	42	4.4	5	26		
TOTAL	247	25.9	26	-	-	-

N° de fincas de la U. P. M. =185; Tamaño de la Unidad Ultima de Muestreo (U. U. M.)=7 fincas; N° de U. U. M. a considerar en la U. P. M. = 185/7=26; N° de viviendas de la U. P. M. = 247; N° de viviendas correspondientes a cada U. U. M. = 247/26=9.5. Interval de muestreo para seleccionar segmento: I=(30) P= (30)X(0.33976)=10,11928; N° de arranque al azar = 01.5087.

28. -Como se dijo antes, no se dispone de un listado con el número de fincas por segmento censal, pero, con facilidad se podía obtener de los mapas censales un listado con el número de viviendas por segmento, en cada U. P. M. seleccionada. Para algunos distritos se disponía de mapas recientes con el número de viviendas actualizado en los últimos tres años, en la preparación del marco de una encuesta de hogares. Asumiendo que por lo menos para los segmentos netamente rurales, la correlación entre el número de viviendas y el número de fincas pequeñas es relativamente elevada (casi solamente los productores grandes viven en las ciudades), se decidió seleccionar los segmentos con probabilidad proporcional al número de viviendas. Los varios pasos de la selección se indican en el cuadro V.

29. -Se comenzó por establecer que si el número de fincas por segmento no hubiera cambiado desde el censo de 1963 se irían a seleccionar 7 fincas por segmento. Este número se tomó como constituyendo una unidad Ultima de Muestreo (U. U. M.) y con ello se calculó el número de U. U. M. contenidas en cada U. P. M. seleccionada. Finalmente, conociendo el número de U. U. M. se determinó el número promedio de viviendas correspondientes a cada una. El número de U. U. M. a formar por segmento se obtuvo dividiendo el número de viviendas contenidas en cada uno entre aquel promedio; el cociente se muestra en la columna 3 del cuadro V y en la columna 4 se dan los números enteros correspondientes a las aproximaciones de la columna 3. En la columna 5 se hizo el acumulado sucesivo de las U. U. M. de la columna 4.

30. -Para determinar la fracción de muestreo a aplicar en la selección de segmentos, hubo que tomar en consideración que se pretendía tener al final una fracción total de muestreo para fincas pequeñas igual a 1/30 y que no se deseaba una concentración de la muestra superior a la dada por la selección de una

U. U. M., por segmento. Estas condiciones eran fácilmente llenadas haciendo f = 1/30P como se puede verificar por la siguiente identidad:

$$PiX(Uij\ XfFi)\ X\ 1/Uij = Pi\ X(Uij/30Pi)\ X\ 1/Uij = 1/30.$$

en donde:

Fi = fracción de muestreo a adoptar en la i-ésima U. P. M.

Pi = probabilidad de selección de la i-ésima U. P. M. en el estrato.

Uij = número de U. U. M. correspondiente al segmento i-ésimo de la i-ésima U. P. M.

31. -Una vez conocida la fracción de muestreo Fi, la selección e identificación de los segmentos de la muestra se hizo por proceso idéntico al empleado en la selección e identificación de las U. P. M. dentro de los estratos y por lo tanto evitamos de repetirlo aquí.

32. -Es evidente que dados los cambios ocurridos desde el censo de 1963 solamente en casos muy raros se seleccionarán las 7 fincas previstas por segmento. Sin embargo, dado que se conoce la fracción de muestreo 1/Uij a adoptar en cada segmento y que el proceso de selección a usar el aleatorio-sistemático, será fácil al supervisar idéntificar la muestra actual de fincas en el listado del segmento previamente elaborado en el campo por el empadronador. Para eso el supervisor seleccionará con anterioridad para cada segmento un número de arranque al azar entre 1 y Uij lo cual identificará la primera finca de la muestra, las restantes serán todas aquellas que comenzando en el número de arranque se encuentran en el listado a intervalos iguales a Uij. Con el propósito de evitar la concentración de las fincas de la muestra el empadronador recibe instrucciones específicas de como hacer el recorrido del segmento en el momento de elaborar el listado.

IV - METODOS DE ESTIMACION

AJUSTES PARA NO RESPUESTAS

33. -Aunque al momento de diseñar una muestra se opta usualmente por el uso de fórmulas simples para la expansión de los datos, sólo en casos muy raros eso se consigue sin previos ajustes de la información recogida en el campo.

Boletas perdidas, boletas inutilizadas por errores de llenado, no-entrevistas por motivos diversos, etc.; requieren ajustes de los datos antes de que se puedan usar las fómrulas simples previstas inicialmente. Por eso, los métodos propuestos para proceder a sus ajustes deben siempre hacer parte del diseño de la muestra.

34. -Los métodos a adoptar, en la presente encuesta, para el ajuste de información incompleta proveniente de varias causas para que se puedan englobar con el título general de "no-respuesta", se dan a continuación, separadamente para cada uno de los tres tamaños distintos de fincas.

(i). - FINCAS GRANDES:- Se elaborará una boleta sumaria por distrito conteniendo para cada renglón el total correspondiente a las fincas empadronadas. Las boletas correspondientes a no-respuestas se llenarán entonces con los promedios por finca, obtenidos de la boleta sumaria. Cuando el número de boletas usables por distrito es inferior a 6 se juntarán las sumarias de dos o más distritos con características agrícolas parecidas hasta completar por lo menos aquel número y las no-respuestas se calcularán como promedios de los valores de las boletas sumarias del respectivo grupo de distrito.

(ii). - FINCAS MEDIAS:- El proceso será idéntico al descrito para fincas grandes, sólo que en este caso se prevé que el número de fincas medias por distrito será suficiente para proceder a todos los ajustes con la boleta sumaria elaborada para cada distrito. Será conveniente notar que tanto en el caso de fincas medias como en el caso de fincas pequeñas, las boletas sumarias una vez corregidas para no-respuestas constituyen un medio simple y apropiado para expandir los datos de la muestra y por lo tanto el trabajo usado en su elaboración no representa una pérdida de tiempo.

(iii). - FINCAS PEQUEÑAS:- Se elaborarán boletas sumarias para cada U. U. M. y los valores correspondientes a no-respuestas se obtendrán como promedios por fincas de esas sumarias. Tal como en el caso de fincas grandes si el número de boletas completas por U. U. M. es inferior a 6 los promedios se calcularán agrupando las U. U. M. de la misma U. P. M.

35. -Es obvio que los ajustes de datos que se puedan llevar a cabo en la oficina de modo alguno sustituyen los datos verdaderos provenientes de entrevistas bien realizadas. Por eso, en encuestas del tipo de entrevista directa al productor, se debe exigir como norma al empadronador hacer el máximo esfuerzo en el campo de modo que el número de no respuestas no sea nunca superior a 5%.

(b). - ESTIMACION DE TOTALES

36. -Aunque en el análisis final de los datos de la presente encuesta se irán a obtener estimaciones totales, promedios, porcentajes, etc., en este párrafo solamente nos vamos a referir a las estimaciones de totales. La razón es doble: - (i). Para evitar de alargar demasiado este escrito, y (ii) porque una gran parte de las restantes estimaciones necesitan el conocimiento previo de totales.

37. -Para cualquiera característica en donde se requiere la estimación total general el respectivo estimador es dado por la siguiente expresión:

$$\hat{Y}' = Y'_G + \hat{Y}'_M + \hat{Y}'_P \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

en donde: $\hat{Y}'$ = estimación del total general.

Y'_G = total correspondiente a fincas grandes.

$\hat{Y}'_M$ = estimación del total de la característica para fincas medias.

$\hat{Y}'_P$ = estimaciones del total de la característica correspondiente a fincas pequeñas.

38. -El total Y'_G correspondiente a fincas grandes aparte los ajustes para no-respuestas no representa realmente una estimación, visto que todas las fincas grandes fueron empadronadas, en cuanto a la estimación del total correspondiente a fincas medias, se compone de dos parcelas distintas:

$$\hat{Y}_M = Y'_{MDAR} + \hat{Y}'_{MDNAR} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

en donde:

Y'_{MDAR} = total de la característica correspondiente a fincas medias empadronadas en U. P. M. auto-representadas.

$\hat{Y}'_{MDNAR}$ = estimación del total para fincas medias localizadas en U. P. M. no auto-representadas.

39. -Dado que el factor de expansión para los totales de las U. P. M. auto representadas es igual a I, el total Y_{MDAR} es simplemente igual a la suma de los valores correspondientes a las fincas medias empadronadas. Por el contrario las U. P. M. no auto-representadas tienen distintos probabilidades de selección y por eso el estimador total Y_{MDNAR} es dado por:

$$\vec{y}'_{MDNAR} = \sum_k \sum_i g_{ki} \cdot y_{ki} = \sum_k \vec{y}'_k \quad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

en donde:

$$g_{ki} = 1/p_{ki}$$

p_{ki} = probabilidad de selección de la i-ésima U. P. M. en el estrato K.

$$y_{ki} = \sum_j y_{kij}$$

y_{kij} = valor de la característica para la finca i-ésima de la i-ésima U. P. M. del estrato k.

41. -También el total, $\hat{Y}'_p$ de la fórmula (1) consta de dos parcelas distintas, una correspondiente a las fincas pequeñas empadronadas en U. P. M. auto-representadas y que fueron seleccionadas en dos etapas de muestreo y la otra a fincas pequeñas localizadas en U. P. M. no auto-representadas a las cuales se aplicó para su identificación tres etapas de muestreo. Sin embargo en los dos casos la fracción total de muestreo para fincas fue la misma y igual a 1/30. En estas condiciones el estimador del total de 7, para este estrato del universo, es dado por:

$$\vec{y}'_p = 30. \left(\sum_h \sum_i \sum_j y_{hij} + \sum_k \sum_i \sum_j \sum_\ell y_{kij\ell} \right) =$$

$$= \sum_h \hat{\vec{y}}'_h + \sum_k \hat{\vec{y}}'_k \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

en donde:

y_{hij} = valor de la característica para la finca i-ésima empadronada en el segmento i-ésimo de la U. P. M. auto-representada h/

$$\hat{y}'_h = 30. \sum_i \sum_j y_{hij}$$

$y_{kij\ell}$ = valor de la característica para la finca s-ésima empadronada en el segmento i-ésimo, perteneciente a la U. P. M. no auto-representada i-ésima, del estrato k.

$$\hat{y}'_k = 30. \sum_i \sum_j \sum_\ell y_{kij\ell}$$

(c). - Estimadores de las variancias de los totales.

42. -Los totales correspondientes a fincas grandes y a fincas medias empadronadas en U. P. M. auto-representadas no están afectados por errores de muestreo dado que en los dos casos las fracciones de muestreo son iguales a 1. Tendremos que considerar, por lo tanto, solamente las variancias correspondientes a los estimadores (3) y (4) dadas en el párrafo anterior.

(i). - Fincas medias en U. P. M. no auto-representadas:- En el estrato K, el estimador del total es, de acuerdo con (3) dado por:

$$\hat{y}'_k = \sum_{i=1}^{2} g_{ki}\, y_{ki} = \sum_{\substack{i=4 \\ 1}}^{2} g_{ki} \sum_{j=1}^{N_{ki}} g_{kij} \quad \ldots\ldots\ldots\ldots (5)$$

La variancia teórica de (5) es:-

$$V(\hat{y}'_k) = \frac{1}{m_k} \sum_1^{M_k} p_{ki} \left(\frac{y_{ki}}{p_{ki}} - y_k \right)^2 \quad \ldots\ldots\ldots\ldots (6)$$

en donde:

M_k = número total de U. P. M. en el estrato k

$m_k = 2$ número de U. P. M. seleccionadas en el estrato k el cual en la presente encuesta fue siempre igual a 2.

$$y_k = \sum_i^{M_k} \sum_j^{N_{ki}} y_{kij}$$

N_{ki} = número de fincas medias en la i-ésima U. P. M. de estrato k.

El estimador de (6) obtenido de la muestra es simplemente:

$$v(\hat{y}'_k) = (\hat{y}'_{k1} - \hat{y}'_{k2})^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots (7)$$

el cual sumado para todos los estratos da para estimación de la variancia de (3):-

$$v\,(\hat{\bar{Y}}'_{MDNAR}) = \sum_{K} (\hat{\bar{Y}}'_{K1} - \hat{\bar{Y}}_{K2})^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots (8)$$

en donde:

$$\hat{\bar{Y}}'_{Ki} = g_{Ki} \sum_{j}^{N_{Ki}} y_{Kij} \quad (i = 1,2)$$

(ii). - Fincas pequeñas en U. P. M. auto-representadas:- Cada una de las U. P. M. auto-representadas funcionan como si fuera un estrato en donde las fincas seleccionadas son identificadas en dos etapas su cesivas con probabilidad total igual a 1/30. El estimador del total de 7 para cada una de las U. P. M. es, de acuerdo con (4) dado por:

$$\hat{\bar{Y}}'_h = 30. \sum_{i} \sum_{j} y_{hij} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (9)$$

La variancia de muestreo de $\hat{\bar{Y}}'_h$ es:

$$V(\hat{\bar{Y}}'_h) = (30)^2 \frac{N_h - n_h}{N_h} \cdot \frac{S_h^2}{n_h} \quad \ldots\ldots\ldots\ldots\ldots (10)$$

en donde:

N_h = número total de U. U. M. en cada una de las U. P. M. auto-representadas

n_h = número de U. U. M. seleccionadas en cada U. P. M.

$$S_h^2 = \sum_{i=1}^{N_h} (y_{hi} - \bar{y}_h)^2 / (N_h - 1)$$

$$y_{hi} = \sum_{j=1}^{Q_{hi}} y_{hij}$$

$$\bar{y}_h = \sum_{i} y_{hi} / N_h$$

Es de notar que la fórmula simplificada del estimador (10) se debe al hecho de que el proceso de selección adoptado es equivalente a un muestreo simples al azar de U. U. M., aunque la identificación de las fincas de la muestra se haga en dos etapas.

La estimación inesgada de la variancia de $\bar{Y}_h$ a partir de los datos de la muestra se obtiene por:-

$$v(\hat{\bar{Y}}'_h) = (30)^2 \frac{N_h - n_h}{N_h} \cdot \frac{s_h^2}{n_h} \quad \ldots\ldots\ldots\ldots\ldots\ldots (11)$$

en donde:

$$s_h^2 = \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2 / (n_h - 1)$$

es una estimación inesgada de S_h^2.

dado que $\frac{N_h - n_h}{N_h} = 1$ la formula (11) se puede todavía simplificar más para de obtener:-

$$\mathcal{V}(\hat{Y}_h') = (30)^2 S_h^2 / n_h \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (12)$$

Sumando (12) para todas las U. P. M. auto-representadas viene finalmente:

$$\mathcal{V}(\hat{Y}'_{PDAR}) = (30)^2 \sum_h (S_h^2 / n_h) \quad \ldots\ldots\ldots\ldots (13)$$

(iii). - Fincas pequeñas en U. P. M. no auto-representadas.

Las fincas pequeñas de U. P. M. no auto-representadas fueron seleccionadas en tres etapas de muestreo empleando un proceso de selección que dá a cada una de las fincas de la muestra la misma probabilidad total, igual a 1/30. Por eso el estimador del total de cualquiera característica, para un estrato es dado por:

$$\hat{Y}_K' = 30. \sum_i \sum_j \sum_s y_{ijs}$$

en donde las letras tienen el mismo significado que en la formula (4).

La formula del estimador de la variancia de $\hat{Y}_K'$ es bastante compleja y por eso evitamos de darla aquí (para los interesados se recomienda ver las páginas 396-399 de Sample Survey Methods and Theory, vol. I, de Morris H. Hansen, William N. Hurwitz, and William G' Madoro).

Sin embargo la estimación de la variancia a partir de la muestra se obtiene por una formula bastante simple, si no se está interesado en conocer cada uno de los tres componentes de la variancia total. Asi:-

$$\mathcal{V}(\hat{Y}_K') = (30)^2 (Y_{K1} - Y_{K2})^2 \quad \ldots\ldots\ldots\ldots\ldots (14)$$

en donde: $Y_{Ki} = \sum_j \sum_s y_{kijs} \qquad (i = 1,2)$

es el total de la característica estudiada para las U. P. M., 1 y 2 seleccionadas en cada estrato.

Sumando sobre todos los estratos tenemos finalmente.

$$\mathcal{V}(\hat{Y}'_{PDNAR}) = (30)^2 \sum_K (Y_{K1} - Y_{K2})^2 \quad \ldots\ldots\ldots\ldots (15)$$

(iv). - Variancia del total y coeficiente de variación:

Dado que el total general de cualquiera característica, Y' dado por (1) también se puede escribir.

$$\hat{\bar{y}}' = \bar{y}_G' + \bar{y}_{MDAR}' + \hat{\bar{y}}_{MDNAR}' + \hat{\bar{y}}_{PDAR}' + \bar{y}_{PDNAR} \quad \ldots\ldots (16)$$

y que las cinco parcelas son independientes en el sentido estadístico, la variancia de Y' es dada por:

$$v(\hat{\bar{y}}') = v(\hat{\bar{y}}_{MDNAR}') + v(\hat{\bar{y}}_{PDAR}') + v(\bar{y}_{PDNAR}') \quad (17)$$

visto que las variaciones de muestreo de Y_G y Y_{Mar} son iguales a cero. Agrupando, por lo tanto los valores obtenidos independientemente por medio del uso de las fórmulas, (8), (13) y (15) llegaremos a la estimación de la variancia total de Y. El coeficiente de variación será dado por:

C. V. (%) = V (Y) / (Y) x 100.....(17)

NOTA SOBRE LA DETERMINACION APROXIMADA DEL LIMITE INFERIOR DE TAMAÑO PARA IDENTIFICAR FINCAS GRANDES EN LA SELECCION DE MUESTRAS

Por :

J. G. Baptista (FAO-GAFICA)

Para una muestra estratificada la forma de repartición de Neyman es:

$$f_k = \frac{n_k}{N_k} = \frac{n S_k}{N_k S_k} \qquad (1)$$

en donde:

f_k = fracción de muestreo en el estrato k-ésimo.

n_k = n° de fincas a seleccionar en el estrato k-ésimo.

N_k = n° de fincas del universo pertenencientes al estrato k-ésimo.

S_k = desviación standard de la característica en estudio dentro de estrato k-ésimo.

En la mayor parte de las poblaciones trabajadas en la práctica se verifica la relación,

$$S_k = c. \bar{Y}_k \qquad (2)$$

en donde:

c = constante.

$\bar{Y}_k$ = promedio de la característica en el estrato k-ésimo.

Sustituyendo (2) en (1) tenemos:

$$f_k = \frac{nc\bar{Y}_k}{c\,N_k Y_k} = n\frac{Y_k}{Y} \qquad (3)$$

Haciendo $f_k = 1$ como pretendemos para fincas grandes y resolviendo en relación a $\bar{Y}_k$ obtenemos:

$$\bar{Y}_k = \frac{Y}{n} \qquad (4)$$

La expresión (4) nos indica que de acuerdo con la repartición de Neyman (usando la aproximación (2) todas las fincas con un promedio de tamaño para la característica en estudio igual o superior al total

del universo dividido entre el tamaño de la muestra, deben entrar con la certeza en la muestra a estudiar para aumentar su eficiencia. En la práctica, usualmente se toma un valor un poco inferior a $\bar{Y}_k$ para compensar las aproximaciones usadas.

Es de notar que la expresión (4) se puede usar para cada característica y dominio de estudio, separadamente. En el caso de los censos en que los dominios de estudio son usualmente las provincias esto significa que podemos tener límites distintos para la misma característica en las diferentes provincias.

Una vez separadas las fincas grandes del universo se puede proceder a la estratificación de las restantes y usar la repartición de Neyman para obtener una distribución óptima de las restantes fincas de la muestra.

Se debe entender que la expresión (4) es solamente una aproximación a la realidad pero que da buenos resultados en la práctica. Siempre que los costos en los diferentes estratos no sean muy desiguales. La solución exacta de la partición óptima del universo en estratos es mucho más compleja de lo que muestra la simple fórmula (4). (véase W. G. Cochran - Sampling Tecnniques, 2nd. edition, páginas 128-135).

Ejemplo:

Supongamos que usando una muestra de aproximadamente 4.000 fincas queremos estimar el total de ganado bovino en Costa Rica que se cree sea aproximadamente de 1.500.000 cabezas.

Usando (4) tenemos

$$\bar{Y}_k = \frac{1.500.000}{4.000} = 375 \text{ cabezas.}$$

En vez de tomar como límite inferior 375 se ha usado el límite 350 para separar las fincas grandes. Esto nos da un número de fincas grandes alrededor de 500, quedando cerca de 3.5000 para ser empadronadas en áreas de muestreo. El efecto ha sido hacer aumentar bastante la precisión de las estimaciones obtenidas como muestra el siguiente razonamiento.

(a) El coeficiente de variación del universo original es aproximadamente de 4.5.

(b) Una vez separadas las fincas grandes el coeficiente de variación pasa a ser aproximadamente 2.3.

(c) Las fincas grandes representan aproximadamente 1% del total del universo pero contienen cerca de 30% del total del ganado, es decir, aproximadamente 450.000 cabezas.

(d) Si no se separan las fincas grandes una muestra de 4.000 fincas daría un coeficiente de variación de la estimación del total igual a:

$$V_{\bar{y}} = \frac{V}{\sqrt{n}} = \frac{4.5 \times 100}{\sqrt{4.000}} \doteq \frac{450}{63} \doteq 7.1\%$$

(e) Al separar las fincas grandes y empadronarlas todas se obtiene un coeficiente de variación para la estimación del universo de fincas pequeñas igual a:

$$V_{\bar{y}_p} = \frac{V'}{\sqrt{n}} = \frac{2.30 \times 100}{\sqrt{3.500}} = \frac{230}{59} = 4.0\%$$

(f) El coeficiente de variación 4% corresponde a un error absoluto igual a:

$$(0.04)\ (1.500,000 - 450.000) = 42.000 \text{ cabezas.}$$

(g) Una vez que los datos de las fincas grandes no están afectadas de error de muestreo, el error relativo de la estimación total para el universo es:

$$\text{error relativo (\%)} = \frac{42.000}{1.500.000} \times 100 = 2.9\%$$

(h) Dadas las características de la muestra en la cual las fincas pequeñas no son seleccionadas por muestreo simples al azar si no que en conglomerados de aproximadamente 10 fincas el coeficiente de variación de la estimación del total de ganado no es tan bajo como 2.8% pero se localiza alrededor de 4.5%.

Impreso en la Sección de PUBLICACIONES - San José, Costa Rica - Fórm. 764 - 1600