Buenas cifras,
mejores vidas

ENEMDU: Cálculo de errores estándar y declaración de muestras complejas

Enero 2021

Marzo, 2021

INEC

Buenas cifras,
mejores vidas

01

Instituto Nacional de Estadística y Censos (INEC)

Dirección

Dirección de Infraestructura Estadística y Muestreo

Elaborado por:

William Constante
Javier Núñez
Jorge Velásquez

Revisado por:

Francisco Céspedes

Aprobado por:

David Sánchez

Quito -Ecuador, 2021

Contenido

ENEMDU: Cálculo de errores estándar y declaración de muestras complejas

Enero 2021

Introducción

La necesidad de analizar los datos de las encuestas dirigidas a hogares es cada vez más frecuente, donde los tabulados descriptivos básicos de los estadísticos de interés no son suficientes para emitir conclusiones con respaldo estadístico. En la actualidad, los usuarios y analistas de información buscan establecer si existen cambios entre dos indicadores en periodos comparables y por tanto generar pruebas de hipótesis y comprobación, o construir modelos a partir de los datos de la encuesta. Por ejemplo, en lugar de estimar simplemente la proporción de una población en situación de pobreza por ingresos, ahora los analistas requieren evaluar el impacto de la política pública o identificar cómo influyen otras variables en el fenómeno estudiado. Para responder a estas inquietudes se precisa de análisis detallados de los datos relativos al hogar o a la persona; por lo cual dichos análisis deben incluir necesariamente medidas apropiadas de la precisión o exactitud de las estimaciones derivadas de los datos de una encuesta que apoyen en la correcta interpretación de los resultados y para la evaluación y mejora de los diseños y procedimientos muestrales.

Por este motivo se hace indispensable que el usuario o analista de información haya revisado previamente la Metodología de Diseño Muestral de cada Operación Estadística y conozca en detalle el procedimiento aplicado, ya que para calcular los errores de muestreo se requiere de instrucciones que tengan en cuenta las características del diseño muestral con el que se generaron los datos, información que a su vez se debe declarar en los programas informáticos utilizados por los analistas de la información.

El presente documento entrega una breve perspectiva general teórica de los diversos métodos empleados para estimar los errores de muestreo y una aplicación práctica con la Encuesta Nacional de Empleo y Desempleo (ENEMDU), que repasa las principales características de su diseño muestral de la encuesta y presenta el cálculo de los errores muestrales con el apoyo de programas estadísticos especializados (SPSS, Stata y R).

Fundamentos teóricos

Diferencias entre el Muestreo Aleatorio Simple y el Muestreo Complejo

El muestreo aleatorio simple es la técnica de muestreo más básica, pero a su vez la menos empleada en las encuestas dirigidas a hogares, por las complicaciones que trae en su organización operativa y por los elevados costos (Hayes, 2008), sin embargo, es importante conocer este diseño debido a que es el fundamento teórico de los diseños muestrales[1] complejos.

La mayoría de los diseños muestrales empleados en las encuestas de hogares son complejos y se diferencian del muestreo aleatorio simple a causa de la presencia de una o más de las siguientes características (Naciones Unidas, 2009):

a) La estratificación en una o más etapas de muestreo; lo que permite formar grupos homogéneos de individuos, que a su vez son heterogéneos entre los diferentes grupos.

b) La conglomeración de las unidades estadísticas en una o más etapas de muestreo, lo que reduce los costos, pero aumenta la varianza de las estimaciones debido a las correlaciones entre las unidades del mismo conglomerado;

c) La ponderación para compensar imperfecciones de la muestra tales como las probabilidades desiguales de selección, la falta de respuesta o de cobertura.

Un error común cuando se trata de un muestreo complejo es suponer que las fórmulas del muestreo aleatorio simple se pueden usar para estimar las varianzas y por tanto los errores muestrales. De esta manera, al analizar los datos de una encuesta de hogares como si hubieran sido generados por un diseño de una muestra aleatoria simple induciría a errores en el análisis y en las conclusiones basadas en esos datos (Hayes, 2008).

Cualquier estimación de varianza que proviene de un diseño complejo siempre debe tener en cuenta su diseño de muestreo.

[1] Son los distintos procedimientos que existen para extraer muestras de poblaciones con el objeto de conocer sus características promedio.

Composición de la varianza en el muestreo

En el diseño de una encuesta a hogares solo se selecciona una muestra, y los valores de la población (parámetros) no se conocen. Ante esta situación, lo recomendable es usar procedimientos matemáticos para calcular la varianza.

La varianza de muestreo de una estimación puede definirse como la desviación cuadrática promedio del valor promedio de la estimación, donde el promedio se obtiene de todas las muestras posibles (Naciones Unidas, 2009).

Con el fin de describir la composición de la varianza, se tomará como ejemplo el muestreo aleatorio simple sin sustitución. La varianza de muestreo de una media estimada ($\hat{Y}$), basada en una muestra con tamaño n, viene dada por la expresión:

$$Var\left(\hat{Y}\right) = \left(1 - \frac{n}{N}\right)\frac{\delta^2}{n}$$

Donde $\delta^2 = \frac{\sum_{i=1}^{N}(Y_i - \hat{Y})^2}{N-1}$ es la media de la variabilidad de la característica de interés (varianza de la población de Y). Por lo general δ se desconoce y tiene que estimarse a partir de la muestra. De la fórmula anterior puede deducirse que la varianza de muestreo depende de los siguientes factores:

a) La varianza de la población de las características de interés;
b) El tamaño de la población;
c) El tamaño de la muestra;
d) El diseño de la muestra y el método de estimación.

La proporción de la población que está en la muestra $^{n}/_{N}$ se denomina la fracción de muestreo (designada como f); y el factor $\left[1 - \left(^{n}/_{N}\right)\right]$, o $1 - f$, que es la proporción de la población que no está incluida en la muestra, se denomina el factor de corrección de población finita (cpf). El cpf representa el ajuste efectuado en el error estándar de la estimación para tener en cuenta el hecho de que la muestra se selecciona sin sustitución a partir de una población finita. Sin embargo, cuando la fracción de muestreo es pequeña el cpf puede ignorarse. En la práctica, el cpf puede ignorarse si no supera el 5% (Cochran, 1997).

La fórmula anterior indica que la varianza de muestreo es inversamente proporcional al tamaño de la muestra. A medida que el tamaño de la muestra aumenta, la varianza de muestreo disminuye, situación que se debe tomar en cuenta para determinar la capacidad de inferencia de la estimación calculada.

Métodos de eliminación de la varianza en el muestreo complejo

En este apartado se realizará una breve descripción de los métodos convencionales para estimar varianzas o errores muestrales para estimaciones basados en muestreo complejo, que es una característica del diseño de las encuestas a hogares. 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), donde los usuarios podrán obtener mayor información sobre estos métodos.

Métodos exactos

Los métodos exactos de estimación de la varianza en diseños de muestras estándares, cuando son aplicables, son la mejor forma de estimar la varianza. Sin embargo, su puesta en marcha en el cálculo de varianzas de estimaciones basadas en muestreos complejos se dificulta por diversos factores.

- Los diseños muestrales empleados en la mayoría de las encuestas de hogares son más complejos que el muestreo aleatorio simple.

- Las estimaciones de interés podrían presentar características de no ser funciones lineales simples de los valores observados, por lo que la varianza de muestreo no siempre puede expresarse por una fórmula de forma cerrada, como sucede al calcular la media de la muestra en el muestreo aleatorio simple o el muestreo estratificado.

- Los métodos exactos dependen del diseño de la muestra, de la estimación de interés y de los procedimientos de ponderación empleados (Naciones Unidas, 2009).

Con la finalidad de compensar esta deficiencia de los métodos se recomienda la aplicación de diferentes métodos, para la estimación de la varianza aplicable en el muestreo complejo de las encuestas a hogares. Estos métodos se analizarán a continuación.

La técnica del último conglomerado

El método del último conglomerado para estimar la varianza (Hansen, Hurwitz y Madow, 1953, págs. 257-259), constituye una aproximación para el cálculo de la varianza basadas en una muestra obtenida a partir de un diseño muestral

complejo. Las estimaciones de la varianza se calculan utilizando solo totales entre Unidades Primarias de Muestreo (UPM), sin tener que calcular los componentes de la varianza en cada etapa de selección, es decir, se considera únicamente la varianza de los estimadores en la primera etapa, y supone que el muestreo en esa etapa fue realizado con reemplazo. Los procedimientos de muestreo en etapas posteriores de la selección se excluyen del cálculo[2] (Naciones Unidas, 2009).

Supongamos que seleccionamos una muestra de n_h UPM del estrato h (con un número cualquiera de etapas dentro de las UPM). En ese caso la estimación del total para el estrato h está dado por:

$$\hat{Y}_h = \sum_{i=1}^{n_h} \hat{Y}_{hi}, \text{ donde } \hat{Y}_{hi} = \sum_{j=1}^{m_i} W_{hijk} Y_{hijk}$$

La estimación $\hat{Y}_{hi}$ al nivel de las UPM es una estimación de $\frac{\hat{Y}_h}{n_h}$. Por tanto, la varianza de las estimaciones individuales a nivel de las UPM viene dada por:

$$v(\hat{Y}_{hi}) = \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi} - \frac{\hat{Y}_h}{n_h} \right)^2$$

La varianza total de $\hat{Y}_h$, el total a nivel de estrato, estimada a partir de una muestra aleatoria de tamaño n_h como estimador del total de la población en el estrato h, viene dada por:

$$v(\hat{Y}_h) = \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi} - \frac{\hat{Y}_h}{n_h} \right)^2$$

Por último, mediante muestreo independiente en los estratos, el estimador de la varianza para el total de toda la población en muestreo complejo se obtiene sumando las varianzas de los totales de cada estrato, es decir:

$$v(\hat{Y}) = \sum_{h=1}^{H} v(\hat{Y}_h)$$

Este procedimiento, propuesto por Hansen, Hurwitz, y Madow (1953), tiende a sobrestimar levemente la varianza verdadera debido a que no contempla todas las etapas del muestreo complejo. Sin embargo, la principal ventaja es que utiliza directamente los factores de expansión que son publicados por los institutos nacionales de estadística.

Aproximación por linealización

La mayoría de las estimaciones de interés en las encuestas de hogares no son lineales. Algunos ejemplos son razones, medias dentro de dominios, cuartiles o funciones de distribución etc. La aproximación por linealización se basa en expresar el estimador como función de estimadores lineales de totales. Esto

[2] Por lo general las encuestas dirigidas a hogares tienen un diseño muestral complejo constituido por dos o más etapas de muestreo.

conlleva expresar la estimación en función de una expansión en serie de Taylor y posteriormente aproximar la varianza de la estimación mediante la varianza de primer orden o parte lineal de dicha expansión aplicando los métodos exactos descritos en apartados anteriores (Naciones Unidas, 2009).

Supongamos que queremos estimar la varianza de una estimación z de un parámetro Z y a la vez siendo z una función no lineal de las estimaciones simples $y_1, y_2, \dots, y_m$ de los parámetros $Y_1, Y_2, \dots, Y_m$

$$z = f(y_1, y_2, \dots, y_m) \ (1)$$

Asumiendo que z se aproxima a Z, la expansión en serie de Taylor de z a términos de primer grado de $z - Z$ equivale a:

$$z = Z + \sum_{i=1}^{m} d_i \, (y_i - Y_i)$$

Donde d_i' son las derivadas parciales de z respecto a y_i', es decir,

$$d_i = \frac{\partial z}{\partial y_i}$$

Que es una función de la estimación básica y_i. Esto significa que la varianza de z puede aproximarse mediante la varianza de la función lineal de la ecuación (1), que se calculan utilizando el método exacto descrito anteriormente:

$$v(z) = v\left(\sum d_i y_i\right) = \sum_{i=1}^{m} d_1^2 \, v(y_i) + \sum_{i \neq j} d_i d_j cov(y_i, y_j) \ (2)$$

La ecuación (2) incluye una matriz $(m * m)$ de covarianza de m estimaciones básicas $y_1, y_2, \dots, y_m$, con m términos de varianza y $m(m-1)/2$ términos idénticos de covarianza, que puede evaluarse a partir de los métodos exactos para estadísticos lineales.

La linealización es ampliamente utilizada debido a que puede aplicarse a casi todos los diseños muestrales y a cualquier estadística que pueda linealizarse, o, lo que es lo mismo, expresarse como una función lineal de las estadísticas comunes como medias o totales, cuyos coeficientes se extraen de derivadas parciales necesarias para la expansión en serie de Taylor. Una vez linealizada, la varianza de la estimación no lineal puede aproximarse mediante los métodos exactos, los cuales ya se amplió en el apartado 3.1.1. (Cochran, 1977, y Lohr, 1999).

Técnicas de replicación

El método de replicación comprende una clase de métodos que consisten en tomar repetidas submuestras, o réplicas, de los datos, recalcular la estimación ponderada en cada replica y en la muestra completa y luego calcular la

varianza como una función de las desviaciones de estas estimaciones replicadas de la estimación de la muestra total (Naciones Unidas, 2009).

La principal ventaja del método de replicación respecto al de linealización es que el método de estimación básico empleado es el mismo, sea cual sea la estadística estimada (porque la estimación de la varianza es una función de la muestra, no de la estimación), mientras que en el caso de la linealización debe desarrollarse analíticamente una aproximación para cada estadística, lo cual puede convertirse en una laboriosa tarea en las grandes encuestas de hogares con un elevado número de características de interés. El uso de las técnicas de replicación resulta práctico también y es aplicable a casi todas las estadísticas, lineales y no lineales.

Las técnicas de replicación más utilizadas son (Naciones Unidas 2009):

- Los grupos aleatorios;
- La replicación repetida equilibrada;
- La replicación Jackknife;
- El bootstrap.

A continuación, se describirá un resumen de los métodos de estimación de la varianza para el muestreo complejo y sus principales características:

- 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[3].
- 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.

Cálculo de los errores muestrales en la ENEMDU

Como se mencionó en los apartados anteriores, es fundamental realizar una revisión previa del diseño muestral aplicado en la operación estadística, ya que esta información es necesaria para realizar un correcto cálculo de las varianzas

[3] Esta es la técnica que por defecto utiliza el software SPSS.

y errores muestrales. Para el caso de la ENEMDU se empezará detallando las principales características de su diseño muestral.

Diseño de la ENEMDU

El diseño muestral de la Encuesta de Empleo y Desempleo (ENEMDU) responde a un diseño muestral complejo en virtud de que cumple con las siguientes características:

a) Es un muestreo estratificado en cada una de sus etapas;
b) Las Unidades Primarias de Muestreo (UPM) son agrupaciones de viviendas, conocidos como "conglomerados de viviendas";
c) Presenta un factor de expansión o ponderación el cual corrige, la falta de respuesta o de cobertura que se presenta en la operación estadística

La representatividad de la ENEMDU que se realiza cada mes a partir del mes de julio del año 2020 es nacional, urbano, rural, considerando que se presenta un tamaño de muestra mensual igual, tanto en número de UPM como de viviendas[4]. La selección muestral de la ENEMDU se realizó considerando el siguiente orden:

- **División política**: Ecuador está dividido políticamente en jurisdicciones (provincias, cantones y parroquias) las cuales se convierten en estratos territoriales.
- **Área**: Cada estrato territorial a la vez se subdivide en área urbana y área rural.
- **Unidades Primarias de Muestreo (UPM)**: Los estratos territoriales por área se encuentran divididos en Unidades Primarias de Muestreo (UPM); las cuales son la agrupación de viviendas (entre 30 y 60 viviendas ocupadas) que comparten características comunes para pertenecer a un mismo nivel socio-económico. La selección de UPM por territorio se la realiza en función del número total de viviendas por nivel socio económico aplicando un Muestreo Aleatorio Simple (MAS), ya que al tener UPM equilibradas en tamaño de viviendas es posible aplicar este método de selección.
- **Viviendas**: Las viviendas dentro de las UPM son seleccionadas de forma aleatoria en un número de diez viviendas ocupadas, donde siete son consideradas como originales y tres como reemplazos.
- **Hogares**: Todos los hogares dentro de una vivienda seleccionada son levantados; y
- **Personas**: Todas las personas fueron seleccionadas en cada hogar.

[4] Para una mayor información del tamaño de muestra, referirse al documento metodológico del diseño muestral de la ENEMDU.

Programas informáticos para la estimación del error de muestreo

Para la estimación de los parámetros de interés y sus correspondientes errores de muestreo, el INEC utiliza diversos programas estadísticos tales como SPSS, Stata y R. En virtud de lo descrito en los apartados anteriores la técnica del último conglomerado en combinación con la linealización de Taylor, 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 réplica. En este sentido, será esta la técnica la que se aplicará y ejemplificará en los diferentes programas estadísticos.

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 1, donde se describe las etiquetas de las variables identificadoras de los conglomerados, estratos y ponderación.

Tabla 1. Variables requeridas para declaración del diseño muestral - ENEMDU

Característica	*Variable*	*Descripción*
Unidad Primeria de Muestreo	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)

A continuación, se describe el cálculo de los errores estándar y varianzas para muestras complejas en los paquetes estadísticos SPSS, Stata y R con un ejemplo de estimación y su sintaxis o código utilizado en cada programa. El ejemplo a utilizar es la estimación promedio del ingreso per cápita y tasa de desempleo para junio 2018. De requerir más información sobre los códigos utilizados, dirigirse al espacio de ayuda de cada uno de los programas.

SPSS

SPSS tiene el módulo adicional "SPSS Complex Samples", que incluye un conjunto de procedimientos que ofrecen la posibilidad de analizar una muestra compleja:

- Plan de muestras complejas (CSPLAN) para especificar un esquema de muestreo y definir el archivo del plan utilizado por los siguientes procedimientos.
- Descriptivos de muestras complejas (CSDESCRIPTIVES) para estimar medias, sumas y razones de variables, cálculos de errores estándar,

efectos de diseño, intervalos de confianza y pruebas de hipótesis para el diseño muestral planteado.

- Tabulación de muestras complejas (CSTABULATE) para mostrar la frecuencia de tablas unidireccionales o tabulaciones cruzadas bidireccionales, relacionadas con la descripción antes mencionada, estas pueden ser solicitadas por subgrupos.
- GLM de muestras complejas (CSGLM) para ejecutar análisis de regresión lineal, y análisis de varianza y covarianza.
- Logística de muestras complejas (CSLOGISTIC) para ejecutar la regresión logística en el análisis en una variable dependiente binaria o multinomial usando el enlace generalizado de función.
- Ordinal de muestras complejas (CSORDINAL) para ajustar un modelo de probabilidades acumulado a una variable dependiente ordinal para los datos que se han recopilado de acuerdo con un diseño de muestreo complejo.

Una vez especificado, el módulo calcula errores estándar a través del método de último conglomerado aplicando la linealización de la serie Taylor. Actualmente no hay ningún módulo disponible en SPSS para manejar automáticamente la técnica de réplicas, aunque el lenguaje de programación SPSS podría permitir la generación de una sintaxis que permita la réplica. Los ejemplos de código proporcionados a continuación se pueden replicar en la interfaz de sintaxis para SPSS.

Ejemplos

Especifique las características de diseño de la muestra (este paso es necesario para cualquiera de los siguientes ejemplos para trabajar).

```
CSPLAN ANALYSIS
  /PLAN FILE='Ubicación del plan de muestreo'
  /PLANVARS ANALYSISWEIGHT=fexp
  /SRSESTIMATOR TYPE=WR
  /PRINT PLAN
  /DESIGN STRATA=estratos CLUSTER=upm
  /ESTIMATOR TYPE=WR.
```

Resumen

			Etapa 1
Variables del diseño	Estratificación	1	Estratos
	Conglomerado	1	Identificador de unidad primaria de muestreo
Información sobre el análisis	Supuestos del estimador		Muestreo con reposición

Variable de ponderación: Factor de expansión
Estimador SRS: Muestreo con reposición

Calcule la estimación de la media del ingreso per cápita, con los errores estándar de la serie Taylor:

```
CSDESCRIPTIVES
/plan file='Ubicación del plan de muestreo'
/summary variables=ingpc
/mean.
```

Estadísticos univariantes

		Estimación	Error típico
Media	Ingreso per cápita	222,51	4,040

Para dividir la media en subgrupos (área, por ejemplo) agregue la siguiente línea al procedimiento anterior.

```
CSDESCRIPTIVES
/plan file='Ubicación del plan de muestreo'
/summary variables= ingpc
/mean
/subpop table=area.
```

Estadísticos univariantes

AREA			Estimación	Error típico
Urbano	Media	Ingreso per cápita	262,01	5,353
Rural	Media	Ingreso per cápita	138,23	3,327

Genere una tabla de frecuencias, con errores estándar, del ingreso per cápita por área con su intervalo de confianza, coeficiente de variación y número de observaciones muestrales.

```
CSDESCRIPTIVES
  /PLAN FILE='Ubicación del plan de muestreo'
  /SUMMARY VARIABLES=ingpc
  /SUBPOP TABLE=area DISPLAY=LAYERED
  /MEAN
  /STATISTICS SE CV COUNT CIN(95)
  /MISSING SCOPE=ANALYSIS CLASSMISSING=EXCLUDE.
```

Estadísticos univariantes

AREA			Estimación	Error típico	Intervalo de confianza al 95%		Coeficiente de variación	Recuento no ponderado
					Inferior	Superior		
Urbano	Media	Ingreso per cápita	262,01	5,353	251,51	272,51	,020	36863
Rural	Media	Ingreso per cápita	138,23	3,327	131,70	144,75	,024	22592

Genere una estimación de la tasa de desempleo, con los errores estándar de la serie Taylor por área con su intervalo de confianza, coeficiente de variación y número de observaciones.

```
IF (pean=1) tdesem=0.
IF (desem=1) tdesem=1.
CSDESCRIPTIVES
  /PLAN FILE='Ubicación del plan de muestreo'
```

```
/RATIO NUMERATOR=tdesem DENOMINATOR=pean
/STATISTICS SE CV COUNT CIN(95)
/SUBPOP TABLE=area DISPLAY=LAYERED
/MISSING SCOPE=ANALYSIS CLASSMISSING=EXCLUDE.
```

Razones 1

AREA	Numerador	Denominador	Estimación de la razón	Error típico	Intervalo de confianza al 95%		Coeficiente de variación	Recuento no ponderado
					Inferior	Superior		
Urbano	tdesem	Población Económicamente Activa (PEA)	,052	,002	,047	,057	,047	17798
Rural	tdesem	Población Económicamente Activa (PEA)	,020	,002	,016	,024	,104	11972

STATA

Stata permite realizar estimaciones de acuerdo a las técnicas anteriormente descritas, pero para el caso de este programa se utilizará la técnica del último conglomerado en combinación con la linealización de la serie Taylor.

Ejemplos

Como primer paso se declara en Stata las características del diseño muestral mediante el comando svyset, de tal forma que, posteriormente, al usar un comando con el prefijo svy: sus errores estándares sean calculados con Taylor tomando en cuenta todas las características del diseño muestral, donde el identificador de cada UPM es la variable upm, el factor de expansión corresponde a la variable fexp; la variable que identifica los estratos es estratos y el método de cálculo de los errores estándar es mediante la linealización de Taylor con la opción linearized. Finalmente, la opción singleunit (certainty) especifica cómo manejar los estratos con una unidad de muestreo. "certainty" hace que los estratos con conglomerados individuales sean tratados como unidades de certeza, es decir estas unidades no contribuyen para el cálculo del error estándar.

```
svyset upm [iw=fexp], strata (estrato) vce(linearized) singleunit(certainty)

iweight: fexp
VCE:    linearized
Single unit:      certainty
Strata 1:         estratos
SU 1:    upm
FPC 1:  <zero>
```

Por ejemplo, si se realiza una estimación de la media del ingreso per cápita, con los errores estándar Linealizados de Taylor, el proceso es el siguiente:

```
svy: mean ingpc
```

	Mean	Linearized Std. Err.	[95% Conf.	Interval]
ingpc	222.509	4.040459	214.5856	230.4323

Para dividir la media en subgrupos (área, por ejemplo) agregue el siguiente comando al procedimiento anterior.

```
svy:  mean ingpc , over (area)
```

Over	Mean	Linearized Std. Err.	[95% Conf.	Interval]
ingpc				
Urbano	262.0088	5.353414	251.5107	272.5069
Rural	138.2265	3.326996	131.7023	144.7508

Para producir una tabla de frecuencias, con errores estándar, del ingreso per cápita por área con su coeficiente de variación.

```
svy:  mean ingpc, over (area)
estat cv
```

Over	Mean	Linearized Std. Err.	CV (%)
ingpc			
Urbano	262.0088	5.353414	2.04322
Rural	138.2265	3.326996	2.40692

De igual forma, si se desea realizar una estimación de la tasa de desempleo, con los errores estándar de la serie Taylor por área con su intervalo de confianza y luego con su coeficiente de variación, el código se incluye a continuación.

```
gen ntdesem=0
replace ntdesem=1 if desem==1
gen dtdesem=0
replace dtdesem=1 if pean==1
svy: ratio ntdesem/dtdesem, over (area)
estat cv
```

Over	Ratio	Linearized Std. Err.	[95% Conf.	Interval]
_ratio_1				
Urbano	.0521242	.002432	.0473552	.0568933
Rural	.0202061	.0021109	.0160667	.0243456

Over	Ratio	Linearized Std. Err.	CV (%)
_ratio_1			
Urbano	.0521242	.002432	4.66568
Rural	.0202061	.0021109	10.4467

R

Una de las posibilidades que presenta R en el análisis de muestras de diseño complejo es el paquete survey. Esta herramienta posibilita la inclusión del tipo de ponderación adecuado (sampling weights, precision weights or frequency weights), unidades primarias de muestreo, estratos, efectos de diseño y demás características intrínsecas a varios tipos de muestreo. Este paquete estadístico permite el análisis de datos obtenidos mediante muestreo aleatorio simple o estratificado, así como por conglomerados, incluyendo diseños multietápicos o de una sola etapa, y diseños complejos.

Se empezará por solicitar las librerías necesarias:

```
rm(list = ls())
library(foreign)
library(survey)
library(dplyr)
library(srvyr)
library(stringr)
```

Para el ejemplo se utilizará la base de datos en formato SPSS, en donde también de definirá las variables a utilizar para el ejemplo.

```
setwd("Ubicación de carpeta de trabajo")
#Lectura de la base en formato CSV
enemdu <- read.spss("C:/Users/cgarces/Desktop/Metodología ENEMDU/Errores
estándar ENEMDU/Ejemplo/201806_EnemduBDD_15anios.sav",
                    to.data.frame = T, use.value.labels = F)
#Recodificación de variables
enemdu <- mutate(enemdu, desem_rc = ifelse(is.na(desem) & pean==1,0,desem))
```

En el paso siguiente se definirá el diseño muestral para la ENEMDU.

```
d1 <- enemdu %>% as_survey_design(ids = upm,
                                  strata = estratos,
                                  weights = fexp,
                                  nest = T)
options(survey.lonely.psu = "certainty")
```

Produzca una estimación de la media del ingreso per cápita, con los errores estándar de la serie Taylor:

```
tabla1<-d1 %>% summarise(ingperc = survey_mean(ingpc, vartype=c("se"), na.rm=T))
View(tabla1)
```

Tabla 1.	media	error estándar
Ingreso per cápita	222,509	4,040459

Para estimar el ingreso per cápita por área y su error estándar de la serie de Taylor, siga el siguiente procedimiento:

```
tabla2<-d1 %>% group_by(area) %>% summarise(ingperc = survey_mean(ingpc,
vartype=c("se"), na.rm=T))
View(tabla2)
```

Tabla 2.	área	media	error estándar
Ingreso per cápita	urbano	262,0088	5,353414
	rural	138,2265	3,326996

Para generar una tabla de frecuencias, con errores estándar, del ingreso per cápita por área con su intervalo de confianza, coeficiente de variación y efecto de diseño, utilice el siguiente código.

```
tabla3<-d1 %>% group_by(area) %>% summarise(ingperc = survey_mean(ingpc,
vartype=c("se","ci","cv"), na.rm=T, deff = T))
View(tabla3)
```

Tabla 3.	área	media	error estándar	lim. Inf	lim. sup	CV
Ingreso per cápita	urbano	262,0088	5,353414	251,5107	272,5069	0,02043219
	rural	138,2265	3,326996	131,7023	144,7508	0,02406916

Para el cálculo de una estimación de la tasa de desempleo, con los errores estándar de la serie Taylor por área con su intervalo de confianza y coeficiente de variación.

```
tabla4 <- d1 %>% group_by(area) %>% summarise(Desempleo =
survey_ratio(desem_rc, pean, vartype=c("se","ci","cv"), na.rm = T, deff = T))
View(tabla4)
```

Tabla 4.	área	ratio	error estándar	lim. Inf	lim. sup	CV
Ingreso per cápita	urbano	0,052	0,00243	0,04736	0,05689	0,04666
	rural	0,020	0,00211	0,01607	0,02435	0,10447

Referencias

Cochran, W. G. (1977). *Sampling Techniques*, 3a. ed. Nueva York: Wiley.

Hansen, M., W. Hurwitz y W. Madow (1953). *Sample Survey Methods and Theory*. Nueva York: Wiley.

Hayes, Clinton (2008). "*HILDA Standard Errors: A Users Guide*" The University of Melbourne - The HILDA Project was initiated, and is funded, by the Australian Government Department of Families, Housing, Community Services and Indigenous Affairs. Melbourne.

Kish, L. y M. R. Frankel (1974). "*Inference from complex samples*". Journal of the Royal Statistical Society: Services B, vol. 36, pags. 1-37.

Lehtonen, R. y E. J. Pahkinen (1995). *Practical Methods for Design and Analysis of Complex Surveys*. Nueva York: Wiley.

Lohr, Sharon (1999). *Sampling: Design and Analysis*. Pacific Grove, California: Duxbury Press.

Naciones Unidas (2009). "*Diseño de muestras para encuestas a hogares: directrices prácticas*" (ST/ESA/STAT/SER.F/98). División de Estadística, Departamento de Asuntos Económicos Y Sociales. Nueva York.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. Nueva York: Springer Verlag.

INEC

Buenas cifras,
mejores vidas

@ecuadorencifras @ecuadorencifras @InecEcuador t.me/ecuadorencifras INEC/Ecuador INECEcuador INEC Ecuador