

Departamento Administrativo Nacional de Estadística



**Dirección de Regulación, Planeación,
Estandarización y Normalización
-DIRPEN-**

**Diseño Muestral
Encuesta Ambiental Industrial
-EAI-**

Diciembre 2010

 DANE	DISEÑO MUESTRAL ENCUESTA AMBIENTAL INDUSTRIAL -EAI-	CÓDIGO: ES-EAI-DIM-01 VERSIÓN : 01 PÁGINA : 1 FECHA:10-12-10
ELABORÓ: EQUIPO TEMÁTICO	REVISÓ: COORDINADOR DE ESTUDIOS ESTADÍSTICOS	APROBÓ :DIRECTOR DIRPEN

TABLA DE CONTENIDO

1. MARCO MUESTRAL
2. TIPO DE MUESTREO
3. DEFINICIÓN TAMAÑO DE MUESTRA

1 Marco Muestral

Con el fin de proponer y aplicar un diseño de muestra se ha tomado como marco el directorio de la Encuesta Anual Manufacturera de 2006. Es un marco conformado por 6.309 establecimientos industriales que cumplieran con las condiciones de pertenecer a algún corredor industrial, reportar información anteriormente o ser un nuevo establecimiento dentro del directorio para el año 2006 y ser un establecimiento dentro perteneciente a algunas de las prioridades 1, 2 o 3.

Mediante ejercicios de reagrupación de las actividades industriales según la CIIU fueron establecidos seis (6) **corredores industriales** que corresponden a la agrupación geográfica por departamentos y municipios, de igual forma se establecieron cuatro nuevas prioridades las cuales indican el tamaño de las empresas por escalas de personal y su respectiva producción anual, reagrupando las cinco iniciales así: **1, 2, 3**, una última conformada por las prioridades 9 y 0 que se conocerá dentro de la investigación como **prioridad 4** la cual no fue considerada dentro de la población objetivo debido a que en la prueba piloto esta tuvo una tasa de no respuesta muy grande. Estadísticamente hablando, a los cruces entre corredores industriales y prioridades dentro del estudio se les denominará **dominios de estudio**. Las cifras obtenidas en la Investigación Ambiental tienen representatividad, respecto a la estrategia muestral adoptada compuesta por un diseño IF-ESTMAS y el estimador π – estimador, a nivel de los corredores industriales y por clases industriales.

2 Tipo de Muestreo

El diseño de muestreo aplicado es **probabilístico estratificado de elementos**. Dado que en cada dominio de estudio la población tiene una distribución asimétrica y por consideraciones de tipo temático, es necesario considerar varios grupos que se representan a si mismos, unidades conocidas en el argot estadístico como de inclusión forzosa (IF), y otros de inclusión probabilística (IP) con las unidades restantes seleccionadas para pertenecer a la muestra mediante diseños de muestreo aleatorio simple (MAS). Para la determinación de estos dos grandes grupos de inclusión forzosa y probabilística fueron tenidos en cuenta criterios proporcionados por parte de expertos en el tema ambiental.

Para el resultado del proceso con las debidas actualizaciones, se entregan 3.038 establecimientos pertenecientes a los 24 dominios de estudio o estratos de presentación.

3 Definición Tamaño de Muestra



**DISEÑO MUESTRAL
ENCUESTA AMBIENTAL INDUSTRIAL
-EAI-**

CÓDIGO: ES-EAI-DIM-01
VERSIÓN : 01
PÁGINA : 2
FECHA:10-12-10

Una vez se define el diseño muestral a utilizar, se procede a fijar los tamaños de muestra pertinentes en cada estrato.

En cada investigación se trata de buscar un tamaño de muestra necesario para obtener de antemano una precisión determinada en las estimaciones, sujeto a restricciones de tipo presupuestal. Para esta investigación en particular el análisis de tamaño de muestra se realizó con la variable **total de costos y gastos en protección ambiental** de un estudio anterior, realizado en 2005, en el cual fueron seleccionadas algunas industrias teniendo en cuenta consideraciones dadas por expertos y conformado por 1.169 establecimientos industriales. El tamaño de muestra total n , se obtuvo por medio del método de **Montecarlo** para simular la población objetivo sin considerar los establecimientos dentro de los estratos de inclusión forzosa. Una vez establecido el valor n del total de la muestra sin incluir los elementos de inclusión forzosa, como se mencionó anteriormente, los tamaños por estrato n_h se asignaron proporcionalmente al tamaño de cada estrato (**asignación proporcional a la variabilidad de la muestra**), puesto que, las dispersiones en cada uno de ellos diferían considerablemente de uno a otro en la prueba piloto.

De acuerdo a lo anterior, el tamaño establecido para la Investigación Ambiental es **$n=3037$** establecimientos industriales, distribuidos en los 24 dominios de estudio según el tipo de inclusión de cada estrato, en la tabla 2 se tienen los 12 estratos de inclusión forzosa determinados por 11 actividades CIIU y la Prioridad 1; en la tabla 3 se tienen los 12 estratos de inclusión probabilística determinados por otros 12 estratos que resultan de la combinación entre los seis corredores industriales y las prioridades 2 y 3 respectivamente. Los tamaños poblacionales de cada uno de los estratos se establecen automáticamente en el momento de la estatificación. Los tamaños muestrales con niveles de confianza no inferiores al 95%, antes de ser el resultado de fórmulas, son el resultado de análisis llevados a cabo por medio de simulación de Montecarlo tratando de encontrar la distribución que minimizara el error cuadrático medio entre la distribución poblacional buscada y la distribución empírica de la variable auxiliar (**total de costos y gastos ambientales**) tomada como referencia para simular la población y asumiendo una distribución Chi Cuadrado en la distribución al interior de éstos datos.

En el caso de la Encuesta Ambiental Industrial los tamaños de muestra están dados por:

Tabla 1. Tamaños de muestra establecimientos de inclusión forzosa 2007

ESTRATO	DISEÑO	N. ESTABLECIMIENTOS
Actividad CIU 1600	I. F.	7
Actividad CIU 2040	I. F.	16
Actividad CIU 2694	I. F.	25
Actividad CIU 2721	I. F.	4
Actividad CIU 3000	I. F.	1
Actividad CIU 3130	I. F.	6
Actividad CIU 3210	I. F.	8
Actividad CIU 3220	I. F.	4
Actividad CIU 3230	I. F.	6
Actividad CIU 3530	I. F.	4
Actividad CIU 3693	I. F.	8
PRIORIDAD 1	I. F.	1813
Total Establecimientos		1902

Fuente: DANE

Tabla 2. Tamaños de muestra establecimientos M. A. S. por corredor industrial y prioridad

ESTRATO	DISEÑO	N. ESTABLECIMIENTOS
Corredor Antioqueño Prioridad 2	M.A.S	199
Corredor Antioqueño Prioridad 3	M.A.S	42
Corredor Eje Cafetero Prioridad 2	M.A.S	46
Corredor Eje Cafetero Prioridad 3	M.A.S	11
Corredor Caribe Prioridad 2	M.A.S	75
Corredor Caribe Prioridad 3	M.A.S	14
Corredor Cundiboyacense Prioridad 2	M.A.S	353
Corredor Cundiboyacense Prioridad 3	M.A.S	120
Corredor Pacifico Prioridad 2	M.A.S	143
Corredor Pacifico Prioridad 3	M.A.S	40
Corredor Santanderes Prioridad 2	M.A.S	71
Corredor Santanderes Prioridad 3	M.A.S	21
Total Establecimientos		1135

Fuente: DANE

Procedimiento de estimación

De acuerdo a la teoría de muestreo, el factor de expansión es la capacidad que tiene cada individuo seleccionado en una muestra probabilística para representar el universo en el cual está contenido. Es decir, es la magnitud de representación que cada selección posee para describir una parte del universo de estudio. Cuando el diseño es M.A.S se asume que individuos dentro de una misma unidad de muestreo tienen la misma capacidad de representar al universo en consideración, en tanto que diferentes unidades de muestreo deben reflejar lo mejor posible la densidad y distribución del universo estudiado.

El factor de expansión por teoría para un diseño M.A.S. sobre k unidades de muestreo está definido por:

$$F_{Exp1k} = \frac{N_{1k}}{n_{1k}}$$

Donde

N_{1k} Denota el tamaño total de elementos en la unidad de muestreo.

n_{1k} Denota el número de elementos a ser seleccionados dentro de la unidad de muestreo.

Metodología de estimación de totales

En el caso del total; el estimador es el siguiente:

$$\hat{t}_{\pi} = \sum_h \hat{t}_{\pi h} \quad (3)$$

Donde

$$\hat{t}_{\pi h} = \frac{N_h}{n_h} \sum_{k \in h} y_k \quad (4)$$

Con varianza

$$V_{EST}(\hat{t}_{\pi}) = \sum_h V_h(\hat{t}_{\pi h}) \quad (5)$$

El cual tiene estimador insesgado de la varianza

$$\hat{V}_{EST}(\hat{t}_{\pi}) = \sum_h \hat{V}_h(\hat{t}_{\pi h}) \quad (6)$$

Donde

$$\hat{V}_h(\hat{t}_{\pi h}) = \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_{y_{sh}}^2 \quad (7)$$

Con

$$S_{y_{sh}}^2 = \frac{1}{n_h - 1} \sum_h (y_k - \bar{y}_{s_h})^2 \quad (8)$$

Para todo $h=1,2,\dots,H$

Metodología de estimación de Razones

Se utilizo la técnica de linealización de Taylor, la cual está dada por:

$$\hat{R} = \frac{\hat{t}_{y\pi}}{\hat{t}_{z\pi}} \quad (9)$$

$$u_k = a_1 y_k + a_2 z_k = \frac{y_k}{t_z} - \frac{R_{zk}}{t_z} = \frac{1}{t_z} (y_k - R_{zk}) \quad (10)$$

Donde

$$a_1 = \frac{\partial \left(\frac{\hat{t}_{y\pi}}{\hat{t}_{z\pi}} \right)}{\partial \hat{t}_{y\pi}} = \frac{1}{\hat{t}_{z\pi}} \quad \text{Evaluada en } \hat{t}_{y\pi} = t_{y\pi} \Rightarrow a_1 = \frac{1}{t_{z\pi}} \quad (11)$$

$$a_1 = \frac{\partial \left(\frac{\hat{t}_{y\pi}}{\hat{t}_{z\pi}} \right)}{\partial \hat{t}_{z\pi}} = \frac{-\hat{t}_y}{\hat{t}_\pi^2} \text{ Evaluada en } (\hat{t}_y, \hat{t}_z) = (t_y, t_z) \Rightarrow a_2 = \frac{-R}{t_z} \quad (12)$$

Con varianza y estimador de la varianza:

$$AV(\hat{R}) = \frac{1}{\hat{t}^2} \sum \sum_U \Delta_{kl} \frac{(y_k - Rz_k)}{\pi_k} \frac{(y_l - Rz_l)}{\pi_l} \quad (13)$$

$$\hat{V}(\hat{R}) = \frac{1}{\hat{t}^2} \sum \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{(y_k - \hat{R}z_k)}{\pi_k} \frac{(y_l - \hat{R}z_l)}{\pi_l} \quad (14)$$

Factor de Expansión Ajustado

Es necesario generar un factor de expansión que considere los establecimientos fuera del universo y la no respuesta, teniendo en cuenta que no se modifique el diseño y tampoco sufra alteraciones.

$$F_{Ajust} = \frac{n_{Ik} - n_{Ifu}}{n_{Ik} - n_{Ifu} - n_{Ir}} \quad (15)$$

Con

n_{Ik} = número total de establecimientos en el estrato I.

n_{Ifu} = número de establecimientos en el estrato I considerados fuera de universo.

n_{Ir} = número de establecimientos en el estrato I considerados rechazo.

Por tanto el factor de expansión final se define como:

$$F_{Exp}^* = F_{Exp_{Ik}} * F_{Ajust} \quad (16)$$

Cálculo de precisión de los resultados

Confianza

Por lo general, se acostumbra a trabajar con estrategias muestrales que aseguren un 95% de confianza respecto a la muestra y 5% de confianza al estadístico de cometer un error cuando se dijo que no lo iba a cometer. Y esta restricción debe interpretarse como una forma de tener precauciones con los resultados de un estudio; es decir, cuando un estudio se lleva a cabo con una excelente estrategia y siguiendo todos los procesos estadísticos pertinentes, por bien que este hecho solo se puede confiar en los resultados del estudio en un 95% y existe el 5% de posibilidad que las conclusiones hayan salido erradas debido a la naturaleza incontrolable de los fenómenos aleatorios presentes en la naturaleza.

Precisión

La precisión es la capacidad de un estimador de dar el mismo resultado en mediciones diferentes realizadas en las mismas. Es decir, cuando una encuesta ha publicado un resultado que proviene de una muestra seleccionada por métodos probabilísticos, este resultado no es un término absoluto sino que puede estar en un rango de valores o puede variar

La Encuesta Ambiental Industrial, específicamente fue diseñada para satisfacer, acorde a la ficha metodológica, los siguientes lineamientos

- Nivel de confianza de 95%

Precisión con errores muestrales menores de 15% (CVE) para el cálculo de totales en categorías de frecuencias