

Roberto Castillo
DIRECTOR EJECUTIVO (e)

Magaly Paredes
SUBDIRECTORA GENERAL (e)

Markus Nabernegg
COORDINADOR GENERAL TÉCNICO DE PRODUCCIÓN ESTADÍSTICA

Christian Garcés
DIRECTOR DE INFRAESTRUCTURA ESTADÍSTICA Y MUESTREO

Equipo técnico:

Salome Borja, Javier Núñez, Francisco Céspedes, Christian Garcés.

Propiedad Institucional

©INEC

Instituto Nacional de Estadística y Censos
Juan Larrea N15-36 y José Riofrío. Casilla postal 135 C
Telf: (02) 2555-701 / 2529-858

Citar como:

INEC (2018). Metodología de Diseño Muestral de la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU). Instituto Nacional de Estadística y Censos, Quito-Ecuador.

Contenidos

1.	Introducción	5
2.	Determinación del Marco de Muestreo	5
3.	Universo de investigación	5
4.	Unidad de observación	6
5.	Unidad de análisis.....	6
6.	Cobertura geográfica	6
7.	Diseño Muestral	6
8.	Equilibrio de las Unidades Primarias de Muestreo (UPM)	6
9.	Estratificación del Marco de Muestreo	7
9.1.	Metodología de Estratificación	8
a)	Análisis de componentes principales mediante mínimos cuadrados alternantes.....	8
b)	Análisis de componentes principales clásico	9
c)	Dimensiones y variables de estratificación	10
d)	Asignación de puntajes para las dimensiones y variables de estratificación	10
e)	Clasificación K-medias	11
9.2.	Aplicación de la estratificación en el Marco de Muestreo.....	12
10.	Dominios de estudio	13
11.	Tamaño y selección de la muestra	13
11.1.	Tamaño de muestra de personas.....	13
11.2.	Tamaño de muestra de viviendas	14
11.3.	Tamaño de muestra de UPM's	14
11.4.	Asignación de la muestra	15
11.5.	Selección de la muestra	16
12.	Rotación de la muestra por paneles.....	16
13.	Cálculo de los factores de expansión	17
13.1.	Ponderación de primera etapa (conglomerado)	18
13.2.	Ajuste por no respuesta a nivel de conglomerado	18
13.3.	Ponderación de segunda etapa (vivienda)	19
13.4.	Ajuste por no respuesta a nivel de vivienda.....	19
13.5.	Ponderadores muestrales totales.....	19
13.6.	Factores de expansión ajustados (población)	20
13.7.	Ajuste de los Factores de Expansión por Post-Estratificación.....	20
14.	Estimación de características de la población	21
15.	El error de muestreo de diseño	21
16.	Métodos de estimación de errores para diseños muestrales complejos	22

Tablas

Tabla 1: <i>Dimensiones y variables de estratificación</i>	10
Tabla 2: Tamaño de muestra a nivel de viviendas y conglomerados por área geográfica	15
Tabla 3: Rotación de paneles periodo 2018 - 2019	16
Tabla 4: Variables requeridas para declaración del diseño muestral - ENEMDU	24

1. Introducción

El Instituto Nacional de Estadística y Censos (INEC), como ente rector del Sistema de Estadística Nacional y dentro de su plan de mejoramiento continuo de las operaciones estadísticas por muestreo, mediante la Dirección de Infraestructura Estadística y Muestreo (DINEM), busca innovar la metodología de los procesos estadísticos que históricamente se han llevado a cabo dentro de la institución.

Así, desde 2018, se plantean algunas mejoras puntuales al diseño muestral de la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU). La técnica de muestreo de la ENEMDU, es similar a la utilizada en años anteriores y corresponde a un muestreo probabilístico en dos etapas, con estratificación geográfica por dominios de estudio y área urbana y rural. Los estimadores asociados al diseño se calibran por una proyección de población calculada según métodos demográficos. Sin embargo, el nuevo diseño contempla la afinación del marco de muestro que incluye la mejora en los siguientes aspectos:

- Equilibrio de las Unidades Primarias de Muestreo (UPM)
- Estratificación acorde a las UPM's equilibradas
- Optimización de la dispersión de la muestra

Los cambios implementados tienen el objetivo de mejorar la precisión de los estimadores y oportunidad de la información.

2. Determinación del Marco de Muestreo.

El marco de la muestra se basa en los resultados definitivos y la cartografía del VII Censo de Población y VI de Vivienda del 2010 (CPV-2010); este Marco tiene la característica de ser constituido por áreas geográficas, que tienen límites perfectamente definidos e identificables sobre el terreno.

El marco ha tenido actualizaciones parciales en función a los cambios presentados en las de unidades de observación (viviendas), en determinadas áreas geográficas, debido principalmente a los movimientos y dinámica demográfica propios de la población, así como los cambios en estructuras habitacionales; que en su conjunto hacen necesario un mantenimiento continuo del Marco de Muestreo.

Se han ejecutado en el periodo 2014 – 2017 diferentes procesos de actualización cartográfica con diferente alcance, las cuales se detallan a continuación:

- Encuesta Condiciones de Vida 2013-2014: 2.425 sectores censales
- Actualización ENEMDU 2014: 5.564 sectores censales
- Proyecto 2015: 548 sectores censales
- Actualización 2017: 1.779 sectores censales

3. Universo de investigación

El universo de estudio de la ENEMDU son personas de 5 y más años de edad, residentes en los hogares del Ecuador, exceptuando la población que reside en viviendas colectivas, viviendas flotantes y población indigente (sin techo).

4. Unidad de observación

La unidad de observación son todas las viviendas particulares ocupadas que se encuentran en territorio nacional, las mismas que tienen ligada su identificación geográfica mediante fuentes cartográficas.

5. Unidad de análisis

Para el caso de los indicadores laborales, la población de referencia son todas las personas de 15 años en adelante.

6. Cobertura geográfica

La cobertura geográfica está definida por las viviendas ocupadas que se encuentren ubicadas dentro del territorio ecuatoriano incluyendo la región insular.

7. Diseño Muestral

El diseño muestral para esta investigación corresponde a un muestreo probabilístico con dos etapas de selección:

- Primera etapa: selección de Unidades Primarias de Muestreo (conglomerados) por estrato.
- Segunda etapa: selección de viviendas ocupadas dentro de cada uno de los conglomerados seleccionados en la primera etapa.

8. Equilibrio de las Unidades Primarias de Muestreo (UPM)

El muestreo es probabilístico y bi-etápico; la Unidad Primaria de Muestreo (UPM) es el conglomerado y la Unidad Secundaria de Muestreo (USM) son las viviendas ocupadas.

El diseño vigente contempla una actualización dentro de la construcción de la Unidad Primaria de Muestreo – (UPM). Para años anteriores, se realizaba la selección de “sectores censales” en función a un criterio operativo para la ejecución y levantamiento de información del Censo de Población y Vivienda. Sin embargo, en función al crecimiento y disminución de la población en ciertas áreas geográficas por el transcurso del tiempo, estas UPM’s pasaron a ser heterogéneas en cuanto al número de viviendas ocupadas que tienen dentro de sus límites (originalmente, en promedio, 150 viviendas ocupadas en el área amanzanada y 80 en el área dispersa), situación que traería como consecuencia probabilidades de selección inadecuadas en la segunda etapa.

Este inconveniente se ve solventado al reconstruir y equilibrar el tamaño de las UPM con respecto al número de viviendas ocupadas. El problema de equilibrar el tamaño de viviendas ocupadas por unidad primaria de muestreo, denominada conglomerado, está asociado a un problema de optimización combinatoria, el mismo que a su vez es un problema de partición de grafos¹, donde los nodos son las manzanas y las aristas están dadas por la existencia de una frontera común entre dos

¹ Típicamente, un grafo se representa gráficamente como un conjunto de puntos (vértices o nodos) unidos por líneas (aristas).

manzanas; es decir, el procedimiento cerciora proximidad entre viviendas tomando en cuenta sus límites físicos y a la vez asegura un número determinado de viviendas para que pueda considerarse como una UPM.

Para resolver este problema se utiliza el "Algoritmo de generación de conglomerados para necesidades operativas -AGCNO". Este algoritmo garantiza que el número de viviendas por conglomerado se encuentre entre un número t de viviendas ocupadas hasta un número $2t$; de esta manera, se asegura que los factores de expansión (ponderación) de diseño pertenezcan a un intervalo que va desde s hasta $2s$.

El número de viviendas ocupadas que tienen un conglomerado (UPM) actualmente es de 30 a 60 viviendas, tanto para el área amanzanada como para el área dispersa.

9. Estratificación del Marco de Muestreo

Una técnica muy extendida para diseñar una encuesta de hogares es la estratificación de la población que se desea encuestar previa a la selección de la muestra. Esta sirve para clasificar a la población en subpoblaciones (estratos) basándose en información auxiliar conocida sobre toda la población. Después, de cada estrato se seleccionan, independientemente, los elementos de la muestra de una forma coherente con los objetivos de medición de la encuesta.

La estratificación se aplica para asegurar la representación adecuada de los grupos de subpoblación importantes sin sesgar la operación de selección.

Con la estratificación se persiguen diversos fines:

- Dar estimaciones separadas para ciertas subpoblaciones
- Agrupar unidades homogéneas entre sí para mejorar la precisión de las estimaciones globales.
- Posibilidad de utilizar diferentes métodos de muestreo en los distintos estratos.

En esta línea, el INEC ha realizado el diseño muestral de la ENEMDU (desde 2018-2019) considerando un muestreo bi-etápico estratificado, en el cual, una vez construidas las UPM's equilibradas en número de viviendas para todo el territorio, corresponde realizar una clasificación de estas por ciertas características e información suministrada por el Censo de Población y Vivienda 2010, del cual se extraen e identifican las principales variables para la determinación de los estratos.

Existen dos normas básicas que deben aplicarse al estratificar una población:

- Los estratos son subconjuntos de población esencialmente independientes y mutuamente excluyentes: cada elemento de una población debe pertenecer a un estrato y solo a uno. Teniendo en cuenta esta característica, cada estrato debe estar incluido en la muestra para que toda la población quede incluida y pueda calcularse una estimación no sesgada de la media de la población.
- Cada estrato debería ser lo más diferente posible de los demás. La heterogeneidad entre estratos y la homogeneidad dentro de cada estrato constituyen la característica básica que rige la creación de los estratos.

A continuación se describirá la metodología y proceso aplicado para el cálculo de la estratificación del Marco de Muestreo.

9.1. Metodología de Estratificación

Existen varias metodologías que permiten, mediante técnicas matemáticas, la creación de subconjuntos aplicando el análisis de variables cuantitativas como cualitativas. Se debe considerar que la fuente principal de información es el Censo de Población y Vivienda 2010 (CPV 2010), dependiendo de las características de sus variables, se ha tomado como alternativa las siguientes técnicas:

- Para las variables cualitativas, se emplea el análisis de componentes principales no lineales. En este caso en particular, se utilizó la técnica del escalamiento óptimo, mediante mínimos cuadrados alternantes; el cual permite transformar las variables originales asignando valores a las categorías de cada una de las variables y luego correlacionarlas para caracterizar o analizar la estructura de los datos, es decir, se transforma las variables cualitativas en cuantitativas con el fin de mejorar la combinación lineal de las variables tratadas y minimizando problemas como subjetividad y manipulación.
- Mientras que para variables cuantitativas, se parte de un análisis de componentes principales, el cual tiene como principal objetivo reducir el grupo de variables seleccionadas que representan la naturaleza observada, por un grupo más pequeño de variables no correlacionadas que presenten la mayor parte de información que se encuentra en las variables originales; en este caso con las variables que pertenecen a los módulos de vivienda, hogar y personas del CPV 2010.

a) Análisis de componentes principales mediante mínimos cuadrados alternantes

La metodología se basa en la aplicación de componentes principales mediante mínimos cuadrados alternantes (escalamiento óptimo). Esta metodología se desprende del modelo de componentes principales clásico y un doble algoritmo de escalamiento óptimo, el mismo que, cuantifica las categorías de las variables para maximizar la correlación entre las variables seleccionadas de manera inicial. Posteriormente, con las cuantificaciones obtenidas se estima los parámetros del análisis de componentes principales lineal.

Por otro lado, se denominan estimaciones de mínimos cuadrados condicionales, debido a que las estimaciones de mínimos cuadrados de los parámetros en un subconjunto asumen que los parámetros en todos los demás subconjuntos son constantes, puesto que la naturaleza de los mínimos cuadrados es condicional a los valores de los otros parámetros en otros subconjuntos.

Se debe considerar que el primer componente principal es la suma ponderada de las variables originales con mayor varianza, es decir, que las variables nuevas generadas por los componentes principales son combinaciones lineales (sumas ponderadas) de las variables originales.

Para generalizar un conjunto de múltiples variables al grupo de índices J de las j variables, se realiza una partición en un subconjunto B , explicado de la siguiente manera:

$$\sigma(x; y_1, \dots, y_j) = B^{-1} \sum_b SSQ(x - \sum_{j \in J(b)} G_j y_j)$$

Como parte del escalamiento óptimo, es importante destacar que las categorías de las variables originales deben pasar por un proceso de recodificación de mayor a menor en su numeración, es decir, si existe en una variable con cinco categorías, éstas según su frecuencia, deben ser recodificadas, considerando que la categoría con mayor frecuencia deberá tener el máximo valor es decir cinco. Este procedimiento se lo debe llevar a cabo con cada una de las variables seleccionadas.

b) Análisis de componentes principales clásico

Esta metodología señala las relaciones que se presentan entre n variables correlacionadas, que se transforman en un nuevo conjunto de variables sin correlación entre sí (que no redundan en información). Este conjunto se denomina conjunto de componentes principales. A su vez son combinaciones lineales de las anteriores variables, construidas por su orden de importancia o según la dimensión a la cual pertenezcan.

Para el análisis de componentes principales, no se requiere el supuesto de normalidad multivariante, no obstante, se debe considerar que las variables fueron escaladas, para evitar subjetividad con sus categorías. Se debe tomar en consideración una serie de variables sobre un grupo de individuos (vivienda, hogar y personas), con lo cual se calcula un nuevo grupo de variables cuya varianza decrezca de una manera progresiva.

Las variables escaladas son clasificadas por dimensiones entre vivienda, hogar y personas, para obtener un mejor resultado y minimizar la varianza en el ACP_TOTAL, el cual está compuesto por los resultantes del ACP_VIVIENDA, ACP_HOGAR y ACP_PERSONAS. De manera general, se obtiene un vector resultante para posteriormente evaluarlo y clasificarlo:

$$\sum_{i=1}^p \text{Var}(y_i) = \sum_{i=1}^p \lambda_i = \text{traza}(\Delta)$$

Dónde:

$\Delta \equiv$ matriz diagonal

Con el fin de contar con el porcentaje de la varianza total, se recoge un componente principal:

$$\frac{\lambda_i}{\sum_{i=1}^p \lambda_i} = \frac{\lambda_i}{\sum_{i=1}^p \text{var}(x_i)}$$

Cabe recalcar que no se suele seleccionar más de tres componentes principales, con el fin de facilitar el análisis, es decir, se debe seleccionar aquellas variables que explican una proporción aceptable de la varianza global.

c) Dimensiones y variables de estratificación

Tras realizar el cálculo del análisis de componentes principales y obteniendo los resultados de la aplicación se obtienen tres dimensiones y un total de 21 variables de estratificación, el cual se puede observar en la tabla 1.

Tabla 1: Dimensiones y variables de estratificación

VIVIENDA	Tipo de viviendas
	Vías de acceso
	Materiales del techo
	Materiales del piso
	Materiales de las paredes
	Procedencia del agua
	Agua que recibe la vivienda
	Servicio higiénico (eliminación de excretas)
	Eliminación de Basura
HOGAR	Disponibilidad de servicio higiénico del hogar
	Disponibilidad de ducha
	Teléfono convencional
	Teléfono celular
	Internet
	Computador
	TV cable
	Equipamiento del Hogar
	Tenencia de la vivienda
POBLACIÓN	Nivel de instrucción
	Grado
	Características económicas

d) Asignación de puntajes para las dimensiones y variables de estratificación

La metodología de componentes principales clásico proporciona una asignación numérica del 1 al 1.000 para observar de mejor manera la posible agrupación que pertenece por sus características, es por ello que se utiliza el siguiente algoritmo de clasificación:

$$S = \frac{ACP_TOTAL - score_{min}}{score_{max} - score_{min}}$$

Dónde:

S	=	Valor del puntaje final
ACP_TOTAL	=	Valor del análisis de componentes principales clásico
score _{min}	=	Valor mínimo obtenido del σ
score _{max}	=	Valor máximo obtenido del σ

A partir de esta puntuación se identifica cada una de las viviendas con sus hogares para realizar la clasificación final, se toman en consideración la presencia de cada una de las dimensiones y variables de estratificación en cada uno de los dominios, los cuales dan como resultado una agrupación única.

e) Clasificación K-medias

Debido a las diferencias en la métrica de las variables, se las transforman antes del análisis, de manera que todas ellas tengan variabilidades similares, con el fin de eliminar del cálculo las distancias.

De esta manera, se aplica el análisis de conglomerados de K-medias, el cual se basa en las distancias euclídeas existentes entre un conjunto de variables, asignando de los K primeros casos a los K conglomerados (centroídes).

$$d_{ii'} = \sqrt{\sum_j (x_{ij} - x_{i'j})^2}$$

Dónde:

X se refiere a las puntuaciones obtenidas por el caso i y el caso i' ($i \neq i'$) en cada una de las $j = 1, 2, \dots, p$ variables incluidas en el análisis (la sumatoria de la expresión incluirá p términos, es decir, tantos p como variables).

9.2. Aplicación de la estratificación en el Marco de Muestreo

El Marco de Muestreo aplicado para la ENEMDU desde 2018-2019, se lo ha dividido por dominios de estudio, y dentro de ellos sus correspondientes conglomerados, a los cuales se asignó un estrato tomando principalmente en entre otras características geográficas, socio-económicas y socio-demográficas, con la finalidad de mejorar la precisión y exactitud de los estimadores, minimizando su varianza.

Diferenciar entre área urbana y rural es necesario, lo cual se considera como estrato implícito dentro del Marco de Muestreo. Para ello, se toma en cuenta la división o identificación de centros poblados²(criterio poblacional) sugerido por la Comunidad Andina de Naciones (CAN); en el – “Seminario: Censos 2000 de Población y Vivienda de los Países Andinos”, clasificándolos de la siguiente manera:

Rural:

Población dispersa y centros poblados con menos de 2000 habitantes.

Urbano:

- 2.000 a 4.999 habitantes
- 5.000 a 9.999 habitantes
- 10.000 a 19.999 habitantes
- 20.000 a 49.999 habitantes
- 50.000 a 99.999 habitantes
- 100.000 a 199.999 habitantes
- 200.000 a 499.999 habitantes
- 500.000 a 999.999 habitantes y más habitantes

Para las encuestas a hogares se considera esta clasificación de área urbana y rural, por ende, la estratificación también usa este criterio para la asignación de los estratos.

La identificación de estrato para cada conglomerado a nivel nacional respetan los límites geográficos de las jurisdicciones a las cuales pertenecen, para que no existan rupturas con la asignación cartográfica preestablecida.

Los estratos creados para los dominios del Marco de Muestreo que se presentan son: (1) Estrato A, (2) Estrato B, (3) Estrato C.

² Es todo lugar del territorio nacional rural o urbano, identificado mediante un nombre y habitado con ánimo de permanencia. Sus habitantes se encuentran vinculados por intereses comunes de carácter económico, social, cultural e histórico. Dichos centros poblados pueden acceder, según sus atributos, a categorías como: caserío, pueblo, villa, ciudad y metrópoli.

10. Dominios de estudio

Los dominios de estudio para la ENEMDU trimestral (marzo, junio, septiembre y diciembre) son los siguientes:

- Trimestral puntual:
 - 5 ciudades: Quito, Guayaquil, Cuenca, Ambato y Machala,
 - 3 restos urbanos: Costa, Sierra y Amazonía; y
 - Resto nacional rural.

11. Tamaño y selección de la muestra

A partir de las expresiones de tamaño de muestra para muestreos complejos, se calcula el tamaño de muestra necesario para lograr una precisión adecuada en la inferencia.

11.1. Tamaño de muestra de personas

Para la determinación del tamaño de la muestra se requiere establecer la característica o características a estimar, el nivel de confianza y la precisión requeridos.

En este sentido, la variable de investigación, en base a la cual se calcula el tamaño de la muestra es el "desempleo". Los valores poblacionales (marco muestral) de esta variable se tomaron de la base de datos de la Encuesta de Empleo y Desempleo (ENEMDU – diciembre 2012).

La expresión matemática utilizada para el cálculo de tamaño de muestra es:

$$n_k \geq \frac{z_{1-\frac{\alpha}{2}}^2 N_k p(1-p)}{(e_r p)^2 (N_k - 1) + z_{1-\frac{\alpha}{2}}^2 p(1-p)} \times \frac{def}{1 - tnr} \quad (1)$$

Donde:

- n_k = Tamaño de la muestra de personas para el dominio de estudio k
- $z_{1-\frac{\alpha}{2}}$ = Nivel de confianza correspondiente al percentil α .
- N_k = Total de personas de 15 años y más en el dominio de estudio k .
- p = Probabilidad de incidencia de la variable a investigar
- e_r = error relativo.
- def = efecto de diseño.
- tnr = tasa de no respuesta.

Los parámetros que se utilizan para el cálculo aplicado son recomendados para los estudios de este tipo, la estrategia asegura por una parte el rigor estadístico de la muestra y por otra, la posibilidad de comparar entre sí los estudios ejecutados. Para que las estimaciones sean comparables, se utilizó un nivel de confianza (95%) y margen de error relativo (0.15).

11.2. Tamaño de muestra de viviendas

Es necesario calcular el número total de viviendas que deben ser seleccionados para lograr entrevistar a todas las personas que serán observadas en el punto anterior. El número de viviendas que deben ser seleccionados estará determinado por la muestra de personas (n_k), número promedio de personas por hogar (b) y el porcentaje de personas con la característica de interés ($r = 15$ años y más), de la siguiente forma:

$$n_{vk} = \frac{n_k}{r_k * b_k}$$

11.3. Tamaño de muestra de UPM's

Los hogares y las personas se observan a partir de las UPM. En este paso final es necesario calcular el número de UPM's que deben ser seleccionadas en el muestreo a partir de la relación:

$$n_{UPMk} = \frac{n_{vk}}{\text{Carga técnica operativa}}$$

La carga técnica operativa se refiere al número de viviendas asignadas a cada encuestador como carga de trabajo; que fue el resultado de un previo análisis de correlación intraclase donde se pudo verificar, mediante simulaciones matemáticas, el número de observaciones necesarias para minimizar la varianza dentro de cada conglomerado. Como resultado de este procedimiento se definió tanto operativa como técnicamente que el número de viviendas por UPM será siete (7).

El total de UPM's seleccionadas de acuerdo a la temporalidad de la ENEMDU-trimestral es:

Tabla 2: Tamaño de muestra a nivel de viviendas y conglomerados por área geográfica

Cód. Provincia	Provincia	Viviendas	Muestra trimestral UPM
1	Azuay	1.456	208
2	Bolívar	266	38
3	Cañar	294	42
4	Carchi	224	32
5	Cotopaxi	504	72
6	Chimborazo	532	76
7	El Oro	1.190	170
8	Esmeraldas	532	76
9	Guayas	3.024	432
10	Imbabura	406	58
11	Loja	490	70
12	Los Ríos	770	110
13	Manabí	1.232	176
14	Morona Santiago	224	32
15	Napo	182	26
16	Pastaza	168	24
17	Pichincha	2.394	342
18	Tungurahua	1.624	232
19	Zamora Chinchipe	182	26
20	Galápagos	168	24
21	Sucumbios	210	30
22	Orellana	196	28
23	Santo Domingo de los Tsáchilas	364	52
24	Santa Elena	266	38
90	Zonas no delimitadas	168	24
Total general		17.066	2.438

11.4. Asignación de la muestra

La muestra es distribuida proporcionalmente por cada estrato dentro de cada dominio de estudio.

$$n_{ki} = n_k \cdot \frac{N_{ki}}{N_k}$$

Donde:

- n_{ki} = Tamaño de la muestra para el estrato i del dominio de estudio k
 n_k = Tamaño de la muestra para el dominio de estudio k
 N_{ki} = Total de viviendas ocupadas en el estrato i del dominio de estudio k
 N_k = Total de viviendas ocupadas en el dominio de estudio k .

11.5. Selección de la muestra

La selección de las UPM's que forman parte de la muestra, de acuerdo al tamaño establecido, se realiza de manera independiente en cada uno de los dominios de forma aleatoria, asignando a cada UPM igual probabilidad de ser seleccionada. De la misma forma, la selección de viviendas es aleatoria dentro de cada UPM seleccionada.

12. Rotación de la muestra por paneles

El objetivo de realizar una encuesta tipo panel, es medir los cambios en los indicadores entre trimestres y años consecutivos, así como la estacionalidad en el tiempo.

Cada UPM pertenece a un panel, el cual está constituido por siete viviendas efectivas u originales, es decir, el panel y su seguimiento será a nivel de conglomerado.

La rotación de paneles sigue el esquema 2-2-2, es decir, un panel es visitado en dos trimestres consecutivos, en los siguientes dos trimestres no se realiza ninguna visita y en los siguientes dos trimestres la segunda visita. Esta metodología se la viene aplicando desde el año 2007 donde cada panel de rotación tiene una duración de dos años.

Una de las bondades de esta propuesta de rotación, es que es compatible con un proceso de actualización cartográfica permanente, donde un panel ingresa con actualización de la condición de ocupación de las viviendas que lo componen, con la finalidad de garantizar cobertura muestral para la investigación, además se asegura que un panel tenga como máximo un periodo de desactualización de 2 años.

Tabla 3: Rotación de paneles periodo 2018 - 2019

mar-18	jun-18	sep-18	dic-18
E	E	L	L
F	I	I	F
Z03	Z06	Z09	Z12

mar-19	jun-19	sep-19	dic-19
E	E	L	L
F	I	I	F
Z03	Z06	Z09	Z12

En la tabla 3 se puede observar la presencia de cada uno de los paneles durante un periodo de dos años. Cada panel (E, F, I y L) representan el 40,7% con respecto al total de la muestra, con excepción del panel Z que le corresponde el 18,6%.

Por ejemplo, entre marzo y junio de 2018; el panel E estará presente en los meses de marzo 2018 y junio 2018, donde es posible analizar la probabilidad de permanencia y transición de la Población en edad de trabajar (PET), para los principales indicadores de mercado laboral.

Por último, se debe considerar que los paneles a los que se les asigna el panel Z son aquellos que se investigan para completar el total de muestra necesaria para garantizar representatividad en los dominios de la investigación y tendrán seguimiento solamente de manera anual.

13. Cálculo de los factores de expansión

El procedimiento de ponderación general para la ENEMDU requiere de dos etapas.

- La **primera etapa** consiste en calcular una ponderación de conglomerado. De ser necesario, se aplica un ajuste por no respuesta a nivel de conglomerado.
- La **segunda etapa** consiste en calcular una ponderación de viviendas dentro del conglomerado. De ser necesario se aplica un ajuste por no respuesta a nivel de vivienda.

La “falta de respuesta” se produce cuando no se llega a obtener respuesta de algunas unidades de la muestra. Resulta útil pensar en la población de la muestra dividida en dos conjuntos, el primero, formado por todas las unidades de la muestra de las que se ha obtenido respuesta; y el segundo, por todas las unidades de la muestra de las que no ha podido obtenerse respuesta.

La tasa de falta de respuesta puede estimarse con exactitud si se llevan a cabo recuentos de todos los elementos que cumplen los requisitos que se incluyen en la muestra. La tasa de respuesta en una encuesta se define como el coeficiente entre el número de cuestionarios completados por las unidades y el número total de unidades de la muestra que cumplen los requisitos³. La falta de respuesta puede deberse a la ausencia del domicilio de las personas seleccionadas, a los cambios en la condición de ocupación de la vivienda, a la negación de dichas personas a participar o a su incapacidad para responder a las preguntas. Otra posible causa de la falta de respuesta puede ser la imposibilidad de llevar a cabo la encuesta en determinadas zonas por razones climatológicas, dificultades del terreno o cuestiones de seguridad.

La ponderación total correspondiente a cada vivienda es el producto de las dos ponderaciones intermedias.

El procedimiento a seguir para el cálculo de los factores de expansión se describe a continuación:

³ Algunas unidades incluidas en la muestra puede que se hallen fuera del ámbito de la encuesta y no cumplan los requisitos. Tal podría ser el caso, por ejemplo, de las viviendas vacías, en ruina o abandonadas.

13.1. Ponderación de primera etapa (conglomerado)

La ponderación de primera etapa representa el inverso de la probabilidad de selección de la primera etapa asignada a cada conglomerado.

Como se mencionó en apartados anteriores, una vez realizado el equilibrio en el número de viviendas que constituyen las UPM, es posible obtener la probabilidad de selección de cada conglomerado de una manera aleatoria simple dentro de cada estrato.

La ponderación básica de primera etapa para el i -ésimo conglomerado muestreado en el estrato k queda definida por:

$$PB_c^{ki} = \frac{C_k}{c_k},$$

donde C_k es el número de conglomerados en el estrato k , c_k es el número de conglomerados muestreados en el estrato k .

13.2. Ajuste por no respuesta a nivel de conglomerado

Para realizar este ajuste, es necesario definir la tasa de no respuesta a nivel de conglomerado, la cual mide la proporción de conglomerados seleccionados que efectivamente participaron en la encuesta. El cálculo de la tasa de no respuesta usa el siguiente término, derivado de la recolección de datos:

c_{kex} = número de conglomerados seleccionados que fueron excluidos en el estrato k

La tasa de no respuesta viene dada por:

$$TNR_k = \frac{c_k - c_{kex}}{c_k}$$

Las ponderaciones de la primera etapa se calculan para los conglomerados seleccionados. Si existieran conglomerados excluidos de la investigación se debe recalculan la ponderación de primera etapa para el i -ésimo conglomerado investigado en el estrato k , queda definida por:

$$PF_c^{ki} = \frac{PB_c^{ki}}{TNR_k}$$

13.3. Ponderación de segunda etapa (vivienda)

La ponderación de segunda etapa representa el inverso de la segunda probabilidad de selección asignada a cada vivienda. Las viviendas fueron seleccionadas con igual probabilidad de selección, por lo que la ponderación básica de segunda etapa para la vivienda j en el conglomerado i viene dada por:

$$PB_v^{ij} = \frac{V_i}{v_i}$$

donde V_i es el número de viviendas en el conglomerado i , v_i es el número de viviendas muestreadas en el conglomerado i .

13.4. Ajuste por no respuesta a nivel de vivienda

La tasa de no respuesta a nivel de vivienda, mide la proporción de viviendas seleccionadas que efectivamente participaron en la encuesta. El cálculo de la tasa de respuesta usa el siguiente término, derivado de la recolección de datos:

v_{iex} = número de viviendas seleccionados que fueron excluidos en el conglomerado c

La tasa de no respuesta viene dada por:

$$TNR_c = \frac{v_i - v_{iex}}{v_i}$$

Las ponderaciones de segunda etapa se calculan para los conglomerados seleccionados. En el caso de que existan viviendas excluidas de la investigación, se debe recalcular la ponderación de primera etapa para la j -ésima vivienda en el conglomerado i , queda definida por:

$$PF_v^{ij} = \frac{PB_v^{ij}}{TNR_i}$$

13.5. Ponderadores muestrales totales

El ponderador muestral total es el producto del ponderador final de primera etapa por el ponderador final de segunda etapa. Para la vivienda j investigada en el conglomerado i del estrato k , el ponderador muestral es:

$$P_T^j = PF_c^{ki} * PB_v^{ij}$$

El uso de ponderadores muestrales es crítico para obtener estimadores a partir de encuestas que usan técnicas distintas del muestreo aleatorio simple.

13.6. Factores de expansión ajustados (población)

Para concluir con los ajustes, a la ponderación calculada en el apartado anterior, se realiza un ajuste que toma en cuenta la proyección oficial de la población.

Este cálculo inicia luego de determinar la población expandida con este primer ajuste. El factor de ajuste por la proyección de la población se calcula de la siguiente manera:

$$\text{Ajuste proyección} = \frac{\text{Población proyección esperada para el periodo en estudio}}{\text{Población expandida de la encuesta}}$$

Este ajuste afecta a cada conglomerado i , por dominios de estimación.

13.7. Ajuste de los Factores de Expansión por Post-Estratificación

El factor de expansión, para un estudio basado en una muestra, es la cifra que actuando como un multiplicador, permite llevar los datos muestrales a la población (expandir la muestra).

La post-estratificación es un procedimiento similar a la estratificación del Marco Muestral, con la diferencia que la post-estratificación es realizada después de haber obtenido los resultados del estudio realizado. Como se señaló anteriormente, se estableció que la estratificación constituye una herramienta para la selección de la muestra para nuestro estudio y que ésta se realiza en las primeras etapas del diseño.

La post-estratificación se realiza una vez que se cuenta con los datos obtenidos en campo. Antes de realizar el estudio no se conocen con detalle muchas características de la población. Posteriormente, ya se cuenta con variables tales como la edad, sexo, entre otras; las mismas que permiten afinar y re-estratificar nuevos grupos con propósitos diversos, es decir, que para la post-estratificación se utilizan variables obtenidas en campo.

De igual manera que en la estratificación tradicional, la post-estratificación consiste en la creación de nuevos grupos de elementos denominados post-estratos y éstos poseen características homogéneas al interior y heterogéneas al exterior de los mismos, basados en variables investigadas en nuestro estudio.

Luego de la post-estratificación, cada uno de los elementos dentro de los datos muestrales será identificado como perteneciente a un post-estrato, según los criterios establecidos, que para el caso de la ENEMDU son la población por área, el dominio de estudio, el sexo y los grupos de edad.

El ajuste de los factores de expansión por post-estratificación consiste en la creación de un componente extra (δ)⁴ en la fórmula de cálculo, para ajustar la población estimada obtenida en el paso anterior, en subgrupos o post-estratos que para el caso de la ENEMDU son el sexo (hombre y mujer) y grupos de edad de interés (menores de

⁴ Este componente "delta" permitirá el ajuste de los factores de expansión a fin de que la Población expandida en el estudio coincida con el dato real, que en este caso es una proyección de la población a la fecha de ejecución de la encuesta.

15 años, 15 a 64 años, 65 años y más). Este ajuste se lo realiza utilizando como dato o parámetro conocido a la proyección de población para la fecha de realización de la encuesta, desagregada en los post-estratos construidos.

Este componente de ajuste resulta del cociente de la población proyectada entre la población expandida.

Para el cálculo del componente de ajuste se utilizó la siguiente fórmula:

$$\delta = \frac{U}{U_{hat}}$$

Donde:

U = Proyección al momento de la encuesta por post-estrato.

Uhat = Expansión de la población por medio de la encuesta por post-estrato.

Contando con los datos del número de personas proyectadas, el número de personas expandidas y el factor de expansión ajustado (población), se aplica la fórmula para cada uno de los post-estratos.

Con los valores obtenidos para los correspondientes componentes de ajuste (δ) para cada uno de los post-estratos, se proceda a multiplicar el factor originalmente calculado por el componente de ajuste, y el resultado obtenido es el nuevo factor de expansión ajustado o calibrado.

14. Estimación de características de la población

Una vez obtenidos los factores de expansión ajustados, se calcula los estimadores provenientes de la ENEMDU, se utiliza el estimador de Horvitz-Thompson, el cual sirve para estimar el valor total de una característica determinada (empleo adecuado, desempleo, etc). Está dado por (Carl-Erik Särndal, 1992).

$$\hat{Y}_{mHT} = \sum_{i=1}^{C_m} \sum_{j=1}^{v_i} P_T^j y_{ij},$$

Donde:

- \hat{Y}_{mHT} = *estimador insesgado para el total Y_k de una variable y en el dominio de estudio m*
- C_m = *número de conglomerados investigadas en el dominio de estudio m.*
- v_i = *número de viviendas investigadas en el conglomerado i del dominio de estudio m*
- y_{ij} = *valor de la variable y para la vivienda j del conglomerado i del dominio de estudio m*

15. El error de muestreo de diseño

Luego de realizar la estimación respectiva para la variable de interés a nivel de dominio de estudio, el error de muestreo podrá ser calculado a partir de la estimación de la varianza del estimador del total \hat{Y}_{mHT} . Para calcular adecuadamente los errores de muestreo de cada estimador, se debe tomar en cuenta los diferentes aspectos del

diseño muestral, es decir, las dos etapas de muestreo, la estratificación presente en los dominios de estudio y los procesos de selección en cada una de las etapas.

Con todos estos elementos, el coeficiente de variación para el estimador \hat{Y}_{kHT} viene dado por la siguiente expresión:

$$CV(\hat{Y}_{kHT}) = \frac{\sqrt{\hat{V}_{2st}(\hat{Y}_{kHT})}}{\hat{Y}_{kHT}}$$

Un estimador insesgado para la varianza está dado por:

$$\hat{V}_{2st}(\hat{t}_{\pi}) = \sum \sum_{s_I} \check{\Delta}_{Iij} \frac{\hat{t}_{i\pi} \hat{t}_{j\pi}}{\pi_{Ii} \pi_{Ij}} + \sum_{s_I} \frac{\hat{V}_i}{\pi_{Ii}}$$

En el cual el \hat{V}_i apropiado es

$$\hat{V}_i = \sum \sum_{s_{IIi}} \check{\Delta}_{IIqr|i} \frac{\hat{t}_{iq\pi} \hat{t}_{ir\pi}}{\pi_{IIq|i} \pi_{IIr|i}}$$

Dónde:

- i, j = Índice que recorre los conglomerados i, j en el dominio de estudio m .
- q, r = Índices que recorren las viviendas del conglomerado i en el dominio de estudio m .
- π_{Ii} = Probabilidad de selección de la Etapa I para el i – ésimo conglomerado en el dominio de estudio m .
- $\pi_{IIq|i}$ = Probabilidad de selección de la Etapa II para la q – ésima vivienda, dada el i – ésimo conglomerado.
- $\check{\Delta}_{Iij}$ = Cantidad Δ expandida asociada a los conglomerados i, j .
- $\check{\Delta}_{IIqr|i}$ = Cantidad Δ expandida asociada a las viviendas q, r dada la selección del i – ésimo conglomerado.

16. Métodos de estimación de errores para diseños muestrales complejos

Aunque la selección del diseño de muestreo y el estimador sean de libre elección para los investigadores, no lo es el cálculo de las medidas de confiabilidad y precisión. Dado que la base científica sobre la cual descansa el muestreo es la inferencia estadística, se deben respetar las normas básicas para la asignación y posterior cálculo del margen de error, que constituye una medida unificada del error total de muestreo el cual cuantifica la incertidumbre acerca de las estimaciones en una encuesta.

Los métodos de estimación de los errores muestrales pueden clasificarse en cuatro categorías:

- a) Métodos exactos
- b) Métodos del último conglomerado
- c) Aproximaciones por linealización
- d) Técnicas de replicación

Para la descripción de los métodos se ha tomado como referencia los textos de Kish y Frankel (1974), Wolter (1985) y Lehtonen y Pahkinen (1995), que se encuentran descritos en el documento **“ENEMDU: Cálculo de errores estándar y declaración de muestras complejas”**, donde se realiza una breve descripción de los métodos convencionales para estimar varianzas o errores muestrales para estimaciones basados en muestreo complejo, que es una característica de la ENEMDU.

A continuación se describirá las principales características de cada uno de los métodos de estimación de errores para el muestreo complejo:

- Los métodos exactos pueden ser utilizadas para estimar totales, medias, tamaños y proporciones.
- La linealización de Taylor debe ser utilizada para estimar parámetros no lineales como razones, medias dentro de dominios, cuartiles o funciones de distribución.
- La técnica del último conglomerado junto con la linealización de Taylor puede ser utilizada para estimar la varianza de los indicadores de interés de las encuestas dirigidas a hogares que tengan diseños muestrales complejos. Esta es la técnica que por defecto utiliza el software SPSS.
- Las técnicas de replicación pueden ser usadas para estimar eficientemente todos los parámetros de interés, sin importar su forma funcional.
- La comparación general entre los métodos de linealización y replicación, es que no generan resultados idénticos del error de muestreo, pero hay que señalar que existen estudios (Kish y Frankel, 1974) que concluyen que las diferencias presentadas no son significativas cuando se trata de grandes muestras.

El INEC utiliza para la estimación de los parámetros de interés y sus correspondientes errores de muestreo diversos programas estadísticos tales como SPSS, Stata y R. En virtud de las características de cada uno de los métodos, es la técnica del último conglomerado en combinación con la linealización de Taylor, la cual induce a una muy buena aproximación del error muestral sobre los indicadores más importantes de las encuestas dirigidas a hogares, además de su facilidad de cálculo y replica. En este sentido, será esta la técnica la utilizada para la estimación de los errores muestrales en la ENEMDU.

Las variables requeridas para declarar el diseño muestral en los programas estadísticos (SPSS, Stata y R) y ejecutar el cálculo de los errores de muestreo son presentadas en la Tabla 4, donde se describe las etiquetas de las variables identificadoras de los conglomerados, estratos y ponderación.

Tabla 4: Variables requeridas para declaración del diseño muestral - ENEMDU

Característica	Variable	Descripción
Conglomerados	upm	Agrupación de viviendas ocupadas en un número entre 30 a 60, próximas entre sí y con límites definidos.
Estratos	estrato	Identificación de estrato muestral (aproximación clasificación socio-económica)
Ponderación	fexp	Factor de expansión calculado y ajustado (no cobertura)

Es importante indicar que los estratos de muestreo están definidos por el cruce entre Provincia (25 grupos) + Área (2 grupos) + estrato socioeconómico de la UPM (3 grupos). Además, las UPM's deben tener identificadores únicos dentro de cada estrato y a través del tiempo. Por último, los hogares deben estar unívocamente identificados, así como su pertenencia a las UPM's, a los estratos de muestreo y a las rondas del panel correspondiente.

Bibliografía

Carl-Erik Särndal, B. S. (1992). *Model Assisted Survey Sampling* (1ra Edición ed.). Nueva York, Nueva York, USA: Springer-Verlag.

Educación, M. d. (2016). *Archivo Maestro de Instituciones Educativas*. Obtenido de <https://educacion.gob.ec/amie/>

Educación, M. d. (s.f.). *Archivo Maestro de Instituciones Educativas*.

Hidiroglu, M., Särndal, C., & Binder, D. (1995). Weighting and Estimation in Business Surveys. En Cox, Binder, Chinnappa, Christianson, Colledge, & Kott, *Business Survey Methods* (págs. 477-502). Jhon Wiley & Sons.

Peña, D. (2002). *ANÁLISIS DE DATOS MULTIVARIANTES*.

Polfeldt, T. (1 de Diciembre de 2003). *CEPAL*. Obtenido de Muestras con rotación de paneles: <http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=oahUKEwiL9eSoguLTAhUBPCYKHfHHAFwQFggsMAE&url=http%3A%2F%2Fwww.cepala.org%2Fdeype%2Fmecovi%2Fdocs%2FTALLER10%2F35.pdf&usg=AFQjCNF7nmrvIBeoGw3NDtiDI3wulQwvJg>

Kish, L. (1972). *Muestreo de Encuestas*. México: Trillas.

Clic para editar título

Clic para editar texto

**CADA
HECHO
DE TU
VIDA**
Cuenta

 @ecuadorencifras

 INEC/Ecuador

 @InecEcuador

 INECEcuador

 t.me/equadorencifras

 INEC Ecuador