Encuesta de Calidad de Vida 1997, DANE

DISEÑO E IMPLEMENTACION DE LA MUESTRA PROBABILISTICA

Luis Carlos Gómez
Asesor Encuestas por Muestreo
Febrero de 1998

La Encuesta de Calidad de Vida 1997, se llevó a cabo en una submuestra de la nueva Muestra Maestra de Población de Propósitos Múltiples del DANE, utilizada por primera vez en 1994 para la Evaluación de la Cobertura del Censo de Población de 1993, e incorporada a partir de 1996 al Sistema de Encuestas Nacionales de Hogares.

1. La Muestra Maestra

Tiene un tamaño total de 70.000 hogares, distribuidos en 219 Unidades Primarias de Muestreo (básicamente municipios), y 3.500 segmentos, de 20 hogares en promedio con una composición urbano-rural similar a la de la población total. El Marco de Muestreo utilizado para la selección se basó en los materiales del Precenso 93, para los 60 principales centros urbanos del país, y en una combinación de materiales del Censo de 1985 y del Catastro Nacional, en las zonas rurales y centros urbanos menores. La representación de la muestra por variables geográficas y socioeconómicas se ajusta para cada encuesta en función de los resultados y proyecciones del Censo de Población 93.

El diseño muestral es probabilístico, de conglomerados, estratificado y polietápico. Toda la superficie del universo poblacional (que excluye solamente las zonas rurales dispersas de la Orinoquía y la Amazonía), tuvo probabilidad de selección. En consecuencia, las zonas despobladas de las áreas rurales y de los centros urbanos, fueron anexadas a una área vecina que tuviera alguna población, con el objeto de garantizar, dentro de un enfoque de muestreo de áreas, la captación dinámica de los cambios poblacionales que se fueran presentando en el futuro.

La selección probabilística de las diferentes unidades de muestreo (conglomerados) estuvo siempre precedida de algún tipo de estratificación, en función de variables altamente correlacionadas con la mayoría de los fenómenos sociales y económicos del país. Las unidades primarias (municipios) se estratificaron dentro de cada departamento, con base en el tamaño de las cabeceras municipales (nivel de urbanización), su composición urbano-rural, y el indicador de necesidades básicas insatisfechas (NBI); las unidades secundarias (secciones censales o manzanas),

según su localización urbana o rural, y cuando se disponía de información, según el estrato socioeconómico en que estaban clasificadas.

Para la selección de las unidades primarias se aplicó la técnica de "Selección Controlada" (ver Goodman y Kish, 1950), que permite, dentro de un esquema totalmente probabilístico, una óptima composición de la muestra según las variables de estratificación, con el fin de garantizar la máxima reducción posible en el error de muestreo de cualquier tipo de estimación, que es el objetivo genérico perseguido por los procesos de estratificación.

Para la precisa delimitación y partición de las áreas últimas de muestreo de las zonas urbanas no incluidas en el precenso, y de las zonas rurales, se llevó a cabo la denominada "Segmentación". Este proceso, consistió en las visita de las citadas áreas últimas seleccionadas a nivel central, con el objeto de recorrerlas minuciosamente, ubicar las viviendas existentes, e identificar con precisión sus límites externos, y todas las particularidades físicas internas que pudieran ser utilizadas para su partición en segmentos múltiples, de acuerdo con su tamaño total. Igualmente se levantó en este proceso toda la información necesaria para la fácil localización y recorrido de los segmentos por parte de los futuros encuestadores.

2. La Submuestra para la Encuesta de Calidad de Vida

Por diseño, está constituida por 10.000 hogares, concentrados en 1.000 segmentos, y en 75 unidades primarias (municipios) ; la distribución urbano-rural de la muestra se ha planteado en partes ligeramente diferentes, 585 segmentos en cabeceras y 415 en el resto; el menor tamaño en la zona rural está dado en función de la menor heterogeneidad de esta. De esta forma se garantizan a ambas zonas las mismas posibilidades analíticas y de desagregación de resultados, con aproximadamente el mismo nivel de precisión. Por definición, este enfoque genera desiguales probabilidades se selección, que obligan el uso de ponderaciones en los resultados, equivalentes al recíproco de las probabilidades, o a factores equivalentes. La tasa global esperada de no respuesta, estimada inicialmente en un máximo del 20%, teniendo en cuenta experiencias recientes, permitió pronosticar un mínimo de 8.000 hogares efectivamente encuestados, por lo cual se diseñaron estrategias especiales de recolección para reducir esta tasa. La muestra finalmente encuestada fue de---- hogares.

La selección de la submuestra descrita, se hizo también, por supuesto, con un diseño probabilístico que dividió la muestra total de unidades primarias en tres submuestras equivalentes, todas ellas representativas a nivel nacional, de grandes regiones, de algunas subregiones, y de los mayores centros urbanos, los mayores municipios del país se repiten en cada submuestra, pero con diferentes segmentos. En la clasificación de las tres submuestras se aplicaron los mismos criterios de estratificación utilizados en el diseño total.

El tamaño, composición y distribución de la muestra, permite diferentes grados de desagregación geográfica y por otras variables independientes, dependiendo de la frecuencia de los fenómenos estudiados, la precisión deseada en las estimaciones (error estándar relativo, Esrel), y el efecto esperado de la conglomeración deff (municipios, segmentos, hogares).

En un ejercicio de cálculo hecho cuando se estaban gestando los parámetros de esta investigación (ver tabla 1), se estimaron las diferentes alternativas de desagregación para tres tamaños, 8.000, 10.000 y 12.000 hogares. En todas ellas se supuso la siguiente desagregación geográfica básica, rutinaria:

- Las grandes regiones geográficas: Atlántica, Oriental, Pacífica, Central, Antioquia y Bogotá D.C.
- Partición urbano-rural de cada región (Bogotá no interesa).
- Dos subregiones especiales: Orinoquía-Amazonía (sólo localidades de más de 800 habitantes) y San Andrés y Providencia.
- Dos áreas metropolitanas especiales: Medellín y Cali.
- Seis niveles de urbanización.

Del análisis de la tabla 1 se pueden deducir conclusiones como la siguiente: con una muestra efectiva de 8.000 hogares, un Esrel de 10% y un deff de 1.5 (los deff 2 y 2.5 son un tanto pesimistas), es factible analizar:

- Fenómenos con una probabilidad de ocurrencia del 5%, en 12 desagregaciones.
- Fenómenos con una probabilidad de ocurrencia del 10%, en 26 desagregaciones.
- Fenómenos con una probabilidad de ocurrencia del 20%, en 60 desagregaciones.
- Fenómenos con una probabilidad de ocurrencia del 30%, en 102 desagregaciones.

La tabla 2 ilustra la distribución de los segmentos de la muestra seleccionada por regiones y dominios de análisis y presentación de resultados.

3. Procedimientos de Estimación de los Resultados

Se refieren a la metodología estadística para garantizar estimaciones insesgadas para el universo estudiado. Los sesgos pueden originarse en las desiguales probabilidades finales de selección de las unidades de estudio, los problemas de medición de las variables de estudio, la cobertura diferencial del operativo de encuesta, y por errores en el procesamiento y edición de los resultados de la investigación. Al mismo tiempo, las estimaciones de los resultados pueden mejorarse sustancialmente en su validez y confiabilidad, mediante el uso de variables exógenas que mejoren la estructura y magnitud de las variables independientes.

Se describen enseguida tres esquemas esenciales de corrección de los valores muestrales.

3.1. Factor Básico de Expansión

Se denomina así el recíproco de la probabilidad final de selección de las unidades últimas de observación. Este factor aplicado rutinariamente a los datos muestrales corrige el sesgo que se generaría en el uso no ponderado de la información recolectada.

Cada una de las cuatro fases de selección de las unidades de la muestra se cumplió aleatoriamente con una determinada probabilidad. Tales fases fueron: selección de unidades primarias de muestreo (municipios) de la muestra maestra; submuestreo de unidades primarias para la encuesta de calidad de vida; selección de unidades secundarias (segmentos) en las unidades primarias de la muestra maestra; y selección de la submuestra de segmentos en las unidades primarias de la submuestra de calidad de vida. Dado que las submuestras de unidades primarias y secundarias se realizaron con el mismo procedimiento y esquema probabilístico, el cálculo de las probabilidades finales se simplificó en dos pasos.

a) Probabilidad de Selección de las Unidades Primarias de Muestreo (UPM)

$$P_1 = \frac{\text{Población tal de la UPM}}{\text{Población total del Superestrato al cual pertenece la UPM}}.$$

b) Probabilidad de Selección de las Unidades Secundarias de Muestreo (USM) o segmentos

$$P_2 = \frac{\text{No de Segmentos Seleccionados en una zona (cabecera o resto) de la UPM}}{\text{No. de Segmentos Existentes (teóricos) en la misma zona de la misma UPM}}$$

c) Probabilidad Final

$$P_1 \times P_2 = P$$

d) Factor Básico de Expansión

$$F_b = \frac{1}{P}$$

3.2. Ajuste por No Cobertura de la Muestra

Dado que la tasa de no cobertura de la muestra varía en las diferentes unidades primarias y los distintos subgrupos (socioeconómicos) de la población, es conveniente corregir el Factor Básico de Expansión por factores que corrijan la cobertura diferencial de segmentos, viviendas y hogares de la muestra, así:

a) Factor de Ajuste de la No Cobertura de Segmentos (completos)

Ac_1 = No. de segmentos seleccionados en una zona de una región o subregión
No. de segmentos encuestados en la misma zona y región

b) Factor de Ajuste de la No Cobertura de Viviendas

Ac_2 = No. de viviendas seleccionadas en un determinado segmento (en los segmentos encuestados)
No. de viviendas encuestadas en el mismo segmento

c) Factor de Ajuste de la No Cobertura de Hogares

Ac_3 = No. de hogares seleccionados en un determinado segmento (en las viviendas encuestadas
No. de hogares encuestados en el mismo segmento

d) Factor Básico de Expansión Ajustado por Cobertura

$F_{bc} = F_b \times A_{c1} \times A_{c2} \times A_{c3}$

3.3. Ajuste por las Proyecciones de Población a la Fecha de la Encuesta

Este factor aplicado a nivel de región o subregión, por zona, corrige la estructura de la población expandida a partir de la muestra con base en el Factor Básico ajustado por cobertura, e iguala la población total expandida a la proyectada con base en los datos del último censo.

La hipótesis es que la estructura urbano-rural originada en el censo es más precisa que la originada en la muestra.

a) Factor de Ajuste

A_p = Población proyectada a la fecha de la encuesta en una zona de una región o subregión
Población expandida a partir de la muestra con base en los factores ajustados por cobertura en la misma zona y región

b) Factor Final de Expansión

$F_{bcp} = F_{bc} \times A_p$

4. Cálculo de la Precisión Observada: Errores de Muestreo

El error estándar, que es el indicador de la precisión de los resultados estimados, refleja la variabilidad del azar, propia de las muestras probabilísticas.

Los errores estándar de los resultados de la Encuesta de Calidad de Vida 1997, pueden ser calculados con fórmulas desarrolladas para el Muestreo Estratificado de Conglomerados Desiguales aquí utilizado. Un modelo apropiado es el de "Propagación de Varianzas".

Las tasas, razones, proporciones y promedios, generadas a partir de este diseño muestral son de la forma de una razón (r), en la cual el numerador y el denominador son variables aleatorias, así:

$$r = \frac{y}{x} = \frac{\sum_{h}^{H}\sum_{\alpha}^{ah} y_{h\alpha} F_{fh\alpha}}{\sum_{h}^{H}\sum_{a}^{ah} x_{h\alpha} F_{fh\alpha}}$$

$h = 1,2.....,H$, son cada uno de los estratos.

$\alpha = 1,2.....,a_h$, son cada uno de los segmentos del estrato h.

$x =$ Es el numero de casos del universo (expandido)

y = Es, en proporciones, el número de casos expandidos que tienen la característica estudiada, y en promedios y, la suma de los valores de la variable estimada.

$F_{fh\alpha} = \frac{1}{P_{h\alpha}}$ es el factor de expansion de cada segmento α en el estrato h

$P_{h\alpha}$ = fraccion de muestreo o probabilidad de seleccion del segmento α en el estrato h

El error estándar de una razón, es :

$$ES(r) = \sqrt{\frac{1}{(x_{h\alpha}F_{fh\alpha})^2}[\sum_h \text{var}(y_h) + r^2 \sum_h \text{var}(x_h) - 2r\sum_h \text{cov}(y_h, x_h)]}$$

En donde:

$$\text{var}(y_h) = \text{varianza de } y_h = \frac{1-p_{h\alpha}}{a_h - 1}[a_h \sum_{\alpha}^{a_h}(y_{h\alpha}F_{fh\alpha})^2 - (\sum_h^H \sum_{\alpha}^{a_h} y_{h\alpha}F_{fh\alpha})^2)]$$

$$\text{var}(x_h) = \text{varianza de } x_h = \frac{1-p_{h\alpha}}{a_h - 1}[a_h \sum_{\alpha}^{a_h}(x_{h\alpha}F_{fh\alpha})^2 - (\sum_h^H \sum_{\alpha}^{a_h} x_{h\alpha}F_{fh\alpha})^2)]$$

$$\text{cov}(y_h, x_h) = \text{covarianza de } y_h, x_h = \frac{1-p_{h\alpha}}{a_h - 1}[a_h \sum_{\alpha}^{a_h}(y_{h\alpha}x_{h\alpha}F_{fh\alpha}) - \sum_h^H \sum_{\alpha}^{a_h} x_{h\alpha}y_{h\alpha}F_{fh\alpha}]$$

$p_{h\alpha}$ = la fraccion de muestreo, o probabilidad de seleccion del segmento α

a_h = numero de conglomerados seleccionados en el estrato h

Tabla 1

TAMAÑOS ALTERNATIVOS DE MUESTRA Y NUMERO POSIBLE DE SUBGRUPOS DE ANALISIS EN:

- Tres alternativas de tamaño muestral (n)
- Dos niveles de precisión (Esrel)
- Tres posibles efectos de los conglomerados (Deff)
- Fenómenos de cuatro magnitudes diferentes (%)

Fenómenos %	n= 8.000						n = 10.000						n = 12.000					
	Esrel = 5% Deff			Esrel = 10% Deff			Esrel = 5% Deff			Esrel = 10% Deff			Esrel = 5% Deff			Esrel = 10% Deff		
	1.5	2.0	2.5	1.5	2.0	2.5	1.5	2.0	2.5	1.5	2.0	2.5	1.5	2.0	2.5	1.5	2.0	2.5
5%	3	2	1-2	12	9	7	4	3	2	15	12	9	4-5	3-4	2-3	19	14	11
10%	6	5	4	26	20	16	8	6	5	33	25	20	10	7	6	40	30	24
20%	15	11	9	60	45	36	18	14	11	75	56	45	22	17	13	90	67	54
30%	26	19	15	102	76	61	32	24	19	128	96	77	38	29	23	154	115	92

Tabla 2

COMPOSICION DE LA MUESTRA DE SEGMENTOS SELECCIONADOS POR REGIONES Y DOMINIOS

Regiones	Dominios	Segmentos
Antioquia	Medellín	50
	Resto U	35
	Rural	85
Pacífica	Cali	50
	Resto U	35
	Rural	85
Central	U	85
	R	85
Oriental	U	85
	R	85
Atlántica	U	85
	R	85
Bogotá	U	90
Orinoquía-Amazonía	U	40
San Andrés	U	30
	R	10
TOTAL	U	585
	R	415
	T	1000