Departamento Administrativo Nacional de Estadística

Dirección de Metodología y Producción Estadística-DIMPE

Metodología de Diseño Muestral
Encuesta Anual de Servicios –EAS

Julio 2007

CONTENIDO

1. INTRODUCCIÓN

Con la intención de dar a conocer la Metodología Estadística de la Encuesta Anual de Servicios, se elaboró este documento en el cual se hace referencia a los objetivos de estudio, el estado de la información, así como a la teoría de muestreo que permitió la escogencia del mejor diseño según los condiciones actuales y los propósitos de la encuesta.

2. OBJETIVOS

2.1.1. OBJETIVOS GENERALES

Conocer la estructura y comportamiento económico del sector de los servicios en estudio a través de valores absolutos.

2.1.2. OBJETIVOS ESPECIFICOS

Obtener la información necesaria para la estimación de los valores absolutos de los principales agregados económicos ingresos, gastos, personal ocupado, remuneraciones y movimiento de activos fijos y algunas regiones.

3. GRADO DE PRECISION

Medida en términos del error de muestreo es igual al 5% para ingresos y personal ocupado, para estimar Totales a Nivel Nacional..

3. COBERTURA

3.1. COBERTURA GEOGRÁFICA

La cobertura geográfica es nacional.

3.2. COBERTURA TEMATICA

Corresponde a las secciones determinadas en la CIIU Rev. 3 A.C. (Clasificación Industrial Internacional Uniforme. Revisión 3 adaptada para Colombia) del Sector Servicios que se presentan en la siguiente tabla:

Código Sección	Descripción Sección
E	Suministro de electricidad, gas y agua
H	Hoteles y restaurantes
I	Transporte, almacenamiento y comunicaciones
K	Actividades inmobiliarias, empresariales y de alquiler
M	Educación
N	Servicios sociales y de salud
O	Otras actividades de servicios comunitarios, sociales y personales

Se excluyen las empresas con 10 o menos personas ocupadas, personas naturales por falta de información de ubicación, no se tuvieron en cuenta las actividades de Actividades de Asociaciones NCP (91), Actividades Comerciales de las Casas de Empeño o Compraventas (5252), Educación no Formal (8060), Transporte Urbano Colectivo Regular de Pasajeros (6021) y las secciones Admón. Pública y Defensa; Seguridad Social de Afiliación Obligatoria (L), Clubes de fútbol (9271), Organizaciones y Órganos Extraterritoriales (Q) e Intermediación Financiera (J), ésta última porque la información se obtendrá por registros administrativos del DANE.

No se tuvieron en cuenta las empresas con menos de once personas ocupadas porque estas se consideran microestablecimientos, los cuales son numerosos, y con comportamientos dinámicos y dispersos, por lo que es difícil su actualización en los marcos de lista; para dichos establecimientos el DANE realiza la Encuesta Nacional de Micorestablecimientos de Comercio, Servicios e Industria.

4. PERIODO DE REFERENCIA

Año anterior al de recolección

5. DISEÑO ESTADISTICO

El diseño estadístico de la investigación, entendido como las técnicas que permiten la observación y análisis de la población objeto de estudio, requiere establecer el universo y lo que se quiere medir

en él para así definir la población objetivo, delimitar el marco de selección y generar propuestas de muestreo.

Antes de determinar la población objetivo y el universo de estudio, se realizó un diagnóstico de la información, en cuanto a la información disponible y su distribución.

- **Variables de estudio.**

Se selecciona la variable personal ocupado como variable de estudio porque tiende a ser estable en el tiempo y su magnitud en general se relaciona con la infraestructura, ventas y la producción de la empresa. También se selecciona la variable ingresos como principal variable de estudio debido a que está relacionada con el personal ocupado, sin embargo no siempre ingresos altos implican personal ocupado alto.

- **Información disponible en las variables de estudio.**

	Personal Ocupado	Ingresos
Sin Información	8,7%	8,7%
Con Cero	78,0%	55,7%
Con Información	13,3%	35,5%

- **Distribución de las variables de estudio.**

De las empresas que tienen información, pocas tienen ingresos y personal ocupado altos y muchas toman valores bajos en estas variables, por lo tanto, es necesario considerar un grupo o estrato de empresas grandes –estrato forzoso- y otro con las unidades restantes –estrato probabilístico-.

El diseño que se expone en este documento se evaluó en comité de expertos, teniendo en cuenta que la muestra permitiera alcanzar los objetivos y que fuera consistente con el presupuesto e información disponible, de esta manera la muestra permite dar resultados desagregados a nivel Sección en el estrato forzoso y para el estrato probabilístico por restricción presupuestal la muestra asegura estimaciones a Total Nacional sin desagregaciones.

5.1. UNIVERSO

Está conformado por empresas formalmente establecidas, residentes en el territorio nacional, cuya principal actividad es la prestación de servicios.

5.2. POBLACIÓN OBJETIVO

Está conformada por empresas formalmente establecidas, residentes en el territorio nacional, cuya principal actividad es la prestación de servicios conforme a la delimitación de la cobertura temática.

5.3. UNIDADES ESTADISTICAS

Unidad estadística: es la empresa con NIT que de manera exclusiva o predominante se dedica a las actividades de servicios dentro del territorio nacional.

Unidad de observación: es la empresa que realiza actividades de servicios dentro del territorio nacional, de la cual existe y puede recopilarse información.

Unidad de análisis: La empresa con NIT que realiza actividades de servicios, de la cual existe y puede recopilarse información.

Unidad de información: El contador, gerente, dueño o persona que maneja los balances de la empresa.

5.4. MARCO ESTADÍSTICO

El marco estadístico es el instrumento que permite la identificación y la ubicación de las unidades que conforman la población objetivo, de esta manera para *Producción de Servicios* el marco es de lista, y para su construcción se partió del Directorio de Servicios, su cobertura es nacional y se construyo con el Censo Económico del año 1990 y sus fuentes se actualizan con información de la Superintendencia de Sociedades de Vigilancia, de Cooperativas, Confecámaras, Gremios, Viceministerio de Turismo, Páginas Amarillas, las Encuestas Anuales del DANE, entre otros. El tamaño del marco de Producción de Servicios se presenta en la siguiente tabla:

Número de Empresas	Sin Información de Personal Ocupado (%)	Sin Información de Ingresos (%)
48.419	86,7	64,4

5.5. VARIABLES

Variables de clasificación

- Sector según clasificación CIIU Rev. 3 A.C.
- Tamaño de la empresa medido en función de Personal Ocupado e Ingresos

Variables de diseño

- Ingresos
- Personal Ocupado

5.6. PARAMETROS A ESTIMAR

Se estiman totales de las variables de estudio, agrupando según las variables de clasificación; se escoge esta medida porque permite conocer la estructura y comportamiento económico del sector de los servicios en valores absolutos.

5.7. DISEÑO MUESTRAL

El diseño muestral es estratificado de elementos –E.S.T. M.A.S.- y probabilístico porque las unidades de muestreo tienen probabilidad conocida y superior a cero de ser seleccionadas.

Estratificar implica particionar la población en conjuntos disjuntos de elementos cuya unión conforma el universo; en la estratificación se tiene en cuenta las diferencias que se presentan entre grupos poblacionales, su eficiencia se obtiene precisamente de considerar en forma separada las particularidades de cada grupo. La razón para estratificar radica en que los estratos presentan características tan diferentes que merecen la consideración en forma separada; la utilización de este método requiere la disponibilidad de información auxiliar que permita detectar diferencias y dividir la población en estratos, en este caso la información auxiliar corresponde por un lado a la Sección Económica, la cuál genera estratos temáticos, y por otra parte las variables de diseño generan estratos según el tamaños de la empresa, lo cuál reduce la varianza y permite contar con estimadores individuales para cada estrato.

De esta manera el primer criterio de estratificación es la sección económica y luego dentro de cada sección se subestratifica por el tamaño de la empresa medido por el valor que tomen las variables Personal Ocupado e Ingresos, creándose dos subestratos.

1. Un primer subestrato de inclusión forzosa con las empresas que se consideran grandes al interior de cada Sección.
2. Otro con el resto de empresas sobre el que se hace la selección de la muestra mediante un Muestreo Aleatorio Simple -M.A.S. –.

Las variables de estratificación se escogieron porque son un buen indicador del tamaño de la empresa y además son las de mayor interés según los objetivos; se estudiaron conjuntamente porque se complementan y para aprovechar al máximo la información disponible de cada una.

Los estratos conformados para esta investigación se presentan a continuación:

Descripción Sección	Inclusión	No.
Suministro de electricidad, gas y agua (Sección E)	Forzoso	1
	Probabilístico	2
Hoteles y restaurantes (Sección H)	Forzoso	3
	Probabilístico	4
Transporte, almacenamiento y comunicaciones (Sección I)	Forzoso	5
	Probabilístico	6
Actividades inmobiliarias, empresariales y de alquiler (Sección K)	Forzoso	7
	Probabilístico	8
Educación (Sección M)	Forzoso	9
	Probabilístico	10
Servicios sociales y de salud (Sección N)	Forzoso	11
	Probabilístico	12
Otras actividades de servicios comunitarios, sociales y personales (Sección O)	Forzoso	13
	Probabilístico	14

5.7.1. TAMAÑO DE MUESTRA

El tamaño de muestra para cada uno de los estratos está dado por el método de Hidiroglou, a continuación se describe el método.

Método de Hidiroglou [1]

Este método trabaja bajo el principio de que para un diseño E.S.T M.A.S. – Muestreo Aleatorio Simple Estratificado – la varianza del estimador tiene un comportamiento parabólico con un mínimo que se puede encontrar mediante iteraciones, este punto determina el límite de los estratos de inclusión forzosa y probabilística, para ello se aplica la siguiente expresión:

$$n(t) = N - \frac{(N-t)\,\mathrm{c}^2\,Y^2}{c^2\,Y^2 + (N-t)\,\mathrm{S}^2_{[\mathrm{N\text{-}t}]}}$$

donde,

n(t)=: tamaño de la muestra

t=: número de empresas de inclusión forzosa

N=: tamaño del estrato en número de empresas

C=: el coeficiente de variación deseado, en este caso 5 %

Y =: el total de la variable de diseño*

S^2 =: varianza de la variable de diseño *

*Esta información se obtuvo del directorio de Producción de servicios.

El algoritmo del método es de manera general el siguiente:

1. Se ordena de mayor a menor la variable de diseño.
2. Se toma la empresa más grande como forzosa y se calcula la varianza de las restantes.
3. Se toman las dos empresas más grandes como forzosas y se calcula la varianza de las restantes.
4. Este proceso se repite aumentando en cada paso el número de empresas forzosas hasta que la varianza sea mínima.
5. El valor de la variable de diseño en este punto se considera el límite entre los dos estratos, uno de empresas que se consideran grandes y otro con el resto, en general el primero se considera de inclusión forzosa y el segundo de inclusión probabilística.

[1] HIDIROGLOU, M.A.. The construction of a self-representing stratum of large units in survey design. 1986.

Aplicado el método descrito anteriormente se generaron los siguientes límites:

Sección	Descripción	Límite Personal Ocupado	Límite Ingresos (Pesos)
E	Suministro de electricidad, gas y agua	13	1.416.382.104
H	Hoteles y restaurantes	32	49.000.000
I	Transporte, almacenamiento y comunicaciones	46	450.000.000
K	Actividades inmobiliarias, empresariales y de alquiler	111	327.453.000
M	Educación superior	21	1.500.000
N	Servicios sociales y de salud	29	458.161.000
O	Otras actividades de servicios comunitarios, sociales y personales	32	220.000.000

El tamaño de muestra obtenido es de 13.000 empresas distribuidas de la siguiente manera:

Sección	Descripción	Número de Empresas		
		Marco	Forzosas	Probabilísticas
E	Suministro de electricidad, gas y agua	217	217	
H	Hoteles y restaurantes	2.452	633	312
I	Transporte, almacenamiento y comunicaciones	10.001	1.030	1.624
K	Actividades inmobiliarias, empresariales y de alquiler	23.467	1.985	3.840
M	Educación superior	276	276	
N	Servicios sociales y de salud	8.415	702	1.354
O	Otras actividades de servicios comunitarios, sociales y personales	3.591	431	596
TOTALES		48.419	5.274	7.726
			13.000	

El diseño de la muestra se realizó para obtener estimaciones, para totales nacionales, con un error de muestreo aproximado de 5%, para las variables ingresos y personal ocupado. Otro nivel de desagregación está sujeto a que su precisión no necesariamente sea buena.

5.7.2. MÉTODO DE SELECCIÓN

Por limitaciones de presupuesto no se pudo seleccionar muestra para cada subestrato estrato probabilístico, por lo cuál se agruparon todos estos en un único estrato probabilístico donde se utilizó el método Muestreo Aleatorio Simple -M.A.S.- para la selección de las unidades.

Para esto se utilizó el método coordinado negativo que consiste en realizar ***N*** (tamaño del estrato probabilístico) ensayos con una distribución de probabilidad uniforme (0,1), asignar estos números a cada uno de los elementos del universo, ordenar los elementos respecto a los valores aleatorios

y considerar como muestra los elementos correspondientes a los ***n*** (tamaño de muestra) valores aleatorios más pequeños. Así, se tiene que la probabilidad de selección de cada unidad muestral esta dada por:

$$P_i = \frac{n}{N}$$

donde, $P_i =$ Probabilidad de selección de la unidad $i = 1, 2, ..., n$

5.7.3. METODOLOGÍA DE ESTIMACIÓN

Teniendo en cuenta que la información obtenida a través de la encuesta es muestral, deben realizarse las expansiones e inferencias del caso, para poder restituir al universo de estudio, a continuación describimos los procedimientos.

Factores de expansión

A todos los individuos de una muestra probabilística se les debe asignar un factor de expansión, el cual, como su nombre lo indica, permite expandir los datos muestrales para obtener la estimación del parámetro en la población.

Es necesario ajustar el factor de expansión según las novedades de la información, las cuales se presentan al momento de recopilarla (empresas liquidadas, cambio de sector, inactivas, sin localizar, deuda, etc.).

Como se trata de un diseño de muestreo aleatorio simple estratificado (E.S.T. M.A.S), donde el mecanismo de selección fue aleatorio simple, el factor de expansión esta dado por:

$$F_k = \frac{N}{n}$$

donde, $F_k =$ factor de expansión de la unidad *k.*

$k = 1, 2, ..., n$ este factor es igual para todos los elementos del estrato probabilístico.

Ajuste del factor de expansión

$$Fajust = F \times \frac{UE}{UE - UENR}$$

donde,

UE = Unidades económicas esperadas en el estrato probabilístico
UENR = Unidades económicas que no respondieron en el estrato probabilístico

A su vez:

No respuesta = Deuda
Unidades económicas esperadas = Tamaño de la muestra

Por lo que, finalmente, el factor ajustado queda así:

$$Fajust = F \times \frac{No.esperado\ de\ entrevistas\ completas}{Nro.\ de\ entrevistas\ completas\ realizadas}$$

Estimadores

Dado que el interés básico de este estudio por muestreo es estimar los totales del universo en ciertos dominios definidos para la EAS, a continuación se explican los conceptos básicos que se tendrán en cuenta para realizar de manera adecuada dichas estimaciones.

Un dominio de estudio es una subpoblación para la cual se requieren estimaciones puntuales con buena precisión y con intervalos de confianza útiles. En este caso los dominios de estudio son: escalas de personal, escala de producción y organización jurídica

Sea la variable Z_{dk} definida como:

$$Z_{dk} = \begin{cases} 1 & si\ \ k \in U_d \\ 0 & si\ \ k \notin U_d \end{cases}$$

donde, k = Unidad económica , U_d = Dominio d

Luego,

$$\sum_U z_{dk} = N_d$$

N_d = Cantidad de elementos en el universo que pertenece al dominio *d*.

Ahora, sea la variable:

$$y_{dk} = x_k * z_{dk}$$

x_k = variable cuantitativa de interés y observada en la muestra

Bajo el diseño de muestreo estratificado, el total de un dominio, es:

t_d = total de la variable x en el dominio d

Parámetro

$$t_d = \sum_U y_{dk}$$

Estimador

$$\hat{t}_d = N_d \bar{y}_{dm} = \sum_m \mathrm{F_d} y_{dk}$$

con $\bar{y}_{dm} = \sum_m y_{dk} / n_d$, donde

U = Universo de estudio.

m = Unidades seleccionadas en la muestra.

y_{dk} = Valor de la variable para el elemento en el dominio.

$\bar{y}_{dm}$ = Promedio de la variable en el dominio.

n_d = Tamaño de la muestra en el dominio.

F = Factor de expansión para los elementos del estrato probabilístico.

La varianza del total de un dominio esta dada por:

$$V(\hat{t}) = N^2 \frac{1-f}{n} S^2_{ydU}$$

donde,

$f = \frac{n}{N}$, es la fracción de muestreo en el estrato probablilístico

$S^2_{ydU} = \frac{1}{N-1} \sum_U (y_{dk} - \bar{y}_{dU})^2$, es la varianza de la variable para el dominio d.

Un estimador insesgado de la varianza, es:

$$\hat{V}(\hat{t}) = N^2 \frac{1-f}{n} S^2_{ydm}$$

Donde la varianza muestral o estimada de la variable y para el dominio d es:

$$S^2_{ydm} = \frac{1}{n-1} \sum_m (y_{dk} - \bar{y}_{dm})^2$$

Errores muestrales

El objetivo de seleccionar una muestra es estimar, a través de ella, características que se desconocen de uno o varios aspectos de la población, como son: frecuencias de presentación de algún suceso, promedios, totales, proporciones, etc. En teoría de muestreo, estos valores se denominan parámetros.

Cuando la magnitud de la variabilidad es muy grande, los parámetros estimados pierden utilidad, pues el valor verdadero del parámetro en el universo, puede estar en un intervalo muy amplio, lo cual no proporciona información útil.

Una muestra probabilística es una parte de un universo, la cual se obtiene mediante selección aleatoria utilizando un diseño muestral p(.)[2], el cual asegura que todos y cada uno de los elementos del universo tiene probabilidad conocida y mayor de cero de ser incluidos en la muestra.

[2] Espacio de las probabilidades de selección

De una población se pueden obtener diferentes muestras y de cada una se obtiene una estimación del parámetro de interés, la forma como se distribuyen las estimaciones para las diferentes muestras se denomina distribución muestral, y la magnitud de la variabilidad de esta distribución debida al azar, es la varianza del estimador; entre menor sea la magnitud de esta variabilidad, se dice que mejor será la precisión de la estimación del parámetro de interés.

Con base en la estimación de dicha variabilidad, se construye el intervalo de confianza para la estimación, de modo que cuando la magnitud de la variabilidad es muy grande, el intervalo es muy amplio, lo que implica que se proporciona información poco útil.

La varianza del estimador esta dada en unidades generalmente de difícil manejo y por ello, se utiliza una medida relativa con base en valores porcentuales, denominada coeficiente de variación estimado *c.v.e* cuya fórmula es:

$$cve = \frac{\sqrt{\hat{V}(\hat{\theta})}}{(\hat{\theta})} \times 100$$

Criterios para utilizar el coeficiente de variación estimado

Una estimación se considera de buena calidad si su *c.v.e.* es menor del 5%; práctica, entre el 5 % y el 10%; aceptablemente practico, si es mayor del 10 % y menor del 15%, y de uso restringido, si es mayor del 15%. Por cuanto con un 95% de probabilidad, el intervalo que cubre el parámetro desconocido esta dado por:

$$\hat{\theta}\left(1-1{,}96\ cve(\hat{\theta}), 1+1{,}96\ cve(\hat{\theta})\right)$$

Para el mejor entendimiento del significado y los diferentes valores que toman los coeficientes de variación en los cuadros presentados, se deben considerar los siguientes aspectos.

- El subestrato de inclusión forzosa tiene la información de todas las unidades económicas que lo conforman y, por lo tanto, cuando la estimación se refiera sólo a este subestrato, el coeficiente de variación será cero pues sólo hay una muestra posible que corresponde exactamente a todas las unidades que conforman el subestrato y en consecuencia, no hay sino una única estimación del parámetro.

- El diseño de la muestra se realizó para obtener estimaciones, para totales nacionales, con un error de muestreo aproximado de 5%, para las variables ingresos y personal ocupado. Otro nivel de desagregación está sujeto a que su precisión no necesariamente sea buena.

BIBLIOGRAFÍA

DEPARTAMENTO ADMINISTRATIVO NACIONAL DE ESTADISTICA –DANE–. Metodología General de la Encuesta Anual de Servicios – EAS.

DEPARTAMENTO ADMINISTRATIVO NACIONAL DE ESTADISTICA –DANE–. Guía para documentar la actividad estadística. SENT, División de Calidad e Interventoría Estadística. 1997.

HIDIROGLOU, M.A.. The construction of a self-representing stratum of large units in survey design. 1986.

F. Mayda y P. Timmons. Survey Maintenance – Philosophy And Practice. 1981

OSPINA, B. David. Introducción al muestreo. Facultad de Ciencias. Universidad Nacional de Colombia. Bogotá. 2001

BAUTISTA, Leonardo. Diseños de muestreo estadístico. Universidad Nacional de Colombia. Bogotá. 1998.