Man examining house under microscope /Veronica Grech/Getty Images

# *Una aproximación metodológica al uso de datos de encuestas en hogares*

# A Methodological Approach to Household-Surveys Data Usage

**Julio César Martínez Sánchez***

**Nota:** el autor agradece los valiosos comentarios y sugerencias de Juan Trejo, César Villagrán, Byanka Barbosa y Sandra Ruiz.

* Instituto Nacional de Estadística y Geografía (INEGI), jcms2665@gmail.com y cesar.martinez@inegi.org.mx

Muchas de las encuestas que se usan para el análisis sociodemográfico tienen un esquema de muestreo complejo, lo cual significa que las observaciones no son por completo independientes ni tienen la misma probabilidad de ser seleccionadas. El objetivo de este artículo es mostrar el efecto que tiene dicho esquema de muestreo al momento de analizar e interpretar los datos de las encuestas. Para lograrlo, se ofrece un panorama sobre el diseño estadístico de las encuestas; luego, una breve exposición de los principales métodos para el cálculo de la varianza y, finalmente, se presenta un ejemplo con los datos de la Encuesta Nacional de Ocupación y Empleo utilizando la linealización por series de Taylor y *Bootstrap*. Los resultados muestran que, al incorporar el esquema de muestreo en el análisis, algunas de las estimaciones no son representativas de la población objetivo, mientras que en los modelos de regresión lineal el efecto se tiene en dos sentidos: se modifican los coeficientes, o bien, existe un cambio en su nivel de significancia.

**Palabras clave:** encuestas complejas; cálculo de varianza; linealización por series de Taylor; *Bootstrap*.

*Many of the surveys that are used in sociodemographic analysis have a complex sampling scheme. This means that observations are not completely independent, and that they do not have the same probability of being selected. In this paper, we explore the effect of this sampling scheme when analyzing and interpreting surveys' data. In order to do this, we offer an overview of the sampling plans, and then we explain the main methods for variance estimation. Finally, we use Taylor series linearization and Bootstrap to analyze data derived from the Occupation and Employment National Survey (ENOE in Spanish). The results suggest that there are some parameters that are not representative of the target population. Moreover, the sampling plan has an impact on linear regression models in two ways: either it changes the coefficients, or their significance level.*

***Key words:** complex survey; variance estimation; Taylor Series; Bootstrap.*

**Recibido:** 19 de agosto de 2016
**Aceptado:** 15 de noviembre de 2016

## Introducción

Muchos de los análisis sociodemográficos usan como fuente de información primaria las encuestas, las cuales, a diferencia de los registros administrativos o los censos de población y vivienda, se caracterizan por levantarse con mayor frecuencia e investigar temas específicos con mayor profundidad; sin embargo, las encuestas son métodos estadísticos que tienen ciertas características, las cuales deben ser tomadas en cuenta al momento de examinar los datos, ya que pasarlas por alto puede llevar a inferencias erróneas.

El objetivo de este artículo es mostrar el efecto que tiene el esquema de muestreo tanto en las estimaciones[1] como en los modelos de regresión lineal. Para lograrlo, se consideran tres etapas: primero se presenta un panorama sobre el diseño estadístico de las encuestas, luego se ofrece una breve exposición de los principales métodos para el cálculo de la varianza y, por último, se expone un ejemplo de cómo analizar los datos de la Encuesta Nacional de Ocupación y Empleo (ENOE) considerando su esquema de muestreo mediante la linealización por series de Taylor y *Bootstrap*.

A lo largo del documento se citan ejemplos y procedimientos que se siguen en el Instituto Nacional de Estadística y Geografía (INEGI), los cuales se ajustan a las normas y recomendaciones internacionales. Para llevar a cabo el análisis cuantitativo, se utiliza *R* versión 3.3.1 y el paquete *survey*; además, el código se encuentra disponible en la plataforma de *Rpubs*[2] por lo que puede ser consultado.

## 1. Diseño estadístico de las encuestas en hogares

### 1.1 Tamaño de muestra

La metodología que siguen las encuestas inicia con la selección de un subconjunto del universo de estudio[3] al cual se denomina *muestra*; su diseño es de suma importancia, pues debe mantener las mismas características que la población objetivo, pero con un menor número de observaciones. A la proporción que forma parte de dicha muestra se le denomina *fracción de muestreo* ($f$), y se estima dividiendo el número de elementos seleccionados entre el tamaño del universo de estudio.[4] En el escenario ideal, este coeficiente tiene un valor cercano a 1, lo cual significa que el número de elementos que están en la muestra es muy parecido al tamaño real del universo de estudio (Heeringa *et al*., 2010; Chambers y Skinner, 2003; INEGI, 2011a).

1 Método para calcular el parámetro de la población (promedios, tasas, totales) a partir de los datos de una encuesta (INEGI, 2011a).
2 https://rpubs.com/jcms2665/MAHSDU
3 Para tener un panorama de los términos usados en muestreo, se recomienda consultar Naciones Unidas (2009) e INEGI (2011a) donde se ofrecen algunas definiciones que son muy útiles.
4 $f=\frac{n}{N}$, donde $f$ es la fracción de muestreo, $n$ son los elementos seleccionados y $N$, el tamaño de la población objetivo.

Sin embargo, el foco de atención se encuentra en aquellas observaciones que no fueron seleccionadas ya que, dependiendo de su magnitud, pueden afectar las estimaciones (Heeringa *et al*., 2010; Chambers y Skinner, 2003). Esta proporción se conoce como *factor de corrección por población finita* ($cpf$)[5] y muestra el efecto de considerar solo una fracción (muestra) de la población objetivo. Cuando la muestra seleccionada es muy grande, el factor se reduce y se puede ignorar si es menor a 5% (Cochran, 1977); esto es visible en los censos de población y vivienda, donde no existe corrección por población finita ($cpf = 0$) ya que el tamaño de muestra es igual al universo de estudio.

Ahora bien, para estimar el tamaño de muestra (el cual se suele denotar como $n$), es necesario apoyarse en el diseño conceptual y determinar cuál es el indicador de mayor relevancia de toda la encuesta (Naciones Unidas, 2009; INEGI, 2011a). A partir de éste, se hace el cálculo correspondiente incorporando otros factores,[6] como: la tasa de no respuesta ($TNR$), nivel de confianza ($z$), error relativo máximo ($r$), coeficiente de variación ($cv$), efecto de diseño ($deff$). Por ejemplo, en la Encuesta Intercensal (EI) 2015, el tamaño de muestra se calculó con base en el indicador del *total de población residente en viviendas particulares habitadas* ($pov$)[7] cuya fórmula es la siguiente (INEGI, 2015a: 68):

$$n = \frac{z^2 * cv^2}{r^2} * \frac{deff}{(1-TNR)*pov} \tag{1}$$

Otros factores que se toman en cuenta al momento de definir el número de observaciones son los objetivos de la encuesta, la cobertura geográfica y la disponibilidad de recursos económicos (Naciones Unidas, 2007). La intención es incorporar todos estos elementos para que el tamaño de muestra ($n$) garantice un nivel de precisión adecuado a un costo razonable (Naciones Unidas, 2009). Luego de que se ha hecho esta valoración, el siguiente paso es definir el esquema de muestreo, en otras palabras, elegir la forma en que se van a colectar los datos.

## 1.2. Selección de la muestra

Existen diferentes técnicas para seleccionar una muestra y se dividen en dos grandes grupos: las que se realizan de forma determinística y aquellas que siguen un esquema probabilístico (ver cuadro 1). Cuando se trata del primer caso, las observaciones son seleccionadas de acuerdo con el criterio del investigador, o bien, se privilegia la conveniencia de elegir ciertos casos de acuerdo con el tema de investigación (INEGI, 2011a). En el segundo caso la situación es diferente, ya que se parte del supuesto de que cualquier elemento de la población objetivo tiene una cierta probabilidad de selección, por lo que se establecen ciertos criterios para determinar qué elementos formarán parte de la muestra (Hansen, 1953; Heeringa *et al*., 2010; Mecatti *et al*., 2014).

5 $cpf = 1 - f$

6 El valor de estos factores se toma de experiencias previas o de otras encuestas que tengan una temática similar y, en el peor de los casos, se usan las cifras de otros países que tengan encuestas similares (INEGI, 2007).

7 Aquí otros ejemplos: Encuesta Nacional sobre Uso del Tiempo (ENUT), *proporción de personas que construyeron o hicieron una ampliación a su vivienda*; Módulo de Condiciones Socioeconómicas (MCS), *promedio del ingreso corriente trimestral por hogar*; Encuesta Nacional de la Dinámica Demográfica (ENADID), tasa de fecundidad (INEGI, 2015b; INEGI, 2015c; INEGI, 2015d).

Cuadro 1

**Tipos de muestreo**

- Tipo de muestreo
  - Probabilístico
    - Aleatorio simple
    - Sistemático
    - Por conglomerados
    - Estratificado
  - Determinístico
    - Convencional o accidental
    - Por cuotas
    - Por juicio

**Fuente:** elaboración propia con base en Heeringa et al. (2010), INEGI (2011a) y Naciones Unidas (2009).

La ventaja que tiene este último procedimiento sobre el primero es que las conclusiones son válidas para todo el universo de estudio y no solo para aquellos casos que fueron seleccionados. Por esta razón, en las encuestas en hogares en las que se pretende analizar el comportamiento de la población, lo más recomendable es optar por un muestreo probabilístico; con ello se garantiza que la muestra tenga todas las características del universo de estudio y las inferencias sean válidas (Naciones Unidas, 2009; INEGI, 2011a).

### 1.2.1 Muestreo aleatorio simple

Existen diferentes métodos de selección dentro del muestreo probabilístico, el más recomendable —y al mismo tiempo difícil de implementar— es el *muestreo aleatorio simple* (mas). Su virtud es que genera observaciones independientes e idénticamente distribuidas; esto significa que todos los elementos de la población objetivo tienen la misma probabilidad de ser seleccionados (Naciones Unidas, 2009; Heeringa *et al*., 2010). Esta cualidad produce estimaciones sin sesgo, por lo que es el escenario ideal para llevar a cabo cualquier tipo de análisis estadístico (Hahs-Vaughn *et al*., 2011).

Sin embargo, a pesar de las ventajas que ofrece este tipo de muestreo, es muy difícil de implementar en proyectos de gran escala ya que se necesita conocer a todos los elementos de la población y hacer una selección aleatoria. Si se aplicara esta técnica en las encuestas en hogares se tendría que visitar a todas las viviendas del país para listarlas, elegir algunas al azar e ir a visitarlas para aplicarles la entrevista. Este procedimiento se tendría que repetir en cada encuesta, lo que significaría una inversión económica muy grande. Ante esta situación, se han desarrollado otros tipos de muestreo donde se busca un equilibrio entre precisión estadística y costo (Naciones Unidas, 2009).

### 1.2.2 Muestreo por conglomerados

Es una técnica que consiste en hacer grupos homogéneos donde sus elementos son muy similares al interior y heterogéneos al exterior (ver figura 1). Estos grupos se pueden formar en una o más etapas y, de acuerdo con ello, adquieren su nombre: aquellos que se generan en la primera etapa reciben el de unidades primarias de muestreo (UPM); luego, están las unidades secundarias de muestreo (USM) y así, sucesivamente (Lumley, 2004; Harden, 2011). En el caso del INEGI, dichas UPM (o conglomerados) se forman agrupando viviendas que se encuentran en una misma zona, y el número óptimo depende del tamaño de cada localidad y el total de viviendas habitadas y deshabitadas (Landeros, 2013; INEGI, 2009); por ejemplo, en la ENOE, las localidades de 100 mil y más habitantes consideran como rango aceptable entre 80 y 160 viviendas en cada UPM, mientras que para las localidades con menos de 100 mil habitantes el intervalo es de 160 a 300 viviendas (INEGI, 2007:44).

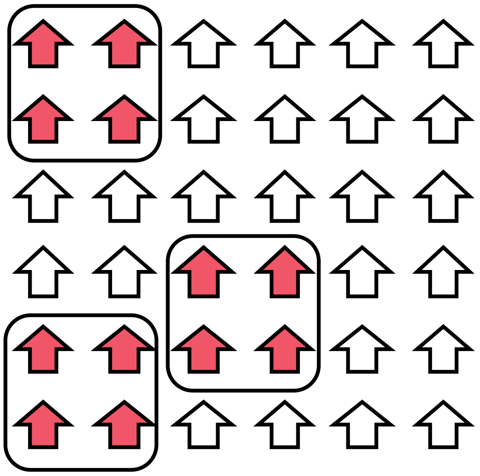
Desafortunadamente, el muestreo por conglomerados modifica la confiabilidad de los datos; esto se debe a que las viviendas que pertenecen a una misma UPM son muy similares y, al momento de hacer las estimaciones, éstas pueden no reflejar las características reales de la población objetivo y estar sesgadas. El impacto que tiene la conglomeración se puede medir a través del coeficiente denominado *efecto de diseño* ($deff$),[8] el cual sirve para indicar qué tan diferentes son las estimaciones si se comparan con una situación ideal (Chambers, 2003; Heeringa *et al*., 2010; Naciones Unidas, 2009).

### 1.2.3 Muestreo estratificado

Consiste en dividir a la población en grupos homogéneos (denominados estratos) y seleccionar un número proporcional de elementos en cada uno de ellos, tal como se observa en la figura 2 (Naciones Unidas, 2009). El ejemplo por excelencia es el de suponer que se requiere una muestra representativa de todos los habitantes del país; entonces, lo más recomendable es tomar en cuenta

Figura 1

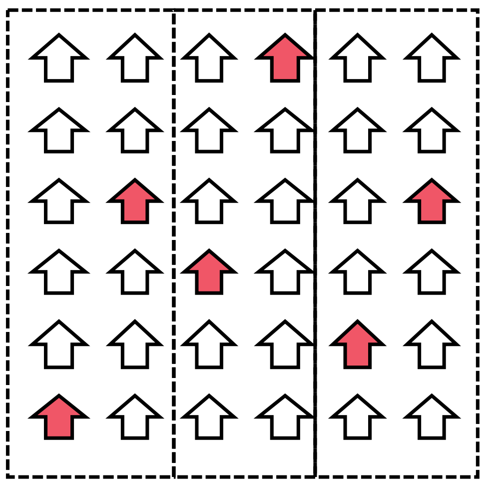
**Muestreo por conglomerados**



**Fuente:** elaboración propia con base en Heeringa *et al*. (2010) y Naciones Unidas (2009).

Figura 2

**Muestreo estratificado**



**Fuente:** elaboración propia con base en Heeringa *et al*. (2010) y Naciones Unidas (2009).

8 Más adelante se explica con detalle cómo se obtiene esta medición y cuál es su significado.

la división política y elegir a algunas personas en cada estado; siguiendo este procedimiento, se puede asegurar que todos contribuyan con, al menos, un elemento y como resultado se tiene una mejor representatividad a nivel nacional.

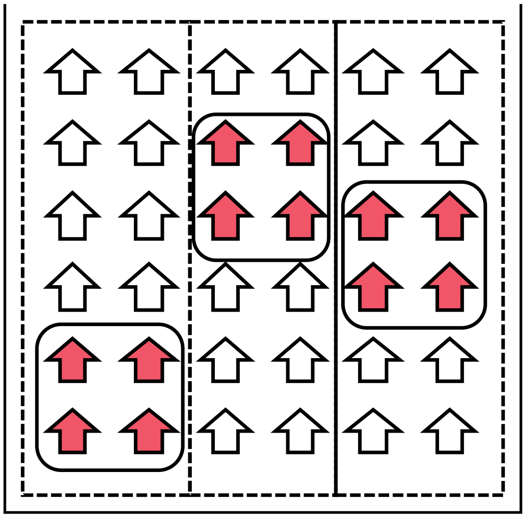
En la analogía anterior, cada estado representaría a un estrato y la división política es información adicional que se requiere para hacer la selección. En el muestreo estratificado se tiene la misma situación, es decir, se requiere de información adicional para poder generar los estratos y hacer la selección. Es por ello que el INEGI usa la información de los eventos censales para construir la estratificación de las viviendas de acuerdo con su nivel socioeconómico y por tamaño de localidad (Landeros, 2013; INEGI, 2009).

Igual que sucede con el muestreo por conglomerados, al momento de hacer la estratificación se altera la calidad de la información; sin embargo, en este caso, el efecto es positivo ya que se reduce el sesgo y se generan estimaciones más confiables debido a que existe una mejor representación de toda la población (Chambers, 2003; Heeringa *et al*. 2010; Naciones Unidas, 2009). Debido a esta característica, es bastante común que los institutos de estadística consideren este tipo de muestreo en combinación con las UPM para mejorar la precisión de las estimaciones y, así, tener una buena representación de la población objetivo; esto da origen a los muestreos estratificados y por conglomerados (ver figura 3).

### 1.3 Número de etapas de selección

Figura 3

**Muestreo estratificado y por conglomerados**

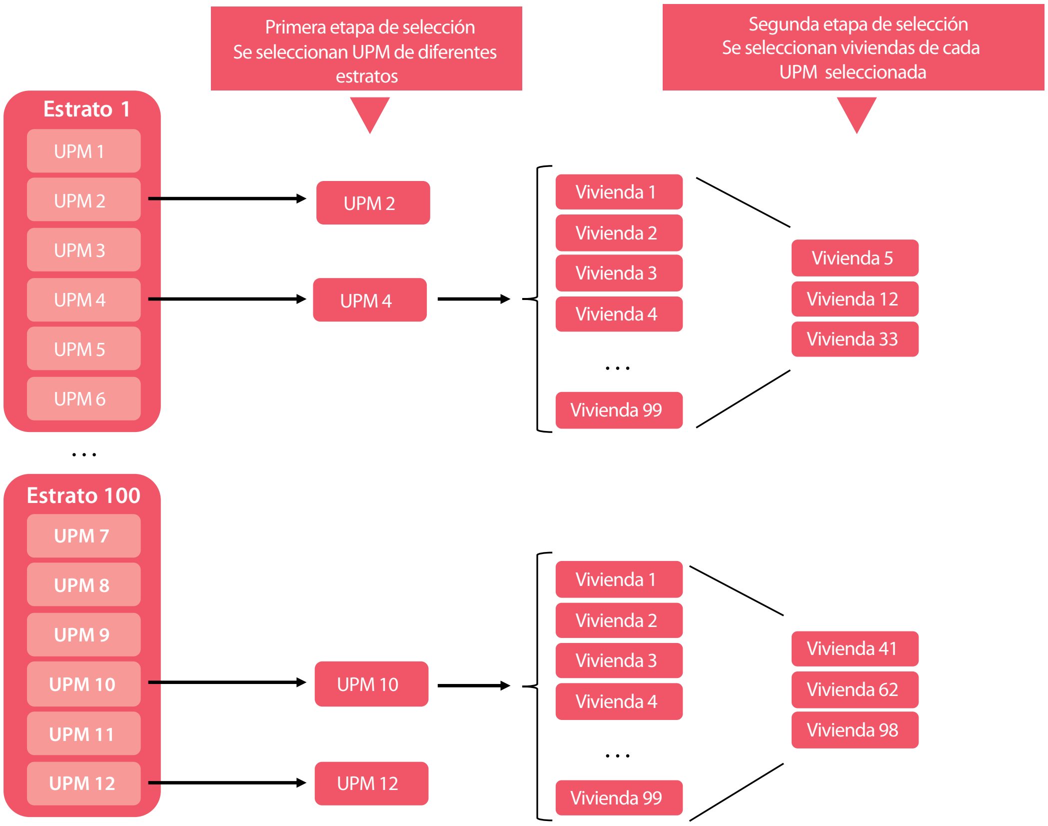


**Fuente:** elaboración propia con base en Heeringa *et al*. (2010) y Naciones Unidas (2009).

Una vez que se tiene el método para seleccionar la muestra, el siguiente paso es identificar a las viviendas que serán entrevistadas, lo cual se lleva a cabo en una o varias etapas de selección. El número puede variar dependiendo de la unidad de análisis y las características de la muestra, sin embargo, cuando son dos o más etapas, cada una de ellas está vinculada con la anterior ya que la muestra en una etapa en particular depende de las unidades seleccionadas en la previa (Heeringa, *et al.*, 2010; Naciones Unidas, 2009); por ejemplo, en la ENOE y la Encuesta Intercensal 2015, las viviendas fueron seleccionadas en dos etapas (ver figura 4): en la primera se escogieron ciertos conglomerados de viviendas (UPM) en cada uno de los estratos; en la segunda, de aquellos que fueron seleccionados se eligieron (de manera aleatoria) las viviendas que serían entrevistadas (INEGI, 2009; INEGI, 2013a).

Figura 4

**Esquema de muestreo bietápico (ENOE y Encuesta Intercensal 2015)**




**Fuente:** elaboración propia con base en INEGI (2009), p. 15.

No obstante, el hecho de que en cada etapa se elijan ciertos casos y se descarten otros, también afecta a las estimaciones ya que se generan probabilidades de selección desigual; por ejemplo, si una vivienda pertenece a una UPM que no fue seleccionada en la primera etapa (en la figura 4, UPM 2), entonces se descarta para la segunda etapa y, de forma automática, queda fuera de la muestra. Para corregir ésta y otras situaciones que se explican a continuación, se requiere del uso de ponderadores.

## 1.4 Ponderación de la muestra

Al momento de configurar una muestra representativa mediante un muestreo complejo surgen algunos inconvenientes. El primero, y más evidente, es que las viviendas no tienen una misma probabilidad de selección, ya que dependen tanto del tamaño de la UPM como del número de etapas de selección. El único caso donde las viviendas tienen una probabilidad de selección igual es en el muestreo aleatorio simple, donde no existe estratificación ni conglomeración, pero en encuestas de gran escala es difícil que se aplique (Heeringa *et al*., 2010, Naciones Unidas, 2009). Además, la identificación de las viviendas no garantiza que la encuesta realmente se llegue a levantar, pues existen factores externos que pueden impedir su aplicación. Esto significa que, del total de viviendas seleccionadas, en algunas no se logra obtener información. Cuando este número es pequeño se puede ignorar, pero en la medida en que esta proporción se incrementa, se deben hacer ajustes para reducir el impacto en el diseño estadístico (Cochran, 1977).

De igual forma, es frecuente que se presenten problemas al momento de captar ciertas características de la población. Cuando éstas son poco frecuentes, o bien, la población que las tiene se encuentra en zonas muy específicas, se pueden presentar problemas de submuestreo. Esto quiere decir que el número de viviendas es insuficiente para dar cuenta de rasgos muy particulares, tal como sería el caso de las personas con nivel de estudios de doctorado.[9] También ocurre el caso opuesto, es decir, tratar de investigar características muy comunes en donde no se requiere de tantas observaciones, lo cual se conoce como sobremuestreo, y un ejemplo típico es la asistencia escolar de niños en zonas urbanas (Hansen, 2011; Graubard y Korn, 2002).

Para corregir estos problemas, se requiere el uso de *ponderadores*, también llamados factores de expansión ($fac$) que, en términos muy generales, se definen como *el inverso de la probabilidad de selección* (Naciones Unidas, 2009:61). A diferencia de aquellos casos donde la muestra se construye a partir de un muestreo aleatorio simple y la ponderación es igual para todos los elementos, cuando se trata de una muestra con varias etapas de selección, la situación es diferente, ya que se requiere hacer varias adecuaciones. De esta manera, el factor de expansión final es el resultado de ajustes por: probabilidad de selección desigual ($fac_{ps,i}$), no respuesta ($fac_{nr,i}$) y problemas con la cobertura de la encuesta ($fac_{cob,i}$). En términos matemáticos, el factor de expansión está dado por:

$$fac_{final,i} = fac_{ps,i} \times fac_{nr,i} \times fac_{cob,i} \qquad (2)$$

Además de estos ajustes, el INEGI hace uno más que está relacionado con las proyecciones demográficas. El objetivo es que los resultados de la encuesta por dominio (ciudades autorrepresentadas, complemento rural y urbano) sean semejantes a la población que estima el Consejo Nacional de Población (CONAPO)[10] en periodos intercensales (INEGI, 2007). Al llevar a cabo este último ajuste ($fac_{proy,i}$), la ecuación anterior se modifica de la siguiente manera:

$$fac_{final,i} = fac_{ps,i} \times fac_{nr,i} \times fac_{proy,i} \times fac_{cob,i} \qquad (3)$$

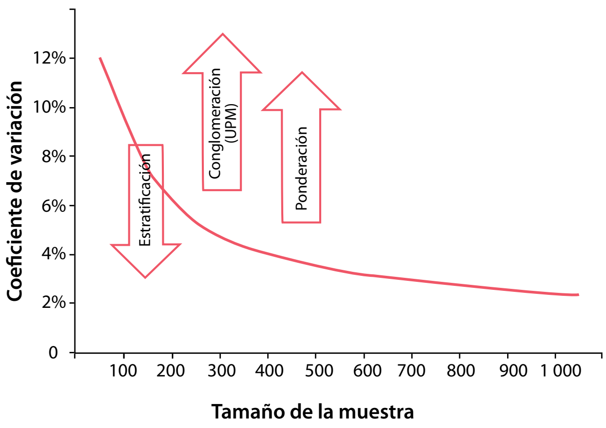
9 Esta situación se va a comprobar más adelante cuando se realicen algunas estimaciones sobre el nivel de escolaridad de la población.

10 El comunicado completo se puede consultar en: http://www.beta.inegi.org.mx/contenidos/proyectos/enchogares/regulares/enoe/doc/Nota_Result_Proy.pdf

Sin embargo, al igual que sucede con la *conglomeración* y la *estratificación*, cuando se aplica una *ponderación* a la muestra se afecta la confiabilidad de los datos. La diferencia está en que estos tres elementos tienen un impacto diferente: mientras que con la conglomeración y ponderación el error estándar aumenta, con la estratificación se reduce. En la gráfica 1, Heeringa *et al.* (2010) muestran esta relación.

**Gráfica 1**

**Efecto de la conglomeración, estratificación y ponderación en el efecto de diseño**




**Fuente:** elaboración propia con base en Heeringa *et al.* (2010), p. 23.

## 1.5 Efecto de diseño (*deff*)

Es una medida que sirve para evaluar el impacto que tiene el esquema de muestreo en la calidad de los datos. En esencia, este indicador compara la varianza de un estimador que se obtiene a partir de un muestreo particular y lo que se obtendría bajo un escenario ideal, es decir, usando un muestreo aleatorio simple ($mas$). De manera formal, si $\hat{\theta}$ es un estimador que se obtiene a partir de un muestreo en particular, denominado $\tau$, entonces el efecto de diseño se define como:

$$deff(\hat{\theta}) = \frac{Var(\hat{\theta})_{\tau}}{Var(\hat{\theta})_{mas}} \qquad (4)$$

donde $\tau$ puede ser: muestreo por conglomerados, estratificado, o bien, una combinación de ambos. Cuando este coeficiente es mayor a 1 significa que existe una mayor variabilidad y, por lo tanto, hay menos precisión en las estimaciones; en el caso contrario, si es menor a 1 entonces el nivel de precisión es mejor, ya que existe menos variación en los datos. Conviene mencionar que el $deff$ no es una medida global de una encuesta, sino que depende de cada estimador en particular, esto quiere decir que dentro de una misma encuesta el efecto de diseño puede variar según el dato que se esté evaluando (Chambers y Skinner, 2003; Naciones Unidas, 2009).

En el caso particular del muestreo por conglomerados existe otra medida cuya fórmula también guarda relación con el $deff$ y que es útil para saber el grado de homogeneidad de los grupos. Se conoce como *coeficiente de correlación intraclase* ($roh$)[11] y su valor oscila entre 1 y -1; mientras más cercano se encuentra de estos valores significa que los elementos son más parecidos, en otras palabras, más homogéneos (Barreto y Raghav, 2015; Naciones Unidas, 2009); su fórmula es la siguiente:

$$roh(\hat{\theta}) = \frac{deff - 1}{\bar{b} - 1} \qquad (5)$$

En esta ecuación, $\bar{b}$ representa el tamaño promedio de las unidades primarias de muestreo. En general, $roh$ es una medida útil para tener una idea de qué tan parecidos son los elementos de un mismo grupo con el supuesto de que si existe una alta correlación y se tienen pocas observaciones, entonces se puede inferir que las estimaciones pueden estar sesgadas; pero lo más recomendable es tomar este valor como una medida secundaria y priorizar el cálculo del efecto de diseño.

### 1.6 Muestra maestra

Para poder llevar a cabo el operativo de campo y levantar la encuesta, se requiere tener información que permita ubicar a las viviendas que serán entrevistadas.[12] Estos datos se encuentran contenidos en la *muestra maestra* que, en términos generales, es el conjunto de UPM de todo el país a partir de la cual se pueden seleccionar submuestras (subconjuntos de UPM) para llevar a cabo las diferentes encuestas (Naciones Unidas, 2009:91). En el caso del INEGI, la muestra maestra se obtiene a partir del Marco Nacional de Viviendas (MNV) y sirve para la planeación de los operativos de campo (Landeros, 2013).

Por otra parte, el MNV es "…un conjunto de materiales que describen áreas perfectamente delimitadas, de las que se tiene información estadística sobre las viviendas y la población que las integran, así como de su ubicación geográfica…" (INEGI, 2014b:33). Tiene una vigencia de 10 años, pero se actualiza de forma constante y se renueva con la información de cada Censo de Población y Vivienda. De esta manera, el MNV 2002 se renovó por completo con el evento censal del 2010 y se generó el nuevo MNV 2012 que está vigente en la actualidad y sirve como referencia para todos los proyectos que lleva a cabo el Instituto (Landeros, 2013; INEGI, 2007; INEGI, 2009; INEGI, 2011a).

## 2. Cálculo de la varianza

Una buena práctica al momento de analizar los datos de una encuesta es conocer su esquema de muestreo. Pasar por alto esta información puede generar inferencias erróneas ya que los datos de encuestas complejas[13] "…no tienen la misma probabilidad de selección ni son independientes…"

11 El problema es que no todos los paquetes estadísticos tienen implementado su cálculo, a excepción de unos cuantos como $R$ o $SAS$.

12 Se recomienda de manera amplia la lectura de INEGI (2009) donde se expone con toda claridad la forma en que se localizan las viviendas en campo; incluso, se presentan algunos ejemplos de los listados que son utilizados por los entrevistadores en sus actividades diarias.

13 Para considerarse como tal, debe tener varias etapas de selección, conformación de estratos y probabilidades de selección desigual (INEGI, 2011b).

(Hans-Vaughn *et al*., 2011:70). Esto significa que las estimaciones pueden estar sesgadas y no reflejar el comportamiento real de la población objetivo.

Ante esta situación, existen algunas medidas de dispersión que son útiles para evaluar la calidad de un dato que se genera a partir de una encuesta compleja. Dentro de éstas se encuentran los errores estándar y de muestreo, el intervalo de confianza y el coeficiente de variación (Carsey, 2014; EUSTAT, 1998; Steven, 1999; Naciones Unidas, 2009; Wolter, 2009). Esta última es de gran importancia, pues refleja la magnitud relativa que tiene dicho error estándar con respecto al estimador de referencia, y entre más pequeño sea este valor, mejor es la precisión. Si bien no existe un consenso unánime sobre qué valores son los más adecuados, el INEGI considera que un dato es de buena calidad si el coeficiente de variación está por debajo de 15%, aceptable si se encuentra entre 15 y 25% y de baja calidad[14] cuando supera 25 por ciento. La fórmula para obtener este coeficiente es:

$$CV = \frac{\sqrt{s}}{|\bar{\theta}|} \qquad (6)$$

El cálculo del *coeficiente de variación* depende de la estimación $\bar{\theta}$ y del error estándar $\sqrt{s}$ que, a su vez, depende de la varianza muestral $s$; por lo tanto, esta última es la medición clave para determinar si un dato es significativo, es decir, si su valor refleja el comportamiento de la población objetivo o no. A diferencia de cómo se estimaría este valor usando un esquema de muestreo aleatorio simple, cuando se trata de muestras complejas, el procedimiento puede resultar más complicado ya que se debe tomar en cuenta la estratificación y conglomeración de las observaciones.

Para resolver estos problemas existen diversos paquetes estadísticos (*R*, *SPSS*, *SAS*, *Stata*, *SUDAAN*) que permiten incorporar el esquema de muestreo al momento de analizar los datos de encuestas complejas.[15] De acuerdo con Siller y Tompkins (2005), varios de ellos producen resultados muy similares, por lo que el foco de atención está en elegir el método para calcular las medidas de dispersión. En este sentido, las técnicas más usadas —y que se encuentran en la mayoría de los programas— se dividen en dos grupos: la replicación de la varianza y linealización por series de Taylor (ver cuadro 2).

## 2.1 Técnicas de replicación de varianza

La idea central de estos métodos es crear *réplicas*[16] de la muestra y usarlas como referencia para calcular los estadísticos correspondientes (ver figura 5). El supuesto es que éstas tienen las mismas características de la muestra original, por lo que se pueden hacer inferencias sobre su comportamiento; además, existen varios submétodos[17] (*BBR, JKn* o *Bootstrap*) que siguen la misma lógica y se ajustan a las características de la muestra. Como es de esperarse, la generación

14 http://www.inegi.org.mx/est/contenidos/proyectos/encuestas/hogares/regulares/enoe/referenciaps_enoe_significancia.aspx

15 Se recomienda consultar a Damico (2009), quien hace una comparación entre *R*, *Stata*, *SAS* y *SUDAAN*.

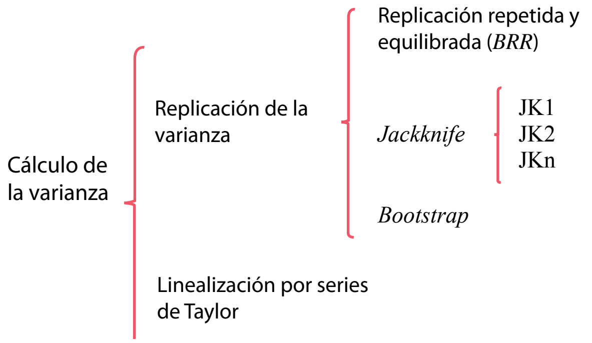
16 Subconjuntos de la población original.

17 En esta sección se ofrece un panorama de estos métodos; sin embargo, para profundizar en el tema, se recomienda consultar a Chernick y LaBudde (2011), quienes no solo dan una explicación matemática con más detalle, sino su implementación en *R*.

de estas réplicas puede presentar problemas para encuestas muy grandes, sin embargo, la virtud de estas técnicas es que resultan muy útiles para estimaciones no lineales, como: intervalos de confianza, medianas o percentiles (Chernick y LaBudde, 2011).

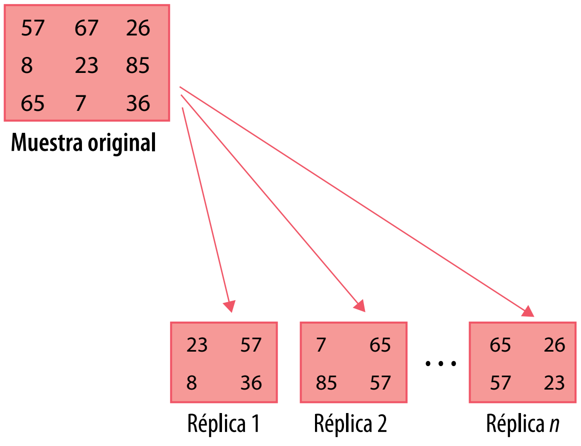
**Cuadro 2**

**Métodos para la estimación de la varianza**




**Fuente:** elaboración propia con base en Carsey y Harden (2014).

**Figura 5**




Las *réplicas balanceadas* (BRR) es una técnica donde se divide a cada estrato en dos UPM[18] (ver tabla 1) y se selecciona un conjunto balanceado de réplicas que permita tener información de la muestra original; a este grupo mínimo se le conoce como *media muestra equilibrada*, y el balanceo se realiza a partir de la matriz de Hadamard (Wolter, 1985). El hecho de considerar menos réplicas de todas las que se pueden formar es el rasgo distintivo de este método; sin embargo, un punto en detrimento de este procedimiento es que no todas las encuestas tienen estratos con dos UPM, pero con ayuda de los paquetes estadísticos es posible crear pseudoestratos y así satisfacer este criterio (Chernick y LaBudde, 2011; Carsey y Harden, 2014).

18 Por ejemplo, si hay $P$ estratos, entonces existen $2^P$ réplicas posibles.

Tabla 1

**Réplicas para una muestra con tres estratos y dos UPM**

| Estrato | UPM | Réplicas | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Estrato 1 | $UPM_1$ | √ | | | √ | | | √ | √ |
| | $UPM_2$ | | √ | √ | | √ | √ | | |
| Estrato 2 | $UPM_1$ | √ | | | | √ | √ | | √ |
| | $UPM_2$ | | √ | √ | √ | | | √ | |
| Estrato 3 | $UPM_1$ | √ | | √ | | | √ | √ | |
| | $UPM_2$ | | √ | | √ | √ | | | √ |
| √ | Sí está en la réplica. | | | | | | | | |

**Fuente:** elaboración propia con base en Wolter (1985), capítulo 3.

Por otro lado, el método conocido como *Jackknife* genera tantas réplicas como sea necesario para que todas las observaciones de la muestra se encuentren en alguna de ellas.[19] A diferencia de la técnica anterior, en ésta existen tres variaciones al método: *Jackknife1 (JK1), Jackknife2 (JK2)* y *Jackknifen (JKn)*. El primero se usa cuando se tiene un esquema de selección aleatorio simple y únicamente consiste en eliminar una observación en cada réplica; el segundo se utiliza cuando se tienen dos unidades primarias de muestreo en cada estrato y, para formar las réplicas, se suprime de manera aleatoria una UPM de cada estrato (similar a la tabla 1); el tercero es la generalización del método y se aplica para muestras que tienen dos o más UPM en cada estrato (Cherkick y LaBudde, 2011; Carsey y Harden, 2014). El *JKn* es muy utilizado en encuestas de gran tamaño ya que ofrece una gran flexibilidad y se adapta a los diferentes esquemas de muestreo; por ejemplo, en la Encuesta Intercensal 2015 se usó para calcular el coeficiente de variación de la tasa global de fecundidad y la tasa de mortalidad infantil (Cherkick y LaBudde, 2011; Carsey y Harden, 2014; INEGI, 2015a).

El *Bootstrap* es una técnica que consiste en seleccionar (con reposición) muestras del mismo tamaño que la original y, con ellas, estimar los estadísticos correspondientes. Igual que en los casos anteriores, este método es muy eficaz para indicadores que son difíciles de calcular desde el punto de vista algebraico,[20] como los percentiles y las medianas. La característica principal de este método es que no se establece un número mínimo de réplicas para hacer los cálculos, sin embargo, entre mayor sea el número, mejor es la precisión (Cherkick y LaBudde, 2011). Para ilustrar esta situación, en la gráfica 2 se comparan los resultados de considerar cien y mil réplicas para obtener el promedio de una variable; es claro que este método requiere procesar un gran volumen de información, aun cuando esto pueda verse limitado por la memoria de las computadoras.[21]

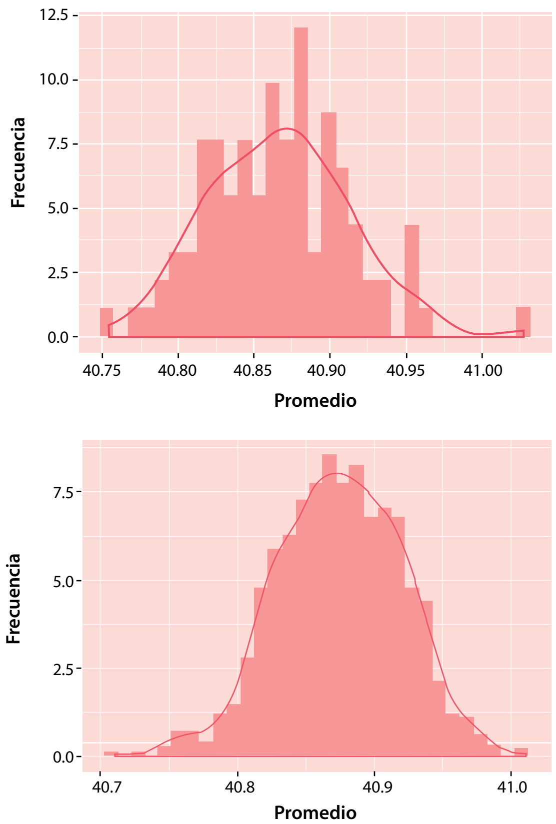
19 En términos formales, la idea es dividir las $n$ observaciones de la encuesta en $k$ grupos de $m$ elementos cada uno de tal forma que $n = km$.

20 La dificultad radica en que estos valores dependen del tamaño de la muestra.

21 El código de la simulación se encuentra disponible en: https://rpubs.com/jcms2665/MAHSDU

Gráfica 2

**Estimación de la media con cien y mil réplicas**




**Fuente:** cálculos propios con base en la ENOE, primer trimestre de 2016.

## 2.2 Linealización por series de Taylor

La idea básica de esta técnica es calcular la varianza de los estimadores a partir de su desagregación mediante los primeros términos de la serie de Taylor, en partes más simples y fáciles de manejar (Heeringa *et al.*, 2010). Con este método, el estimador a evaluar se expresa en términos de numerador ($nm$) y denominador ($d$); por ejemplo, si se requiere calcular la varianza de un estimador $z$, éste se puede expresar de la forma:

$$z = \frac{nm}{d} \qquad (07)$$

Así, la ecuación a resolver es la siguiente:[22]

$$v(z) = \frac{1}{x^2}\{v(d) + r^2 * v(nm) - 2r * cov(nm, d)\} \qquad (8)$$

22 Para una explicación más formal de este método, se recomienda consultar Chambers y Skinner (2003) o bien Naciones Unidas (2009).

Entonces, en lugar de evaluar de forma directa la varianza de $z$, se obtiene la de $nm$ y $d$ por separado, la covarianza entre ambas y, por último, la varianza del coeficiente $z$.

## 3. Aplicación

En esta sección se presenta un ejemplo del manejo de muestras complejas utilizando la ENOE, cuyo objetivo es ofrecer información sobre las características de la fuerza de trabajo de la población y en donde el indicador de referencia para calcular su tamaño de muestra es la *tasa de desocupación*. Además, un rasgo distintivo de esta encuesta es que sirve para el levantamiento de módulos que, en esencia, son encuestas de menor tamaño que usan la infraestructura operativa de la ENOE. En cuanto al esquema de muestreo, se trata de una muestra compleja, la cual es estratificada, por conglomerados, con dos etapas de selección (bietápico) y cuyo marco muestral es el MNV 2012 (INEGI, 2007).

El análisis empírico se lleva a cabo en *R*[23] con el paquete *survey*, el cual tiene varias funciones que permiten definir el esquema de muestreo,[24] crear pseudoestratos para el tratamiento de estratos con una sola UPM[25] y obtener los estadísticos descriptivos[26] para el análisis de las estimaciones. En este paquete se pueden hacer las estimaciones vía linealización por series de Taylor, sin embargo, tiene la facilidad de hacer el ajuste a otros métodos,[27] como: *Bootstrap, Jackknife* y *BRR*, entre otros.

### 3.1 Razones

En este ejercicio se plantea el análisis de la población ocupada que, de acuerdo con los resultados publicados,[28] asciende a 50 778 629 con un coeficiente de variación[29] de 0.43. De los ocupados, un dato que sirve para conocer las características laborales es el promedio de horas a la semana que trabajan las personas. Para analizar el efecto que tiene el diseño de la muestra, se plantea el análisis de esta variable suponiendo cuatro escenarios de muestreo: aleatorio simple, por conglomerados, estratificado, estratificado y por conglomerados; además, cada uno de ellos se evalúa mediante dos técnicas diferentes: linealización por series de Taylor (ver tabla 2) y *Bootstrap* (ver tabla 3).

---

23 Se recomienda ampliamente la lectura del Lumley (2010), donde se describe paso a paso el manejo de muestras complejas con $R$, así como la documentación del paquete que se encuentra disponible en https://cran.r-project.org/web/packages/survey/index.html

24 *svydesign*.

25 *survey.lonely.psu*.

26 *svymean*.

27 Los ajustes se pueden hacer con el comando *as.svrepdesign*.

28 Las cifras corresponden al primer trimestre del 2016.

29 Las pruebas de significancia estadística de los tabulados se encuentran disponibles en la página http://www3.inegi.org.mx/sistemas/tabuladosbasicos/tabtema.aspx?s=est&c=33699

Tabla 2

**Cálculo de la media por series de Taylor**

| Método | Tipo de muestreo | Estadísticos | | | | |
|---|---|---|---|---|---|---|
| | | Estimación | $SE_x$ | *deff* | *cv* (%) | *roh* |
| Series de Taylor | Aleatorio simple | 40.83 | 0.0459 | 1.00 | 0.11 | 0.00 |
| | Por conglomerados | 40.83 | 0.1034 | 5.17 | 0.25 | 0.52 |
| | Estratificado | 40.83 | 0.0667 | 2.15 | 0.16 | 0.14 |
| | Estratificado y por conglomerados | 40.83 | 0.0972 | 4.57 | 0.24 | 0.45 |

**Fuente:** elaboración propia con datos de la ENOE, primer trimestre del 2016.

Tabla 3

**Cálculo de la media por *Bootstrap* (cien réplicas)**

| Método | Tipo de muestreo | Estadísticos | | | | |
|---|---|---|---|---|---|---|
| | | Estimación | $SE_x$ | *deff* | *cv* (%) | *roh* |
| *Bootstrap* | Aleatorio simple | 40.83 | 0.0438 | 1.00 | 0.11 | 0.00 |
| | Por conglomerados | 40.83 | 0.0910 | 4.01 | 0.22 | 0.38 |
| | Estratificado | 40.83 | 0.0567 | 1.56 | 0.14 | 0.07 |
| | Estratificado y por conglomerados | 40.83 | 0.0944 | 4.32 | 0.23 | 0.41 |

**Fuente:** elaboración propia con datos de la ENOE, primer trimestre del 2016.

Usando la linealización por series de Taylor, la primera observación es que el promedio de horas es el mismo (40.83) en todos los casos, sin embargo, para los demás estadísticos, la situación es diferente y su variación depende del tipo de muestreo que se está suponiendo. En el caso del coeficiente de variación, todos los valores se encuentran por debajo de 15%, lo cual significa que el promedio de horas es un dato de buena calidad; no obstante, conviene observar que su valor más alto se encuentra en el muestreo por conglomerados y es el resultado de que las observaciones que están en una misma UPM comparten ciertas características y el sesgo de la estimación aumenta.

El efecto de diseño tiene un comportamiento similar y, al igual que en el caso anterior, su valor más alto se encuentra en el escenario donde los datos provienen de un muestreo por conglomerados; mientras que el más bajo es 1 y se obtiene cuando las observaciones se producen suponiendo un muestreo aleatorio simple. Sin embargo, la cifra que debe tomarse como referencia es 4.57, ya que la ENOE tiene un diseño estratificado y por conglomerados; este valor se encuentra entre 5.17 (conglomeración) y 2.15 (estratificación), por lo tanto, se observa el efecto mostrado en la gráfica 1. Por último, el coeficiente de correlación intraclase también muestra que cuando se supone una conglomeración su valor es mayor y se reduce cuando la muestra solo es estratificada.

En la tabla 3 se replican los mismos escenarios, pero esta vez usando la técnica denominada *Bootstrap*. Como se puede observar, los resultados son muy similares a los que se obtuvieron con la linealización por series de Taylor, tal como sucede en los estudios de Damico (2009) y Siller y Tomkins (2005). Además, de acuerdo con Chernick y LaBudde (2011), las diferencias se deben a que esta técnica requiere de muchas réplicas para mejorar las estimaciones. Para este ejercicio, se consideraron cien, pero, como ya se ha mencionado, lo recomendable es incrementar el número de réplicas.

## 3.2 Totales

El análisis no solo se limita a indicadores de razón, también se puede aplicar a otro tipo de estimadores, como los totales. Para seguir con el ejemplo de la población ocupada y ejemplificar esta situación, ahora se analiza el nivel de estudios de las mujeres que se declaran ocupadas en la Encuesta Nacional de Ocupación y Empleo. En este caso, se trata de una población más específica (19 millones, aproximadamente, para el primer trimestre del 2016), por lo que vale la pena saber si la información que se genera a partir de esta nueva subpoblación sigue siendo confiable desde el punto de vista estadístico.

Para hacer esta valoración, se obtienen los mismos estadísticos de los cuadros anteriores, donde se observa que la categoría *Doctorado* tiene un coeficiente de variación de 15.03%, lo cual significa que este dato tiene una calidad aceptable; sin embargo, la de *Preescolar* supera el umbral de 25%, lo que la convierte en un indicador de baja calidad. Si bien la decisión de usar estos datos depende del contexto y del juicio de cada investigador, se debe tener presente que estos valores pueden no reflejar las características reales de la población (ver tabla 4).

**Tabla 4**

**Estimación de totales por series de Taylor**

| Categoría | Estadísticos | | | |
|---|---|---|---|---|
| | Estimación | $SE_x$ | *deff* | $cv\,(\%)$ |
| Ninguno | 733 720 | 26 493 | 3.40 | 3.61 |
| Preescolar | 6 319 | 1 704 | 1.57 | 26.96 |
| Primaria | 4 182 075 | 60 191 | 3.78 | 1.44 |
| Secundaria | 5 145 469 | 62 013 | 3.48 | 1.21 |
| Preparatoria o bachillerato | 3 581 350 | 51 403 | 3.10 | 1.44 |
| Normal | 79 750 | 7 229 | 2.25 | 9.06 |
| Carrera técnica | 1 220 871 | 30 565 | 2.79 | 2.50 |
| Profesional | 3 928 343 | 56 445 | 3.48 | 1.44 |
| Maestría | 369 488 | 18 611 | 3.27 | 5.04 |
| Doctorado | 38 883 | 5 842 | 3.01 | 15.03 |
| No sabe | 10 459 | 2 791 | 2.55 | 26.69 |

Aceptable.

Baja calidad.

**Fuente:** elaboración propia con datos de la ENOE, primer trimestre del 2016.

### 3.3 Modelos estadísticos

Como se ha visto hasta ahora, al momento de incorporar el diseño estadístico en el análisis de los datos se obtiene otra perspectiva sobre la calidad de la información. En otras palabras, se tiene una idea si los datos en realidad están reflejando el comportamiento de la población objetivo o se trata de estimaciones sesgadas; sin embargo, el esquema de muestreo no solo afecta a las estimaciones puntuales, como razones o totales, sino también a los modelos estadísticos.

Retomando el ejemplo de la población ocupada, a continuación se propone modelar el impacto que tienen las características sociodemográficas de las personas sobre el tiempo que le dedican al trabajo mediante un modelo de regresión lineal.[30] De esta manera, la variable dependiente es el número de horas trabajadas a la semana ($hrsocup$) y las covariables son sexo ($sexo$), nivel de escolaridad ($niv\_esc$), estado conyugal ($\mathbf{e\_con}$), edad ($edad$) y posición en la ocupación ($pos\_ocu$). Con este modelo, el objetivo primario es mostrar el efecto de la estratificación y conglomeración en los resultados de la regresión lineal cuya ecuación es:

$$\begin{aligned} hrsocup = \beta_0 + \beta_1 sexo + \beta_2 niv_{esc} + \beta_3 e_{con} \\ + \beta_4 edad + \beta_5 pos\_ocu + e \end{aligned} \qquad (9)$$

A diferencia de los ejercicios anteriores, en éste se plantean dos escenarios de muestreo (aleatorio simple y complejo) y dos métodos para estimar la varianza (por series de Taylor y *Bootstrap*). Al comparar los resultados de ambos modelos, si bien es cierto que algunos son muy similares, en otros se presentan diferencias en nivel de significancia ($\rho$-$value$) y en la magnitud de los coeficientes ($\beta$). En la tabla 5 se muestran de manera más específica las variaciones de los coeficientes entre un modelo y otro; además se identifica a la categoría que tiene la mayor diferencia en cada una de las variables. En este sentido, la mayor diferencia se encuentra en la categoría de nivel medio superior con 0.7 unidades; le siguen la de divorciada(o) y empleadores con 0.5 unidades cada uno.

En lo que se refiere a las diferencias en el nivel de significancia, la mayor de ellas se presenta en la variable de estado conyugal,[31] pues dos de las seis categorías resultan ser no significativas al momento de incorporar el diseño de la muestra, lo cual es el resultado de la estratificación, conglomeración y el número de etapas de selección. Esta situación sí modifica la interpretación general del modelo, ya que solo tres de las opciones son susceptibles de ser analizadas. En el resto de las variables se presentan únicamente diferencias en la magnitud del coeficiente.

También, se evaluó el modelo de regresión lineal usando la técnica de *Bootstrap* (ver tabla 6) considerando cien réplicas.[32] Los resultados son muy similares, tanto en la magnitud de los

---

**30** Se recomienda la lectura de Manzi *et al.* (2010), quienes presentan el efecto que tienen las muestras complejas en un modelo de regresión logística.

**31** Conviene mencionar que, si bien existen diferencias en las categorías *No sabe* y *No especificados*, éstas son categorías residuales y no se espera que sean significativas.

**32** Para profundizar en el análisis de los modelos estadísticos a partir del método *Bootstrap*, se recomienda ver Fox y Weisber (2012), quienes también describen su implementación en *R*.

coeficientes como en la significancia de las variables, a los que se obtuvieron con la linealización por series de Taylor, por lo que optar por una técnica u otra depende de la implementación de cada paquete.

Tabla 5

**Resultados del modelo de regresión lineal por series de Taylor**

| Variable | Modelo suponiendo un muestreo aleatorio simple (omitiendo el esquema de muestreo) | | | | | Modelo suponiendo un muestreo estratificado y por conglomerados | | | | | Diferencia en $\beta$ | Diferencia en *p-value* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\beta$ | $SE_x$ | $P$ | IC (90%) LI | IC (90%) LS | $\beta$ | $SE_x$ | $P$ | IC (90%) LI | IC (90%) LS | | |
| **Sexo** | | | | | | | | | | | | |
| Mujer | -7.473 | 0.093 | 0.000 | -7.656 | -7.290 | -7.083 | 0.145 | 0.000 | -7.367 | -6.799 | 0.4 | 0.0 |
| **Nivel de instrucción** | | | | | | | | | | | | |
| Primaria completa | 1.902 | 0.178 | 0.000 | 1.553 | 2.250 | 1.911 | 0.276 | 0.000 | 1.371 | 2.451 | 0.0 | 0.0 |
| Secundaria completa | 2.314 | 0.167 | 0.000 | 1.986 | 2.643 | 2.596 | 0.279 | 0.000 | 2.049 | 3.144 | 0.3 | 0.0 |
| Medio superior y superior | 0.464 | 0.167 | 0.005 | 0.137 | 0.792 | 1.188 | 0.297 | 0.000 | 0.605 | 1.771 | 0.7 | 0.0 |
| No especificado | 3.684 | 1.777 | 0.038 | 0.202 | 7.166 | 0.104 | 3.141 | 0.974 | -6.053 | 6.261 | 3.6 | 0.9 |
| **Estado conyugal** | | | | | | | | | | | | |
| Está separada(o) | -0.508 | 0.237 | 0.032 | -0.973 | -0.043 | -0.321 | 0.355 | 0.365 | -1.017 | 0.374 | 0.2 | 0.3 |
| Está divorciada(o) | -0.805 | 0.326 | 0.014 | -1.445 | -0.165 | -0.341 | 0.523 | 0.514 | -1.365 | 0.683 | 0.5 | 0.5 |
| Está viuda(o) | -2.637 | 0.307 | 0.000 | -3.238 | -2.037 | -2.322 | 0.495 | 0.000 | -3.291 | -1.352 | 0.3 | 0.0 |
| Está casada(o) | -1.561 | 0.130 | 0.000 | -1.817 | -1.306 | -1.377 | 0.212 | 0.000 | -1.793 | -0.961 | 0.2 | 0.0 |
| Está soltera(o) | -2.898 | 0.141 | 0.000 | -3.174 | -2.622 | -2.586 | 0.223 | 0.000 | -3.024 | -2.149 | 0.3 | 0.0 |
| No sabe | -7.266 | 6.827 | 0.287 | -20.647 | 6.114 | 5.673 | 11.317 | 0.616 | -16.509 | 27.855 | 12.9 | 0.3 |
| **Posición en la ocupación** | | | | | | | | | | | | |
| Empleadores | 2.012 | 0.216 | 0.000 | 1.589 | 2.435 | 1.488 | 0.342 | 0.000 | 0.818 | 2.158 | 0.5 | 0.0 |
| Cuenta propia | -5.188 | 0.118 | 0.000 | -5.419 | -4.957 | -4.740 | 0.209 | 0.000 | -5.148 | -4.331 | 0.4 | 0.0 |
| Trabajadores sin pago | -9.003 | 0.222 | 0.000 | -9.438 | -8.568 | -9.125 | 0.382 | 0.000 | -9.873 | -8.376 | 0.1 | 0.0 |
| **Edad** | -0.026 | 0.004 | 0.000 | -0.034 | -0.018 | -0.034 | 0.007 | 0.000 | -0.047 | -0.020 | 0.0 | 0.0 |
| Constante | 46 540 | 0.252 | 0.000 | 46.047 | 47.034 | 46.059 | 0.450 | 0.000 | 45.177 | 46.941 | | |

Mayor diferencia en *p-value*.

Mayor diferencia en $\beta$.

Tabla 6

## Resultados del modelo de regresión lineal por *Bootstrap*

| Variable | Muestreo estratificado y por conglomerados | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Series de Taylor | | | | | *Bootstrap* | | | | |
| | $\beta$ | $SE_x$ | $P$ | IC (90%) | | $\beta$ | $SE_x$ | $P$ | IC (90%) | |
| | | | | LI | LS | | | | LI | LS |
| **Sexo** | | | | | | | | | | |
| Mujer | -7.083 | 0.145 | 0.000 | -7.367 | -6.799 | -7.083 | 0.153 | 0.000 | -7.364 | -6.801 |
| **Nivel de instrucción** | | | | | | | | | | |
| Primaria completa | 1.911 | 0.276 | 0.000 | 1.371 | 2.451 | 1.911 | 0.260 | 0.000 | 1.386 | 2.435 |
| Secundaria completa | 2.596 | 0.279 | 0.000 | 2.049 | 3.144 | 2.596 | 0.267 | 0.000 | 2.016 | 3.176 |
| Medio superior y superior | 1.188 | 0.297 | 0.000 | 0.605 | 1.771 | 1.188 | 0.311 | 0.000 | 0.568 | 1.809 |
| No especificado | 0.104 | 3.141 | 0.974 | -6.053 | 6.261 | 0.104 | 3.370 | 0.976 | -5.931 | 6.138 |
| **Estado conyugal** | | | | | | | | | | |
| Está separada(o) | -0.321 | 0.355 | 0.365 | -1.017 | 0.374 | -0.321 | 0.352 | 0.364 | -1.063 | 0.420 |
| Está divorciada(o) | -0.341 | 0.523 | 0.514 | -1.365 | 0.683 | -0.341 | 0.538 | 0.527 | -1.340 | 0.658 |
| Está viuda(o) | -2.322 | 0.495 | 0.000 | -3.291 | -1.352 | -2.322 | 0.516 | 0.000 | -3.405 | -1.238 |
| Está casada(o) | -1.377 | 0.212 | 0.000 | -1.793 | -0.961 | -1.377 | 0.217 | 0.000 | -1.795 | -0.958 |
| Está soltera(o) | -2.586 | 0.223 | 0.000 | -3.024 | -2.149 | -2.586 | 0.211 | 0.000 | -3.043 | -2.130 |
| No sabe | 5.673 | 11.317 | 0.616 | -16.509 | 27.855 | 5.673 | 14.628 | 0.699 | -21.372 | 32.717 |
| **Posición en la ocupación** | | | | | | | | | | |
| Empleadores | 1.488 | 0.342 | 0.000 | 0.818 | 2.158 | 1.488 | 0.329 | 0.000 | 0.836 | 2.139 |
| Cuenta propia | -4.740 | 0.209 | 0.000 | -5.148 | -4.331 | -4.740 | 0.232 | 0.000 | -5.207 | -4.273 |
| Trabajadores sin pago | -9.125 | 0.382 | 0.000 | -9.873 | -8.376 | -9.125 | 0.401 | 0.000 | -9.972 | -8.277 |
| **Edad** | -0.034 | 0.007 | 0.000 | -0.047 | -0.020 | -0.034 | 0.008 | 0.000 | -0.048 | -0.020 |
| **Constante** | 46.059 | 0.450 | 0.000 | 45.177 | 46.941 | 46.059 | 0.460 | 0.000 | 45.245 | 46.873 |

Mayor diferencia en $p\text{-}value$.
Mayor diferencia en $\beta$.

## 4. Reflexiones finales

Los datos de encuestas complejas presentan dos grandes problemas: las observaciones no se obtienen de forma aleatoria ni son independientes. Por este motivo, una buena práctica al momento de analizar las encuestas en hogares es conocer el esquema de muestreo, el número de etapas de selección y el factor de expansión para extrapolar los datos. Ignorar estas características sería equivalente a suponer que los datos se obtienen de manera aleatoria, lo cual puede llevar a inferencias erróneas.

En la actualidad, gran parte de los programas estadísticos permite no solo incorporar el diseño de la muestra, sino calcular los estadísticos descriptivos para evaluar la calidad de los datos. En este documento se utiliza $R$ versión 3.3.1 y los métodos de linealizacion por series de Taylor y *Bootstrap*, los cuales ofrecen resultados muy similares. Lo importante es reconocer que los diferentes esquemas de muestreo producen resultados distintos que afectan tanto a las estimaciones como a los modelos de regresión lineal.

## Fuentes


Barreto, H. & M. Raghav. *Understanding and Teaching Within-Cluster Correlation in Complex Surveys. Working Papers*. DePauw University, Department of Economics and Management, 2015.

Canty, A. *Boot: Bootstrap Functions. R package version 3.31*. 2015. Obtenido de https://cran.r-project.org/web/packages/boot/citation.html

Carsey, T. & J. Harden. *Monte Carlo Simulation and Resampling Methods for Social Science*. Thousand Oaks, SAGE Publications, Inc., 2014.

Chambers, R. & C. Skinner. *Analysis of Survey Data*. Chichester, Reino Unido, John Wiley & Sons, 2003.

Chernick, M. & R. LaBudde. *An introduction to bootstrap methods with applications to R*. Hoboken, New Jersey, John Wiley & Sons, 2011.

Cochran, W. *Sampling Techniques*. Massachusetts, John Wiley & Sons, 1977.

Dabbish, L.; C. Stuart; J. Tsay & J. Herbsleb. "Social Coding in GitHub: Transparency and Collaboration in an Open Software Repository", en: *School of Computer Science and Center for the Future of Work*. Pittsburgh, Heinz College, 2012, pp. 1277-1286.

Damico, A. "Transitioning to $R$: Replicating SAS, Stata, and SUDAAN Analysis Techniques in Health Policy Data", en: *The R Journal*. 1(2), 2009, 37-44.

Dembe, A.; J. Partridge & L. Geist. "Statistical software applications used in health services research", analysis of published studies in the U.S. *BMC Health Services Research*. 2011, 252-258.

EUSTAT. *El método de replicación para la estimación de errores de muestreo*. País Vasco, Instituto Vasco de Estadística, 1998.

Fox, J. & S. Weisberg. *Bootstrapping Regression Models in R*. 2012. Obtenido de https://socserv.socsci.mcmaster.ca/jfox/Books/Companion/appendix/Appendix-Bootstrapping.pdf

Graubard, B., & E. Korn. *The Use of Sampling Weights in the Analysis of Survey Data*. Nueva York, División de Estadística de las Naciones Unidas, 2002.

Hahs-Vaughn, D.; C. McWayne; R. Bulotsky-Shearer; X. Wen & A.-M. Faria. *Methodological Considerations in Using Complex Survey Data: An Applied Example With the Head Start Family and Child Experiences Survey*. Orlando, Florida, SAGE, 2011.

Hansen, M.; W. Hurwitz & W. Madow. *Sample Survey Methods and Theory*. Nueva York, John Wiley & Sons, 1953.

Harden, J. "A bootstrap method for conducting statistical inference with clustered data", en: *State Politics & Policy Quarterly*. 2011, 223-246.

Heeringa, S.; B. West & P. Berglund. *Applied Survey Data Analysis*. Boca Raton, FL, CRC Press, 2010.

Holt, D.; T. Smith & P. Winter. "Regression Analysis of Data from Complex Survey", en: *Journal of the Royal Statistical Society*. 143(4), 1980, 474-487.

INEGI. *Cómo se hace la ENOE*. *Métodos y procedimientos*. Aguascalientes, Ags., INEGI, 2007.

________ *Manual para la identificación de viviendas seleccionadas*. Aguascalientes, Ags., INEGI, 2009.

________ *Diseño de la muestra en proyectos de encuesta*. Aguascalientes, Ags., INEGI, 2011a.

________ *Manual del supervisor e instructor supervisor de la ENOE 2011*. Aguascalientes, Ags., INEGI, 2011b.

________ *Manual del entrevistador de la ENOE.* Aguascalientes, Ags., INEGI, 2013a.

________ *Diseño muestral de la Encuesta Nacional sobre Salud y Envejecimiento en México (ENASEM) 2012*. Aguascalientes, Ags., INEGI, 2013b.

_________ *Encuesta Intercensal 2015. Manual del entrevistador*. Aguascalientes, Ags., INEGI, 2014a.

_________ *Encuesta Nacional sobre Uso del Tiempo 2014*. Manual del *entrevistador*. Aguascalientes, Ags., INEGI, 2014b.

_________ *Encuesta Intercensal 2015. Síntesis metodológica y conceptual*. Aguascalientes, Ags., INEGI, 2015a.

_________ *Encuesta Nacional sobre Uso del Tiempo (ENUT) 2014. Documento metodológico*. Aguascalientes Ags., INEGI, 2015b.

_________ *Encuesta Nacional de Ingresos y Gastos de los Hogares 2014*: *diseño muestral*. Aguascalientes, Ags., INEGI, 2015c.

_________ *Encuesta Nacional de la Dinámica Demográfica (ENADID) 2014. Síntesis metodológica*. Aguascalientes, Ags., INEGI, 2015d.

Landeros, A. M. "El Marco Nacional de Viviendas", en: *Reunión Nacional de Estadística*. Aguascalientes, Ags., INEGI, 2013.

Lumley, T. "Analysis of complex survey samples", en: *Journal of Statistical Software*. 9(1), 2004, 1-19.

_________ *Complex Surveys A Guide to Analysis Using R*. Hoboken, New Jersey, John Wiley & Sons, 2010.

_________ *Survey: Analysis of Complex Survey Samples*. 2014. Obtenido de *R package version 3.31*: https://cran.r-project.org/web/packages/survey/index.html

Manzi, A.; F. Munyaneza; F. Mujawase; L. Banamwana; F. Sayinzoga; D. Thomson; ... B. Hedt-Gauthier. "Assessing predictors of delayed antenatal care visits in Rwanda: a secondary analysis of Rwanda demographic and health survey 2010", en: *BMC Pregnancy and Childbirth*. 2014.

Mecatti, F.; P. Luigi Conti & M. Giovanna Ranalli. *Contributions to Sampling Statistics*. New York, Springer, 2014.

Miller, R. "The Jackknife, A review", en: *Biometrika*. 61(1), 1974, 1-15.

Naciones Unidas. *Encuestas de hogares en los países en desarrollo y en transición*. New York, Naciones Unidas, 2007.

_________ *Diseño de muestras para encuestas de hogares: directrices prácticas*. Nueva York, Naciones Unidas, 2009.

Orange, A. "Jackknife Estimation of Sampling Variance of Ratio Estimators in Complex Samples: Bias and the Coefficient of Variation", en: *Educational Testing Service*. 2006, 1-23.

Ruiz de los Santos, S. R. "Deambulando entre los vagoneros del Metro de la Ciudad de México", en: *Iztapalapa*. 2009, 115-135.

Siller, A., & L. Tompkins. "The big four: Analyzing complex sample survey data using SAS, SPSS, STATA, and SUDAAN", en: *SUGI 31*. *Paper 172-31*. 2005.

Steven, P. "Comparison of Variance Estimation Methods for the National Compensation Survey", en: *Proceedings of the Section on Survey Research Methods, American Statistical Association*. 1999.

US Department of Education. National Center for Education Statistics. *Strengths and Limitations of Using SUDAAN, Stata, and WesVarPC for Computing Variances from NCES Data Sets. Working Paper*. 2000.

Wegman, E. & J. Solka. *Statistical Software for Today and Tomorrow*. 2005. Recuperado el 12 de 08 de 2016, de http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.83.1025&rep=rep1&type=pdf

Wolter, K. *Introduction to Variance Estimation*. New York, Springer, 2007.

Wu, C. "Jackknife, bootstrap and other resampling methods in regression analysis", en: *Annals of Statistics*. 14(4), 1986, 1261-1295.